

# On the Prediction of a Nonstationary Geometric Distribution Based on Bayes Decision Theory

Daiki Koizumi <sup>a</sup>

Otaru University of Commerce, 3-5-21, Midori, Otaru-city, Hokkaido, 045-8501, Japan

**Keywords:** Probability Model, Bayes Decision Theory, Nonstationary Geometric Distribution, Hierarchical Bayesian Model, Time Series Analysis.


**Abstract:** This paper considers a prediction problem with a nonstationary geometric distribution in terms of Bayes decision theory. The proposed nonstationary statistical model contains a single hyperparameter, which is used to express the nonstationarity of the parameter of the geometric distribution. Furthermore, the proposed predictive algorithm is based on both the posterior distribution of the nonstationary parameter and the predictive distribution for data, operating with a Bayesian context. Each predictive estimator satisfies the Bayes optimality, which guarantees a minimum mean error rate with the proposed nonstationary probability model, a loss function, and a prior distribution of the parameter in terms of Bayes decision theory. Furthermore, an approximate maximum likelihood estimation method for the hyperparameter based on numerical calculation has been considered. Finally, the predictive performance of the proposed algorithm has been evaluated in terms of both the model selection theory and the predictive mean squared error by comparison with the stationary geometric distribution using real web traffic data.

## 1 INTRODUCTION

The geometric distribution (Johnson and Kotz, 1969) (Hogg et al., 2013) is one of significant discrete probability distributions with at least two definitions. One is that the probability distribution of the number of failures before the first success, with the success probability as the parameter. The other is that the discrete probability distribution of the number of Bernoulli trials needed to get one success given the same parameter. This paper is based on the former definition. Some important characteristics of the geometric distribution are that it is the discrete version of the exponential distribution; that it has the memoryless property; and that it is a special case of the negative binomial distribution. Based on the above definitions and characteristics of the geometric distribution, many applications have been reported, including quality control (Frank C. Kaminsky and Burke, 1992), queueing theory (Winsten, 1959), biology (Ewens, 2004), epidemiology (O. Diekmann, 2000), communication theory (G. Gallager, 1995), computer networks (Bianchi, 2000), and so forth.

In the field of Bayesian statistics (Berger, 1985) (Bernardo and Smith, 1994), on the other hand, the parameter estimation or prediction problems often become intractable. This is because these problems require integral calculations in the denominator of the Bayes theorem depending on a known prior distribution of parameter. However, if the specific distribution of parameter is assumed to be the prior, complex integral calculations can be avoided. In Bayesian statistics, this specific class of prior is called a *conjugate family* (Berger, 1985, pp. 130–132) (Bernardo and Smith, 1994, pp. 265–267). The beta distribution is the natural conjugate prior of the stationary geometric distribution (Bernardo and Smith, 1994).

The above results are limited within the stationary geometric distribution. If the nonstationary probability distributions are assumed, the Bayesian estimation problems become more difficult and more intractable. In such cases, there is no guarantee of the existence of a natural conjugate prior. In this regard, at least two nonstationary probability models have been proposed. One is the Bayesian entropy forecasting (BEF) model (Souza, 1978) in which the Shannon's entropy function and Jaynes' principle of maximum entropy are applied to the model formulation. The other is referred to as the Simple Power Steady Model (SPSM)

<sup>a</sup>  <https://orcid.org/0000-0002-5302-5346>

(Smith, 1979). The SPSM is a time-series model and they have shown certain illustrative probability distributions called linear expanding families in which natural conjugate priors exist (Smith, 1979). Recently, a new similar and particular nonstationary parameter classes with hyperparameter estimation methods have been proposed (Koizumi, 2020; Koizumi, 2021; Koizumi, 2023). Among the aforementioned results, a single hyperparameter is identified as the expression of nonstationarity of the parameter, and its estimation can be achieved through the approximate maximum likelihood estimation with numerical calculation. These results contribute new aspects to the field of empirical Bayes methods (Carlin and Louis, 2000). Furthermore, a Bayesian problem in the context of Bayes decision theory (Berger, 1985) (Bernardo and Smith, 1994) has been considered. Using this approach, the predictive estimator satisfies *Bayes optimality*, which guarantees a minimum mean error rate for predictions.

In this paper, the aforementioned approach to the nonstationary geometric distribution is presented. The proposed nonstationary class of parameter has only a single hyperparameter. This hyperparameter can be estimated from observed data by an approximate maximum likelihood estimation with numerical calculations. Under condition with the known (or estimated) hyperparameter, the posterior distribution of parameter with specific prior can be tractably obtained with simpler arithmetic calculations. This point would be the generalization of the natural conjugate prior with stationary geometric distribution to avoid heavy integral calculations under the equation of Bayes theorem. Moreover, a Bayes optimal prediction algorithm is proposed, which guarantees the a minimum mean error rate for predictions in terms of Bayes decision theory. Finally, evaluation of the predictive performances of the proposed algorithms are via comparison with the results of the stationary geometric distribution using real web traffic data is detailed.

The rest of this paper is organized as follows. Section 2 provides the basic definitions of the proposed nonstationary geometric distribution and some lemmas and corollaries in terms of Bayesian statistics. Section 3 proposes the Bayes optimal predictive algorithm in terms of Bayes decision theory. Section 4 presents numerical examples using real web traffic data. Section 5 presents a discussion on the results of this paper. Section 6 presents the conclusion.

## 2 PRELIMINARIES

### 2.1 Hierarchical Bayesian Modeling with Nonstationary Geometric Distribution

Let  $t = 1, 2, \dots$  be a discrete time index and  $X_t = x_t \geq 0$  be a discrete random variable at  $t$ . Assume that  $x_t = 0, 1, 2, \dots, N$  represents count data with known  $N$  and  $X_t \sim \text{Geometric}(\theta_t)$ , where  $0 < \theta_t \leq 1$ , is a nonstationary parameter at  $t$ . Thus, the probability function of the nonstationary geometric distribution  $p(x_t | \theta_t)$  is defined as follows:

**Definition 2.1.** Nonstationary Geometric Distribution

$$p(x_t | \theta_t) = (1 - \theta_t)^{x_t} \theta_t, \quad (1)$$

where  $x_t = 0, 1, 2, \dots, N$  and  $0 < \theta_t \leq 1$ .  $\square$

**Definition 2.2.** Function for  $\Theta_t, A_t$ , and  $B_t$

Let  $\Theta_t = \theta_t, A_t = a_t$ , and  $B_t = b_t$  be random variables where  $A_t$  and  $B_t$  are mutually independent, then a function for  $\Theta_t$  is defined as,

$$\Theta_t = \frac{A_t}{A_t + B_t}, \quad (2)$$

where  $0 < a_t, 0 < b_t$ .  $\square$

**Definition 2.3.** Nonstationarity of  $A_t, B_t$

Let  $C_t = c_t, D_t = d_t$  be random variables, then the nonstationary functions for  $A_t$  and  $B_t$  are defined as,

$$A_{t+1} = C_t A_t, \quad (3)$$

$$B_{t+1} = D_t B_t, \quad (4)$$

where  $0 < c_t < 1, 0 < d_t < 1$  and they are sampled from the following two types of beta distributions:

$$C_t \sim \text{Beta}[k\alpha_t, (1-k)\alpha_t], \quad (5)$$

$$D_t \sim \text{Beta}[k\beta_t, (1-k)\beta_t], \quad (6)$$

where  $k$  is a real valued constant and  $0 < k \leq 1$ .  $\square$

**Definition 2.4.** Conditional Independence for  $A_t, C_t$  (or  $B_t, D_t$ ) under  $\alpha_t$  (or  $\beta_t$ )

$$p(a_t, c_t | \alpha_t) = p(a_t | \alpha_t) p(c_t | \alpha_t), \quad (7)$$

$$p(b_t, d_t | \beta_t) = p(b_t | \beta_t) p(d_t | \beta_t). \quad (8)$$

$\square$

**Definition 2.5.** Initial Distributions for  $A_1, B_1$

$$A_1 \sim \text{Gamma}(\alpha_1, 1), \quad (9)$$

$$B_1 \sim \text{Gamma}(\beta_1, 1), \quad (10)$$

where  $0 < \alpha_1$  and  $0 < \beta_1$ .  $\square$

**Definition 2.6.** Initial Distributions for  $C_1, D_1$

$$C_1 \sim \text{Beta}[k\alpha_1, (1-k)\alpha_1], \quad (11)$$

$$D_1 \sim \text{Beta}[k\beta_1, (1-k)\beta_1]. \quad (12)$$

□

**Definition 2.7.** Gamma Distribution for  $q$   
Gamma distribution of  $\text{Gamma}(r, s)$  is defined as,

$$p(q | r, s) = \frac{s^r}{\Gamma(r)} q^{r-1} \exp(-sq), \quad (13)$$

where  $0 < q, 0 < r, 0 < s$ , and  $\Gamma(r)$  is the gamma function defined in Definition 2.9. □

**Definition 2.8.** Beta Distribution for  $q$   
Beta distribution of  $\text{Beta}(r, s)$  is defined as,

$$p(q | r, s) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} q^{r-1} (1-q)^{s-1}, \quad (14)$$

where  $0 < q < 1, 0 < r, 0 < s$ . □

**Definition 2.9.** Gamma Function for  $q$

$$\Gamma(q) = \int_0^{+\infty} y^{q-1} \exp(-y) dy, \quad (15)$$

where  $0 < q$ . □

## 2.2 Lemmas

**Lemma 2.1.** Transformed Distribution for  $A_t$

For any  $t \geq 1$ , the transformed random variable  $A_{t+1} = C_t A_t$  in Definition 2.3 follows the following Gamma distribution:

$$A_{t+1} \sim \text{Gamma}(k\alpha_t, 1). \quad (16)$$

□

**Proof of Lemma 2.1.**

See APPENDIX A. □

**Lemma 2.2.** Transformed Distribution for  $B_t$

For any  $t \geq 1$ , the transformed random variable  $B_{t+1} = D_t B_t$  in Definition 2.3 follows the following Gamma distribution:

$$B_{t+1} \sim \text{Gamma}(k\beta_t, 1). \quad (17)$$

□

**Proof of Lemma 2.2.**

The proof is exactly same as Lemma 2.1, replacing  $A_{t+1}$  by  $B_{t+1}$ ,  $C_t$  by  $D_t$ , and  $\alpha_t$  by  $\beta_t$ .

This completes the proof of Lemma 2.2. □

**Lemma 2.3.** Transformed Distribution for  $\Theta_t$

For any  $t \geq 2$ , the transformed random variable  $\Theta_t = \frac{A_t}{A_t + B_t}$  in Definition 2.2 follows the following beta distribution:

$$\Theta_t \sim \text{Beta}(k\alpha_{t-1}, k\beta_{t-1}). \quad (18)$$

□

**Proof of Lemma 2.3.**

See APPENDIX B. □

**Corollary 2.1.** Transformed Initial Distribution for  $\Theta_1$

The transformed random variable  $\Theta_1 = \frac{A_1}{A_1 + B_1}$  in Definition 2.2 follows the following beta distribution:

$$\Theta_1 \sim \text{Beta}(\alpha_1, \beta_1). \quad (19)$$

□

**Proof of Corollary 2.1.**

From Definition 2.5,

$$A_1 \sim \text{Gamma}(\alpha_1, 1),$$

$$B_1 \sim \text{Gamma}(\beta_1, 1).$$

If Lemma 2.3 is applied to the above  $A_1$  and  $B_1$ , then the following holds,

$$\Theta_1 \sim \text{Beta}(\alpha_1, \beta_1). \quad (20)$$

This completes the proof of Corollary 2.1. □

## 3 PREDICTION ALGORITHM BASED ON BAYES DECISION THEORY

### 3.1 Preliminaries

**Definition 3.1.** Loss Function

In this paper, the predictive error for  $x_{t+1}$  is measured by the following squared-error loss function (Berger, 1985, 2.4.2, I., p. 60),

$$L(\hat{x}_{t+1}, x_{t+1}) = (\hat{x}_{t+1} - x_{t+1})^2. \quad (21)$$

□

**Definition 3.2.** Risk Function

The risk function is defined as the expectation of the loss function  $L(\hat{x}_{t+1}, x_{t+1})$  with respect to the sampling distribution  $p(x_{t+1} | \theta_{t+1})$ ,

$$R(\hat{x}_{t+1}, \theta_{t+1}) = \sum_{x_{t+1}=0}^N L(\hat{x}_{t+1}, x_{t+1}) p(x_{t+1} | \theta_{t+1}), \quad (22)$$

where  $p(x_{t+1} | \theta_{t+1})$  is from Definition 2.1. □

**Definition 3.3.** Bayes Risk Function

Let  $\mathbf{x}^t = (x_1, x_2, \dots, x_t)$  be observed sequence and  $p(\theta_t | \mathbf{x}^t)$  be the posterior distribution of parameter  $\theta_t$  under  $\mathbf{x}^t$ . Then, the posterior distribution after parameter transitions to nonstationary becomes  $p(\theta_{t+1} | \mathbf{x}^t)$  and the Bayes Risk Function  $BR(\hat{x}_{t+1})$  is defined as,

$$BR(\hat{x}_{t+1}) = \int_0^1 R(\hat{x}_{t+1}, \theta_{t+1}) p(\theta_{t+1} | \mathbf{x}^t) d\theta_{t+1}. \quad (23)$$

□

**Definition 3.4.** Bayes Optimal Prediction

The Bayes optimal prediction  $\hat{x}_{t+1}^*$  is defined as the minimizer of the Bayes risk function,

$$\hat{x}_{t+1}^* = \operatorname{arg\,min}_{\hat{x}_{t+1}} BR(\hat{x}_{t+1}). \quad (24)$$

□

**3.2 Main Theorems**

**Theorem 3.1.** Posterior Distribution after transitions to nonstationary for  $\theta_t$

Let the prior distribution of parameter  $\theta_1$  of the nonstationary geometric distribution in Definition 2.1 be  $\Theta_1 \sim \text{Beta}(\alpha_1, \beta_1)$ . For any  $t \geq 2$ , let  $\mathbf{x}^{t-1} = (x_1, x_2, \dots, x_{t-1})$  be the observed data sequence. Then, the posterior distribution of  $\Theta_t \mid \mathbf{x}^{t-1}$  can be obtained as the following closed form:

$$\Theta_t \mid \mathbf{x}^{t-1} \sim \text{Beta}(\alpha_t, \beta_t), \quad (25)$$

where the parameters  $\alpha_t, \beta_t$  are given as,

$$\begin{cases} \alpha_t = k^{t-1}\alpha_1 + \sum_{i=1}^{t-1} k^i; \\ \beta_t = k^{t-1}\beta_1 + \sum_{i=1}^{t-1} k^i x_{t-i}. \end{cases} \quad (26)$$

□

**Proof of Theorem 3.1.**

For any  $t \geq 2$ , the posterior of parameter distribution  $p(\theta_t \mid \mathbf{x}^{t-1})$  remains in the closed form  $\Theta_t \sim \text{Beta}(\alpha_t, \beta_t)$  if  $X_t \sim \text{Geometric}(\theta_t)$  in Definition 2.1 and  $\Theta_1 \sim \text{Beta}(\alpha_1, \beta_1)$  in Corollary 2.1 according to the nature of *conjugate families* (Bernardo and Smith, 1994, 5.2, p.265) (Berger, 1985, 4.2.2, p.130).

Furthermore, assuming that  $x_{t-1}$  is the observed data,

$$\begin{cases} \alpha_t = \alpha_{t-1} + 1; \\ \beta_t = \beta_{t-1} + x_{t-1}, \end{cases} \quad (27)$$

holds for  $t \geq 2$  by conjugate analysis (Bernardo and Smith, 1994,  $n = 1, r = 1$  for *Negative-Binomial model*, p.437). This is the proof of Eq. (25).

In this paper, nonstationary parameter model is assumed. Therefore, if both Lemma 2.1, and Lemma 2.2 are recursively applied to Eq. (27), then,

$$\begin{cases} \alpha_t = k(\alpha_{t-1} + 1); \\ \beta_t = k(\beta_{t-1} + x_{t-1}), \end{cases} \quad (28)$$

holds for  $t \geq 2$ .

Finally, Eq. (26) is ultimately derived by recursively applying Eq. (28) backwards until the initial conditions  $\alpha_1, \beta_1$  in both Definition 2.5 and Corollary 2.1 are reached.

This completes the proof of Theorem 3.1. □

**Remark 3.1.**

The right hand sides of Eqs. (26) have structures called as *Exponentially Weighted Moving Average* (EWMA) (Harvey, 1989). □

**Theorem 3.2.** Predictive Distribution for  $x_{t+1}$

$$p(x_{t+1} \mid \mathbf{x}^t) = \frac{\alpha_{t+1} \prod_{i=0}^{x_{t+1}-1} (\beta_{t+1} + i)}{\prod_{i=0}^{x_{t+1}} (\alpha_{t+1} + \beta_{t+1} + i)}, \quad (29)$$

where  $\alpha_{t+1}$  and  $\beta_{t+1}$  are formulated in Eq. (26). □

**Proof of Theorem 3.2.**

$$p(x_{t+1} \mid \mathbf{x}^t) = \int_0^1 p(x_{t+1} \mid \theta_{t+1}) p(\theta_{t+1} \mid \mathbf{x}^t) d\theta_{t+1} \quad (30)$$

$$= \frac{\Gamma(\alpha_{t+1} + \beta_{t+1}) \Gamma(\alpha_{t+1} + 1) \Gamma(\beta_{t+1} + x_{t+1})}{\Gamma(\alpha_{t+1}) \Gamma(\beta_{t+1}) \Gamma(\alpha_{t+1} + 1 + \beta_{t+1} + x_{t+1})} \quad (31)$$

$$= \frac{\alpha_{t+1} \prod_{i=0}^{x_{t+1}-1} (\beta_{t+1} + i)}{\prod_{i=0}^{x_{t+1}} (\alpha_{t+1} + \beta_{t+1} + i)}. \quad (32)$$

Note that the second term in Eq. (31) is obtained from the definition of the beta function and that Eq. (32) is obtained from Eq. (31) by applying the following property of gamma function:  $\Gamma(x + 1) = x\Gamma(x)$ .

This completes the proof of Theorem 3.2. □

**Theorem 3.3.** Bayes Optimal Prediction  $\hat{x}_{t+1}^*$

$$\hat{x}_{t+1}^* = \frac{\beta_{t+1}}{\alpha_{t+1} - 1}, \quad (33)$$

where  $\alpha_{t+1}$  and  $\beta_{t+1}$  are formulated in Eq. (26). □

**Proof of Theorem 3.3.**

For parameter estimation problem under the squared-error loss function, the posterior mean is the optimal (Berger, 1985, Result 3 and Example 1, p. 161). For the prediction problem, the predictive mean, i.e. the expectation of the Bayes predictive distribution is identically the optimal under the squared-error loss function. Therefore,

$$\hat{x}_{t+1}^* = E[x_{t+1} \mid \mathbf{x}^t] \quad (34)$$

$$= \sum_{x_{t+1}=0}^N [x_{t+1} p(x_{t+1} \mid \mathbf{x}^t)] \quad (35)$$

$$= \frac{\beta_{t+1}}{\alpha_{t+1} - 1}. \quad (36)$$

Note that Eq. (36) is derived from Eq. (35) by the expectation of the *Negative-Binomial-Beta distribution* (Bernardo and Smith, 1994, p. 429).

This completes the proof of Theorem 3.3. □

### 3.3 Hyperparameter Estimation with Empirical Bayes Method

Since a hyperparameter  $0 < k \leq 1$  in Eqs. (5) and (6) is assumed to be known, it must be estimated in practice. In this paper, the following maximum likelihood estimation in terms of empirical Bayes method (Carlin and Louis, 2000) is considered.

Let  $l(k)$  be a likelihood function of hyperparameter  $k$  and  $\hat{k}$  be the maximum likelihood estimator. Then, those two functions are defined as,

$$\hat{k} = \arg \max_k l(k), \quad (37)$$

$$l(k) = p(x_1 | \theta_1) p(\theta_1) \prod_{i=2}^t p(x_i | \mathbf{x}^{i-1}, k) \quad (38)$$

$$= \prod_{i=1}^t \left[ \frac{\alpha_i \prod_{j=0}^{x_i-1} (\beta_i + j)}{\prod_{j=0}^{x_i} (\alpha_i + \beta_i + j)} \right], \quad (39)$$

where  $\alpha_i$  and  $\beta_i$  are formulated in Eq. (26).

Eq. (39) can not be solved analytically and then the approximate numerical calculation method should be applied. The detail is described in Subsection 4.3.

### 3.4 Proposed Predictive Algorithm

The proposed predictive algorithm that calculates the Bayes optimal prediction  $\hat{x}_{t+1}^*$  in Theorem 3.3 is described as the following Algorithm 3.1.

**Algorithm 3.1.** Proposed Predictive Algorithm.

1. Estimate hyperparameter  $\hat{k}$  by Eq. (39) from training data.
2. Define hyperparameters  $\alpha_1, \beta_1$  for the initial prior  $p(\theta_1 | \alpha_1, \beta_1)$  in Eqs. (9) and (10).
3. Using the observed sequence  $\mathbf{x}^t$  of test data, calculate the Bayes optimal prediction  $\hat{x}_{t+1}^*$  from Eq. (33) where  $\alpha_{t+1}$  and  $\beta_{t+1}$  are formulated in Eq. (26), and  $k$  is replaced by  $\hat{k}$  in step 1. in Eq. (26).

□

## 4 NUMERICAL EXAMPLES

### 4.1 Data Specifications

In order to evaluate the efficacy of the proposed Algorithm 3.1, the real web traffic data is utilized. This data is extracted from the http (Hyper Text Transfer Protocol) request arrival time stamps at three-minute intervals from a web server, spanning a twelve-day period between March 20 and March 31, 2005.

Tables 1 and 2 illustrate a portion of the specifications of both training and test data. Table 1 explains an overview of the training data on March 25, 2005, while Table 2 provides an overview of the test data on March 26, 2005. Figure 1 depicts both plots with line graphs. In figure 1, the vertical axis represents the number of request arrivals, the horizontal axis represents the time interval index, the blue line represents the training data, and the red line represents the test data.

The remaining characteristics of the data are illustrated in Tables 7 and 8 in Appendix C.

Table 1: Training Data Specifications.

Items	Values
Date	Mar. 25, 2005
Time Interval	Every 3 minutes
Total Request Arrivals	11,527
Total Time Intervals $t_{max}$	305
Auto Correlation Coefficient (lag=1)	0.821

Table 2: Test Data Specifications.

Items	Values
Date	Mar. 26, 2005
Time Interval	Every 3 minutes
Total Request Arrivals	6,369
Total Time Intervals $t_{max}$	291
Auto Correlation Coefficient (lag=1)	0.670

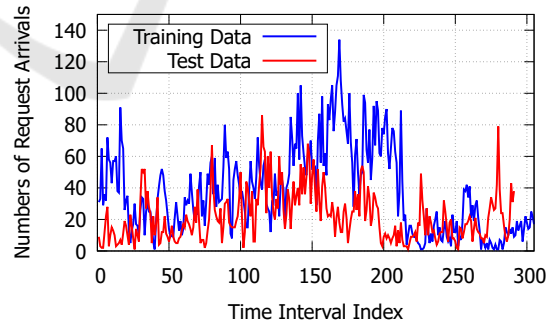


Figure 1: Training and Test Data Plots for Web Traffic Data on Mar. 25–26, 2005.

### 4.2 Conditions and Criteria for Evaluation

The performance of Algorithm 3.1 with the real data is evaluated. Note that the training data is used only for hyperparameter estimation of  $\hat{k}$  in Eq. (39). With the estimated  $\hat{k}$ , the Bayes optimal prediction of  $\hat{x}_{t+1}^*$

is calculated from the test data. For the comparison, two types of prediction  $\hat{x}_t^*$  are considered. The first is from the proposed algorithm with nonstationary geometric distribution in Theorem 3.3 and the second is from a conventional algorithm with stationary geometric distribution.

#### 4.2.1 Initial Prior Distribution of Parameter

According to Corollary 2.1, the class of the initial prior distribution of parameter is beta distribution. If the non-informative prior (Berger, 1985; Bernardo and Smith, 1994) is considered under beta prior  $p(\theta_1 | \alpha_1, \beta_1)$ , it should correspond to the uniform distribution and each of two hyperparameters of  $\alpha_1$  and  $\beta_1$  equals to one. Their settings are shown in Table 3.

Table 3: Defined Hyperparameters for Prior Distribution  $p(\theta_1 | \alpha_1, \beta_1)$ .

Items	$\alpha_1$	$\beta_1$
Values	1	1

#### 4.2.2 Criteria

For the criteria for evaluations, the following mean squared error based on the squared-error loss function in Definition 3.1 is defined.

**Definition 4.1.** Mean Squared Error

$$\frac{1}{t_{max}} \sum_{t=1}^{t_{max}} L(\hat{x}_t, x_t) = \frac{1}{t_{max}} \sum_{t=1}^{t_{max}} (\hat{x}_t - x_t)^2. \quad (40)$$

□

### 4.3 Results

Table 4 presents the estimated hyperparameter  $\hat{k}$  from the training data. Figure 2 depicts the loglikelihood function  $\log l(k)$  with base of  $10^3$ , as calculated numerically using R version 4.4.1 (R Core Team, 2024).

Table 9 in Appendix C illustrates the extended results of the hyperparameter estimation of the training data.

Table 5 illustrates the values of predictive mean squared errors for both the proposed and stationary models in Definition 4.1. Figure 3 depicts its graphical result. In Figure 3, the horizontal and vertical axes are the index of time interval  $1 \leq t \leq 291$  and the number of request arrivals, respectively. Furthermore, the orange bar is real request arrivals  $x_t$  from test data, the blue solid line is the predictions  $\hat{x}_t^*$  from the proposed nonstationary geometric model, the red solid line is the predictions from the stationary geometric model.

The second column from Left in Table 10 in Appendix C illustrates the extended results of mean squared errors for both the proposed and stationary models from the test data.

Table 4: Hyperparameter Estimation from Training Data.

Item	$\hat{k}$
Value	0.889

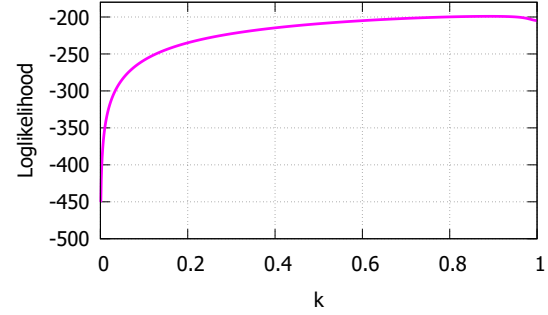


Figure 2:  $\log l(k)$  Plot for  $0 \leq k \leq 1$ .

Table 5: Mean Squared Error for Two Models.

Items	MSE	
	Proposed	Stationary
Values	186.8	254.1

## 5 DISCUSSIONS

From Table 4, the estimator is  $\hat{k} = 0.889$ . If  $k = 1$ , then the second parameters of beta distributions in Eqs. (5),(6),(11), and (12) become zero. This means that the variances of beta distributions are also zero, and that the parameter  $\theta_t$  of geometric distribution is *stationary*. Therefore  $\hat{k} = 0.889 \neq 1.000$  means that training data is *nonstationary*. Furthermore, Figure 2 show that the likelihood function  $l(k)$  is empirically upward convex. Hence the estimated value for  $\hat{k}$  can be considered reliable as an estimator.

Table 5 shows that the performance of prediction with the proposed model is superior to that of the stationary model. The improvement of MSE is more than 25%. In fact, Figure 3 also shows that the blue line follows the orange bar better than the red line. Similarly, Figure 4 compares the expectation values of the posterior parameter distributions  $E[\theta_t | \mathbf{x}^{t-1}]$  between two models. In Figure 4, the blue line of the proposed model shows more intense dynamic fluctuations than the red line of the stationary model. These results suggest that the proposed empirical Bayesian method with the nonstationary geo-

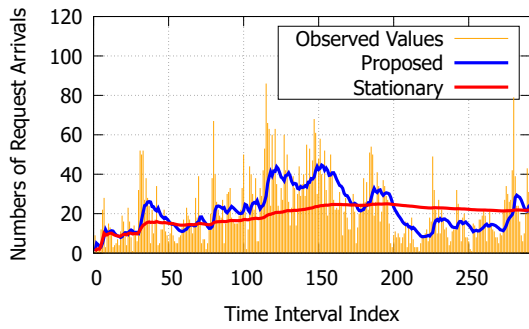


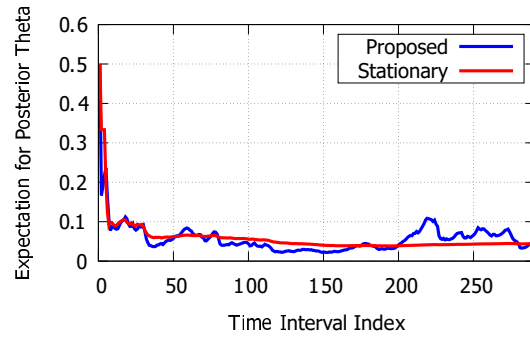
Figure 3: Prediction Result for Test Data.

ometric model works to some extent compared to the stationary model. However, Figure 3 also shows that the blue line did not necessarily follow the rapid increase or decrease in the orange bars. If one sets the hyperparameter  $k = 0.50$ , then the nonstationarity of parameter  $\theta_t$  becomes larger and the MSE of the proposed model becomes 148.5. In this case, the improvement is about 40% better than that of the stationary model. This desirable situation did not occur because the training and test data were very different as shown in Figure 1.

Finally, the values of Akaike Information Criteria (AIC) (Akaike, 1973) between two models in terms of the model selection were calculated and shown in Table 6. Note that the value of AIC is calculated by  $(-2\log_2 L + 2m)$ , where  $L$  and  $m$  represent the likelihood value and the number of parameters in the statistical model, respectively. In this paper,  $m = 4$  for the proposed nonstationary model ( $\theta_t, \alpha_t, \beta_t$ , and  $k$ ) and  $m = 3$  for the stationary model ( $\theta_t, \alpha_t$ , and  $\beta_t$ ). Therefore, the parametric penalty in AIC for the proposed model did not become so larger than that of the stationary model. As a result, the value of AIC in Table 6 for the proposed model is slightly smaller than that of the stationary model. It means that the proposed model is relatively suitable than the stationary model. Furthermore, the third and fourth columns of Table 9 in Appendix C present the extended results of the AIC values for the training data. The overall results indicate that the the proposed nonstationary geometric model is comparatively more suitable than the stationary geometric model based on the observed data in terms of the model selection.

Table 6: Akaike Information Criteria (AIC) for Two Models.

Items	AIC	
	Proposed	Stationary
Values	3976.4	4090.1

Figure 4:  $E[\theta_t | \mathbf{x}^{t-1}]$  for the Posterior Distribution from Test Data.

## 6 CONCLUSION

In this paper, a special class of nonstationary geometric distributions has been proposed. It has been proved that the Bayes optimal prediction related by the nonstationary geometric distribution and squared-error loss function can be obtained by the simple arithmetic calculations if its nonstationary hyperparameter is known. Using the real web traffic data, the predictive performance of the proposed algorithm was shown to be superior to that of the stationary algorithm in terms of both model selection theory and predictive mean squared error.

For the nonstationary hyperparameter estimation, the approximate maximum likelihood estimation is considered. It has been observed that the likelihood function for the hyperparameter is empirically upward convex for certain data. The theoretical proof of the general convexity should be an open problem. Moreover, the generalization for the nonstationary negative binomial distribution and the other Bayes optimal predictive algorithms with loss functions other than the squared-error loss function should also be considered for future research.

## REFERENCES

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 267–281, Budapest.
- Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.
- Bernardo, J. M. and Smith, A. F. (1994). *Bayesian Theory*. John Wiley & Sons, Chichester.
- Bianchi, G. (2000). Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE Journal on Selected Areas in Communications*, 18(3):535–547.

- Carlin, B. and Louis, T. (2000). *Bayes and Empirical Bayes Methods for Data Analysis (Second Edition)*. Chapman & Hall, New York.
- Ewens, W. J. (2004). *Mathematical Population Genetics I. Theoretical Introduction*. Springer, New York.
- Frank C. Kaminsky, James C. Bennenyan, R. D. D. and Burke, R. J. (1992). Statistical control charts based on a geometric distribution. *Journal of Quality Technology*, 24(2):63–69.
- G.Gallager, R. (1995). *Discrete Stochastic Processes*. Kluwer Academic Publishers, Boston.
- Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Marsa, Malta.
- Hogg, R. V., McKean, J. W., and Craig, A. T. (2013). *Introduction to Mathematical Statistics (Seventh Edition)*. Pearson Education, Boston.
- Johnson, N. L. and Kotz, S. (1969). *Discrete Distributions*. John Wiley & Sons, New York.
- Koizumi, D. (2020). Credible interval prediction of a nonstationary poisson distribution based on bayes decision theory. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, pages 995–1002. INSTICC, SciTePress.
- Koizumi, D. (2021). On the prediction of a nonstationary bernoulli distribution based on bayes decision theory. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, pages 957–965. INSTICC, SciTePress.
- Koizumi, D. (2023). On the prediction of a nonstationary exponential distribution based on bayes decision theory. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART*, pages 193–201. INSTICC, SciTePress.
- O. Diekmann, J. (2000). *Mathematical epidemiology of infectious diseases : model building, analysis and interpretation*. John Wiley & Sons, New York.
- R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Smith, J. Q. (1979). A generalization of the bayesian steady forecasting model. *Journal of the Royal Statistical Society - Series B*, 41:375–387.
- Souza, R. C. (1978). A bayesian entropy approach to forecasting. PhD thesis, University of Warwick.
- Winsten, C. B. (1959). Geometric Distributions in the Theory of Queues. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(1):1–22.

## APPENDIX

### A: Proof of Lemma 2.1

If  $t = 1$ , suppose  $A_1 = a_1$  and  $C_1 = c_1$  are defined as,

$$A_1 \sim \text{Gamma}(\alpha_1, 1), \quad (41)$$

$$C_1 \sim \text{Beta}[k\alpha_1, (1-k)\alpha_1], \quad (42)$$

according to Definition 2.5 and Definition 2.6, respectively.

Since  $A_2 = C_1 A_1$  from Definition 2.3, and  $A_t$  and  $C_t$  are conditional independent from Definition 2.4, the joint distribution of  $p(c_1, a_1)$  becomes,

$$\begin{aligned} p(c_1, a_1) &= p[c_1 | k\alpha_1, (1-k)\alpha_1] p(a_1 | \alpha_1, 1) \\ &= \frac{\Gamma(\alpha_1)}{\Gamma(k\alpha_1)\Gamma[(1-k)\alpha_1]} c_1^{k\alpha_1-1} (1-c_1)^{(1-k)\alpha_1-1} \\ &\quad \cdot \frac{a_1^{\alpha_1-1}}{\Gamma(\alpha_1)} \exp(-a_1) \\ &= \frac{c_1^{k\alpha_1-1} (1-c_1)^{(1-k)\alpha_1-1}}{\Gamma(k\alpha_1)\Gamma[(1-k)\alpha_1]} a_1^{\alpha_1-1} \exp(-a_1). \end{aligned}$$

Now, denote the two transformation as,

$$\begin{cases} v = a_1 c_1 \\ w = a_1 (1 - c_1), \end{cases} \quad (43)$$

where  $0 < v, 0 < w$ .

Then, the inverse transformation of Eq. (43) becomes,

$$\begin{cases} a_1 = \frac{v+w}{v+w} \\ c_1 = \frac{v}{v+w}, \end{cases} \quad (44)$$

The Jacobian  $J_1$  of Eq. (44) is,

$$\begin{aligned} J_1 &= \begin{vmatrix} \frac{\partial a_1}{\partial v} & \frac{\partial a_1}{\partial w} \\ \frac{\partial c_1}{\partial v} & \frac{\partial c_1}{\partial w} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ \frac{w}{(v+w)^2} & -\frac{v}{(v+w)^2} \end{vmatrix} \\ &= -\frac{1}{v+w} = -\frac{1}{a_1} \neq 0. \end{aligned}$$

Then, the transformed joint distribution  $p(v, w)$  is obtained by the product of  $p(c_1, a_1)$  and the absolute value of  $J_1$ .

$$\begin{aligned} p(v, w) &= p(c_1, a_1) \left| -\frac{1}{a_1} \right| \\ &= \frac{\left(\frac{v}{v+w}\right)^{k\alpha_1-1} \left(\frac{w}{v+w}\right)^{(1-k)\alpha_1-1}}{\Gamma(k\alpha_1)\Gamma[(1-k)\alpha_1]} \\ &\quad \cdot (v+w)^{\alpha_1-1} \exp[-(v+w)] \cdot \frac{1}{v+w} \\ &= \frac{v^{k\alpha_1-1} w^{(1-k)\alpha_1-1}}{\Gamma(k\alpha_1)\Gamma[(1-k)\alpha_1]} \exp[-(v+w)]. \end{aligned} \quad (45)$$

Then,  $p(v)$  is obtained by marginalizing Eq. (45)



with respect to  $w$ ,

$$\begin{aligned}
p(v) &= \int_0^\infty p(v, w) dw \\
&= \frac{v^{k\alpha_1-1} \exp(-v)}{\Gamma(k\alpha_1) \Gamma[(1-k)\alpha_1]} \\
&\quad \cdot \int_0^\infty w^{(1-k)\alpha_1-1} \exp(-w) dw \\
&= \frac{v^{k\alpha_1-1} \exp(-v)}{\Gamma(k\alpha_1) \Gamma[(1-k)\alpha_1]} \cdot \Gamma[(1-k)\alpha_1] \\
&= \frac{1}{\Gamma(k\alpha_1)} v^{k\alpha_1-1} \exp(-v). \quad (46)
\end{aligned}$$

Eq. (46) exactly corresponds to  $Gamma(k\alpha_1, 1)$  according to Definition 2.7. Recalling  $v = a_1 c_1$  from Eq. (43) and  $A_2 = C_1 A_1$  from Definition 2.3,

$$A_2 \sim Gamma(k\alpha_1, 1).$$

Thus if  $t = 1$ , then  $A_{t+1} \sim Gamma(k\alpha_t, 1)$  holds.

For  $t \geq 2$ , by substituting  $\alpha_t = k\alpha_{t-1}$ ,  $A_t = a_t$  and  $C_t = c_t$  are defined as,

$$A_t \sim Gamma(\alpha_t, 1), \quad (47)$$

$$C_t \sim Beta[k\alpha_t, (1-k)\alpha_t]. \quad (48)$$

Eqs. (47) and (48) correspond to Eqs. (41) and (42), respectively. Therefore the same proof can be applied for the case of  $t \geq 2$  and it can be proved that,

$$\forall t, A_{t+1} \sim Gamma(k\alpha_t, 1).$$

This completes the proof of Lemma 2.1.  $\square$

## B: Proof of Lemma 2.3

From Lemma 2.1 and 2.2,

$$\forall t \geq 2, A_t \sim Gamma(k\alpha_{t-1}, 1),$$

$$\forall t \geq 2, B_t \sim Gamma(k\beta_{t-1}, 1).$$

According to Definition 2.2, two random variables  $A_t$  and  $B_t$  are mutually independent. Therefore, the joint distribution pf  $p(a_t, b_t)$  becomes,

$$\begin{aligned}
p(a_t, b_t) &= p(a_t | k\alpha_{t-1}, 1) p(b_t | k\beta_{t-1}, 1) \\
&= \left[ \frac{a_t^{k\alpha_{t-1}-1} \exp(-a_t)}{\Gamma(k\alpha_{t-1})} \right] \cdot \left[ \frac{b_t^{k\beta_{t-1}-1} \exp(-b_t)}{\Gamma(k\beta_{t-1})} \right] \\
&= \frac{a_t^{k\alpha_{t-1}-1} b_t^{k\beta_{t-1}-1}}{\Gamma(k\alpha_{t-1}) \Gamma(k\beta_{t-1})} \exp[-(a_t + b_t)].
\end{aligned}$$

Denoting the two transformations,

$$\begin{cases} \lambda = a_t + b_t \\ \mu = \frac{a_t}{a_t + b_t}, \end{cases} \quad (49)$$

where  $0 < \lambda, 0 < \mu$ .

The inverse transformation of Eq. (49) becomes,

$$\begin{cases} a_t = \lambda \mu \\ b_t = \lambda (1 - \mu). \end{cases} \quad (50)$$

Then, the Jacobian  $J_2$  of Eq. (50) is,

$$\begin{aligned}
J_2 &= \begin{vmatrix} \frac{\partial a_t}{\partial \lambda} & \frac{\partial a_t}{\partial \mu} \\ \frac{\partial b_t}{\partial \lambda} & \frac{\partial b_t}{\partial \mu} \end{vmatrix} = \begin{vmatrix} \mu & \lambda \\ 1 - \mu & -\lambda \end{vmatrix} \\
&= -\lambda = -(a_t + b_t).
\end{aligned}$$

Then, the transformed joint distribution  $p(\lambda, \mu)$  is obtained by the product of  $p(a_t, b_t)$  and the absolute value of  $J_2$  as the following,

$$\begin{aligned}
p(\lambda, \mu) &= p(a_t, b_t) \cdot |-(a_t + b_t)| \\
&= \frac{(\lambda \mu)^{k\alpha_{t-1}-1} [\lambda (1 - \mu)]^{k\beta_{t-1}-1}}{\Gamma(k\alpha_{t-1}) \Gamma(k\beta_{t-1})} \exp(-\lambda) \cdot \lambda \\
&= \frac{\mu^{k\alpha_{t-1}-1} (1 - \mu)^{k\beta_{t-1}-1}}{\Gamma(k\alpha_{t-1}) \Gamma(k\beta_{t-1})} \lambda^{k\alpha_{t-1} + k\beta_{t-1} - 1} \exp(-\lambda). \quad (51)
\end{aligned}$$

Then,  $p(\mu)$  is obtained by marginalizing Eq. (51) with respect to  $\lambda$ ,

$$\begin{aligned}
p(\mu) &= \int_0^\infty p(\lambda, \mu) d\lambda \\
&= \frac{\mu^{k\alpha_{t-1}-1} (1 - \mu)^{k\beta_{t-1}-1}}{\Gamma(k\alpha_{t-1}) \Gamma(k\beta_{t-1})} \\
&\quad \cdot \int_0^\infty \lambda^{k\alpha_{t-1} + k\beta_{t-1} - 1} \exp(-\lambda) d\lambda \\
&= \frac{\mu^{k\alpha_{t-1}-1} (1 - \mu)^{k\beta_{t-1}-1}}{\Gamma(k\alpha_{t-1}) \Gamma(k\beta_{t-1})} \cdot \Gamma(k\alpha_{t-1} + k\beta_{t-1}) \\
&= \frac{\Gamma(k\alpha_{t-1} + k\beta_{t-1})}{\Gamma(k\alpha_{t-1}) \Gamma(k\beta_{t-1})} \mu^{k\alpha_{t-1}-1} (1 - \mu)^{k\beta_{t-1}-1}. \quad (52)
\end{aligned}$$

Eq. (52) exactly corresponds to  $Beta(k\alpha_{t-1}, k\beta_{t-1})$  according to Definition 2.8.

Recalling  $\mu = \frac{a_t}{a_t + b_t}$  from Eq. (49) and  $\Theta_t = \frac{A_t}{A_t + B_t}$  from Definition 2.2,

$$\forall t \geq 2, \Theta_t \sim Beta(k\alpha_{t-1}, k\beta_{t-1}),$$

holds.

This completes the proof of Lemma 2.3.  $\square$

### C: Extended Results of Numerical Examples

The following five tables are presented in Appendix C. Tables 7 and 8 illustrate the specifications of training data, which encompasses a twelve-day period between March 20 and March 31, 2005. Table 7 illustrates the request arrivals and the total time intervals ( $t_{max}$ ). Table 8 illustrates the auto correlation coefficients with lag=1. Tables 9 presents the estimated values of the hyperparameter  $\hat{k}$ , obtained through the approximate maximum likelihood estimations, along with the values of Akaike Information Criteria (AIC) for both the proposed nonstationary and the conventional stationary models, derived from the training data with a eleven-day period. Table 10 illustrates the predictive performances of both the proposed and stationary models, evaluated on the test data. It depicts the mean squared errors under both the past observed test data and the estimated values of the hyperparameters from the training data.

Table 7: Extended Training Data Specifications #1.

Date	Total Request	
	Arrivals	$t_{max}$
Mar.20, 2005	4,669	279
Mar.21, 2005	6,742	312
Mar.22, 2005	9,767	329
Mar.23, 2005	11,672	333
Mar.24, 2005	17,329	332
Mar.25, 2005	11,527	305
Mar.26, 2005	6,369	291
Mar.27, 2005	26,325	314
Mar.28, 2005	17,994	341
Mar.29, 2005	17,874	336
Mar.30, 2005	7,267	295
Mar.31, 2005	11,260	329

Table 8: Extended Training Data Specifications #2.

Date	Auto Correlation Coefficient
Mar.20, 2005	0.615
Mar.21, 2005	0.667
Mar.22, 2005	0.708
Mar.23, 2005	0.782
Mar.24, 2005	0.864
Mar.25, 2005	0.821
Mar.26, 2005	0.670
Mar.27, 2005	0.839
Mar.28, 2005	0.873
Mar.29, 2005	0.862
Mar.30, 2005	0.712
Mar.31, 2005	0.714

Table 9: Extended Results of Training Data.

Date	$\hat{k}$	AIC	
		Proposed	Stationary
Mar.20, 2005	0.932	3086.2	3098.4
Mar.21, 2005	0.936	3663.1	3693.9
Mar.22, 2005	0.924	4155.0	4191.4
Mar.23, 2005	0.921	4333.2	4392.7
Mar.24, 2005	0.918	4643.6	4758.4
Mar.25, 2005	0.889	3976.4	4090.1
Mar.26, 2005	0.924	3417.2	3456.7
Mar.27, 2005	0.937	4862.5	4927.6
Mar.28, 2005	0.900	4788.0	4904.0
Mar.29, 2005	0.911	4692.5	4832.2
Mar.30, 2005	0.923	3547.3	3601.5
Mar.31, 2005	0.928	4290.0	4320.9

Table 10: Mean Squared Errors of Test Data.

Date	MSE		
	Proposed(A)	Stationary(B)	$\frac{A}{B}$
Mar.21, 2005	142.9	192.0	0.744
Mar.22, 2005	289.5	376.7	0.768
Mar.23, 2005	288.3	499.1	0.578
Mar.24, 2005	492.8	1263.0	0.390
Mar.25, 2005	365.8	756.2	0.484
Mar.26, 2005	186.8	254.1	0.735
Mar.27, 2005	1083.1	2261.5	0.479
Mar.28, 2005	761.7	1504.1	0.506
Mar.29, 2005	631.4	1491.3	0.423
Mar.30, 2005	201.9	302.8	0.667
Mar.31, 2005	338.7	482.4	0.702