stats

MDPI

Brief Report

Statistical Framework: Estimating the Cumulative Shares of Nobel Prizes from 1901 to 2022

Xu Zhang [1,*], Bruce Golden [2] and Edward Wasil [3]

[1] Department of Mathematics, University of Maryland, College Park, MD 20742, USA
[2] Robert H. Smith School of Business, University of Maryland, College Park, MD 20742, USA; bgolden@umd.edu
[3] Kogod School of Business, American University, Washington, DC 20016, USA; ewasil@american.edu
* Correspondence: xuzhang306@gmail.com

Abstract: Studying trends in the geographical distribution of the Nobel Prize is an interesting topic that has been examined in the academic literature. To track the trends, we develop a stochastic estimate for the cumulative shares of Nobel Prizes awarded to recipients in four geographical groups: North America, Europe, Asia, Other. Specifically, we propose two models to estimate how cumulative shares change over time in the four groups. We estimate parameters, develop a prediction interval for each model, and validate our models. Finally, we apply our approach to estimate the distribution of the cumulative shares of Nobel Prizes for the four groups from 1901 to 2022.

Keywords: Nobel Prizes; cumulative share; log–log transformation; prediction interval; estimation

1. Introduction

Nobel Prizes are awarded, according to Alfred Nobel's will, to "those who, during the preceding year, have conferred the greatest benefit to humankind" [1]. They were first awarded in 1901 in physics, chemistry, medicine, literature, and peace. The economics prize was awarded for the first time in 1969. The Nobel Prize is widely considered the most prestigious award in these fields.

Since 1901, Nobel Prizes have been awarded in almost every year. Due to its worldwide attention, studying and forecasting trends in the geographical distribution of the Nobel Prize has received attention in the academic literature. A simple logistic function [2] provided an excellent fit for the number of Nobel Prize recipients in the United States from 1901 to 1987, and forecasted a declining number of U.S. recipients in the future. In this paper [2], the logistic function that described the cumulative number of Nobel Prizes, $P(t)$, by time t is given by

$$P(t) = \frac{M}{1 + exp(-\alpha(t - t_0))},$$

where α, t_0, and M are parameters that must be estimated. According to the results of this model, the U.S. was forecasted to receive approximately 235 prizes by the end of 2002. This prediction was off. In fact, the U.S. received 270 prizes by the end of 2002 [3].

Two papers [3,4] showed that Europe's share of Nobel Prizes would continue to decline while North America's share of prizes would continue to increase in the early 2000s. One paper [3] used a nonlinear least-squares model with a logistic function to show that the number of Nobel Prizes awarded to the U.S. was underpredicted by the logistic function model from [2]. Specifically, the authors used the Gauss–Newton method of Hartley [5] to perform the nonlinear least-squares minimization along with other optimization methods, including the Levenberg–Marquardt algorithm [6,7] with different starting values. Different starting values were specified to check the accuracy of the solution. A second paper [4] used a polynomial smoothing spline analysis to show the rise of prizes in North America and the fall of prizes in Europe since the 1930s. In this paper, the smoothing parameter λ

Citation: Zhang, X.; Golden, B.; Wasil, E. Statistical Framework: Estimating the Cumulative Shares of Nobel Prizes from 1901 to 2022. *Stats* **2024**, *7*, 95–109. https://doi.org/10.3390/stats7010007

Academic Editor: Wei Zhu

Received: 6 December 2023
Revised: 5 January 2024
Accepted: 16 January 2024
Published: 19 January 2024

and coefficients in the polynomial smoothing spline model $f(t)$ by time t were estimated to minimize the penalized residual sum of squares

$$\lambda \int \{D^m f(t)\}^2 dt + \sum_{t=1901}^{2003} \{y(t) - f(t)\}^2$$

where $D^m f(t)$ is the mth derivative of the function $f(t)$ with respect to t and $y(t)$ is the cumulative share. In the minimization, the smoothing parameter λ controlled the trade-off between the data fit and the variability of the polynomial smoothing spline model $f(t)$. More recently, a third paper [8] proposed a Volterra–Lotka model to fit data from 1901 to 2009 with two groups (United States and other nationalities denoted by X and Y, respectively)

$$\begin{cases} X' = a_x X - b_x X^2 + c_{xy} XY \\ Y' = a_y Y - b_y Y^2 + c_{yx} XY, \end{cases}$$

where a_x, b_x, c_{xy}, a_y, b_y, c_{yx} are parameters in the equations. The equations model the derivatives of the annual data with respect to the United States and other nationalities, that is, X' and Y', with logistic growth equations and cross terms, which are XY. As the author notes, the parameters, c_{xy} and c_{yx}, decide how one's rate of growth depends on the presence of the other, which describes any possible interaction between X and Y. The results of this model revealed an interesting insight: The U.S. and other nationalities were locked in an even-money competition where each side was expected to win about half of all future Nobel Prizes. The author suggested updating the data in a couple of decades with more than two groups.

The models used in all four papers only considered deterministic scenarios. It might be helpful to take stochastic estimates into consideration in order to allow variations in the trend and forecast. In general, the models only considered two geographical groups. It would be more realistic to consider the interactions of more than two groups. In this paper, we consider four geographical groups: North America, Europe, Asia, and Other (countries not included in the first three groups). Figure 1 displays the number of Nobel Prizes awarded to each geographical group from 1901 to 2022. In terms of the number of Nobel Prizes based on Figure 1, it is difficult to distinguish the trend among the groups. Therefore, we use an alternative metric, the cumulative share of prizes, as suggested in [4].

In this study, our goal is to estimate the distribution of the cumulative shares under a stochastic scenario using Nobel Prize data from 1901 to 2022. Our estimation problem has several challenging characteristics. First, we need to define an appropriate distribution of cumulative shares. In particular, we would like to take four geographical groups into consideration. The second challenge is to identify a stochastic metric to estimate the trend and forecast. Moreover, we need to develop an appropriate distribution in order to estimate the stochastic metric. To overcome these challenges, we use multiple distributions for the cumulative shares and prediction interval to show where the future cumulative shares will fall. We use a log–log transformation so that the estimate for the cumulative share is contained in the interval $[0, 1]$.

The remaining sections of this paper are organized as follows. In Section 2, we formulate our problem mathematically and present our models. We describe our methodology for estimating parameters and developing a prediction interval using a log–log transformation. In Section 3, we validate the performance of our estimation approach using simulation. In Section 4, we use Nobel Prize data from 1901 to 2022 to test our estimation approach for the cumulative shares of the four geographical regions. In Section 5, we conclude with a discussion of our results and directions for future research.

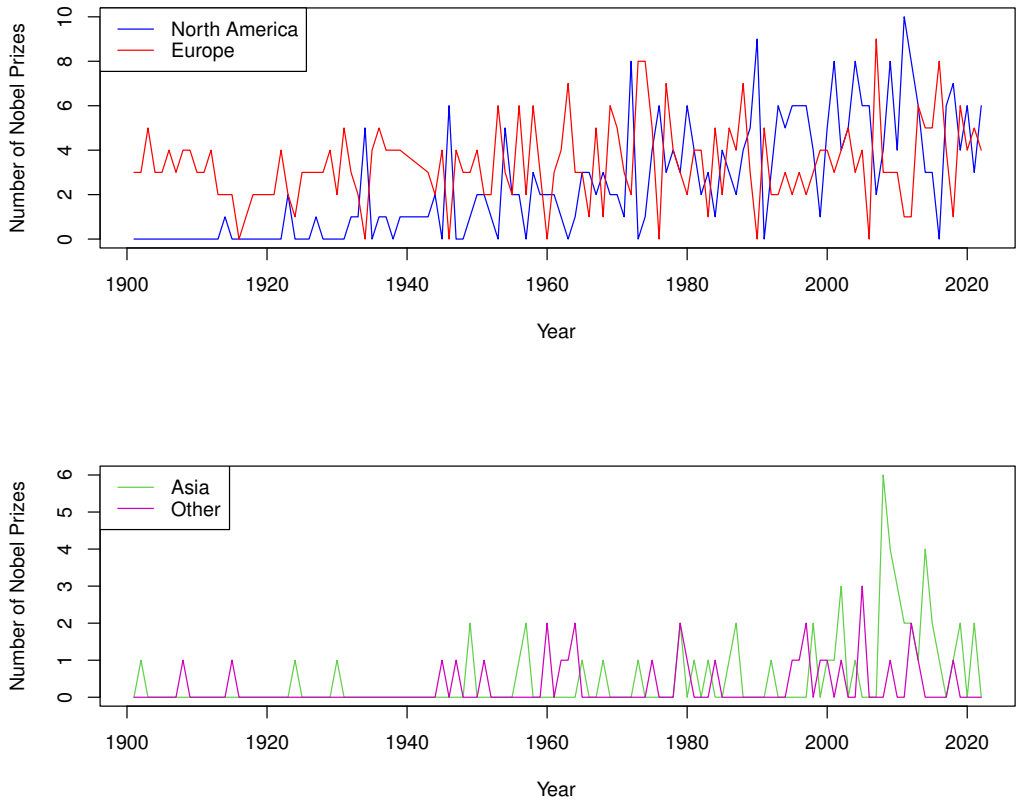

Figure 1. The number of Nobel Prizes awarded to each geographical group from 1901 to 2022.

2. Methods

In this section, we introduce two ways of modeling the cumulative shares of Nobel Prizes. To quantify the cumulative shares, we let $N(t)$ be the total number of Nobel Prizes awarded by time t, where $1901 \leq t \leq 2022$. We let $N_i(t)$ denote the cumulative number of Nobel Prizes awarded to group i by time t. Thus, the cumulative share of Nobel Prizes by time t for group i is given by

$$f_i(t) = N_i(t)/N(t).$$

The data used for this paper are taken from the Nobel Prize website at https://www.nobelprize.org/ (accessed on 6 December 2023) for the period 1901 to 2022. We categorize the data into four geographical groups: North America, Europe, Asia, and Other. In Figure 2, we display the cumulative share for each geographical group from 1901 to 2022.

To account for variability in estimation at time t, we assume a normal distribution with mean $f_i^*(t)$ and variance σ_i^2 for the purpose of simplicity, that is, $f_i(t) \sim N(f_i^*(t), \sigma_i^2)$, $i = 1, 2, 3, 4$. Estimating parameters for the normal distribution is essential in fitting cumulative shares. We provide point estimates of parameters using least-squares estimation (LSE). We also provide parameter estimates using a prediction interval for a cumulative share through maximum likelihood estimation (MLE). We use a log–log transformation when the upper or lower bound in a prediction interval for a cumulative share is outside $[0, 1]$.

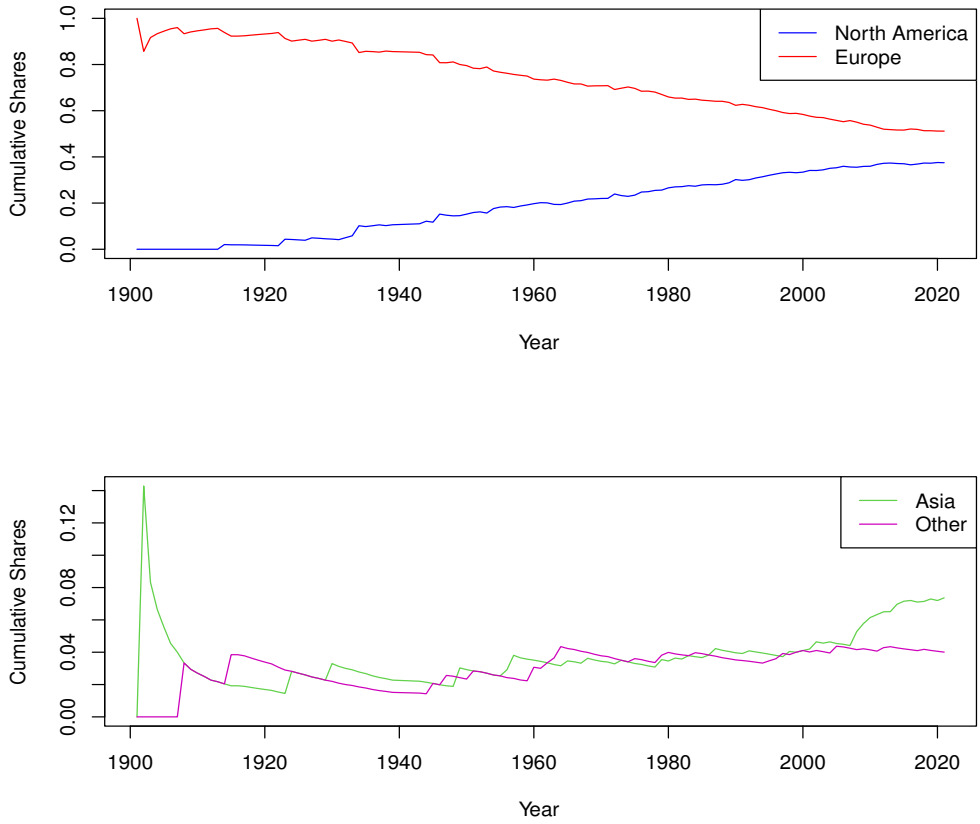

Figure 2. The cumulative share for each geographical group from 1901 to 2022.

2.1. Independent and Competitive Models

2.1.1. Independent Model

In the independent model (denoted by a), we set $f_i^*(t) = f_i^a(t)$ where

$$f_i^a(t) = \frac{\alpha_i}{1 + e^{-\beta_i(t-\gamma_i)}}, i = 1, 2, 3, 4.$$

In addition, we assume logistic growth as a mean. We assume that the cumulative share for a group does not impact the other three groups (that is, each group is independent). Therefore, when we estimate the three parameters $(\alpha_i, \beta_i, \gamma_i)$ for this model using LSE or MLE, we do so for each group independently.

2.1.2. Competitive Model

In the competitive model (denoted by b), we set $f_i^*(t) = f_i^b(t)$ where

$$[f_i^b(t)]' = f_i^b(t)[a_i - \sum_{j=1}^{4} a_j f_j^b(t)], \quad i = 1, 2, 3, 4; \quad \sum_{i=1}^{4} f_i^b(t) = 1.$$

We note that

$$[f_i^b(t)]' = \underbrace{a_i f_i^b(t) - a_i [f_i^b(t)]^2}_{\text{logistic growth}} - \underbrace{\sum_{j \neq i} a_j f_i^b(t) f_j^b(t)}_{\text{interaction with other groups}}.$$

Thus,

$$[f_i^b(t)]' = f_i^b(t)[b_i - \sum_{j=1}^{3} b_j f_j^b(t)]$$

where $b_i = a_i - a_4$. After we solve this equation, we have

$$f_i^b(t) = \frac{e^{F_i(t)}}{1 + \sum_{j=1}^{3} e^{F_j(t)}}$$

where $F_i(t) = b_i t + c_i = (a_i - a_4)t + c_i$. Therefore,

$$f_i^b(t) = \begin{cases} \frac{e^{b_i t + c_i}}{1 + \sum_{j=1}^{3} e^{b_j t + c_j}}, & i = 1, 2, 3 \\ \frac{1}{1 + \sum_{j=1}^{3} e^{b_j t + c_j}}, & i = 4 \end{cases}$$

where $c_i = \log(\frac{f_i(0)}{1 - \sum_{j=1}^{3} f_j(0)})$ and $i = 1, 2, 3$. The details on how to solve for $f_i^b(t)$ and c_i are presented in Appendix A and B, respectively.

In the competitive model, the mean is motivated by the Lotka–Volterra models, and all cumulative shares compete with each other, since the summation of cumulative shares is equal to 1 [9]. Therefore, we use cumulative shares from all four groups to estimate the parameters. There are fewer parameters to estimate in this model compared to the independent model. The competitive model takes both logistic growth and interaction with other groups into consideration.

2.2. Estimation

2.2.1. Least-Squares Estimation

In the independent model, we denote the parameters by $\theta_i^{a_1} = \{\alpha_i, \beta_i, \gamma_i\}$. The objective of the LSE is to minimize the sum of squares of the residuals, which is the difference between an observed cumulative share and the fitted value generated by a model. We have

$$\hat{\theta}_i^{a_1} = \underset{\theta_i^{a_1}}{\operatorname{argmin}} \sum_k [f_i(t_k) - f^a(t_k)]^2.$$

In the competitive model, we denote the parameters by $\theta^{b_1} = \{b_1, b_2, b_3\}$. We have

$$\hat{\theta}^{b_1} = \underset{\theta^{b_1}}{\operatorname{argmin}} \sum_k \sum_{i=1}^{4} [f_i(t_k) - f_i^b(t_k)]^2.$$

2.2.2. Maximum Likelihood Estimation

Maximum likelihood estimation has excellent performance among all estimation methods. MLE provides a prediction interval by taking variance into account. In MLE, the contribution to the likelihood L is the product of the density for $f_i(t)$ over all years. The log-likelihood is then maximized over parameters $\theta_i^{a_2} = \{\alpha_i, \beta_i, \gamma_i, \sigma_i\}$ for the independent model and parameters $\theta^{b_2} = \{b_1, b_2, b_3, \sigma\}$ for the competitive model. We have

$$\hat{\theta}_i^{a_2} = \underset{\theta_i^{a_2}}{\operatorname{argmax}}\, l \quad \text{and} \quad \hat{\theta}^{b_2} = \underset{\theta^{b_2}}{\operatorname{argmax}}\, l.$$

In the independent model, the log-likelihood function is

$$l = \log L \propto -\frac{1}{2\sigma^2} \sum_{k=1}^{n} \left(f_i(t_k) - f_i^a(t_k) \right)^2 - n\log(\sigma).$$

In the competitive model, the log-likelihood function is

$$l = \log L \propto -\frac{1}{2\sigma^2} \sum_{k=1}^{n} \sum_{i=1}^{4} \left(f_i(t_k) - f_i^b(t_k) \right)^2 - 4n\log(\sigma).$$

2.2.3. Prediction Interval

After estimating σ using MLE, we are able to produce a prediction interval for cumulative shares based on normality. A 100 $(1-\alpha)$% prediction interval for $f_i(t)$ with $\hat{\sigma}$ is given by

$$\left(\hat{f}_i^*(t) - z_{\alpha/2}\sigma_i, \hat{f}_i^*(t) + z_{\alpha/2}\sigma_i\right).$$

We note that the upper or lower bound of the prediction interval may fall out of $[0,1]$. Transforming $f_i(t)$ onto a $(-\infty,\infty)$ scale is one possible solution. For example, we could use a log–log transformation given by

$$g_i(t) = \log[-\log(f_i(t))].$$

Applying the Delta method [10], we have

$$Var(g_i(t)) = \sigma[\frac{1}{f_i^*(t)\log(f_i^*(t))}]^2.$$

The Delta method helps us derive an asymptotically normal distribution of a function of a random variable. Therefore, the distribution of $g_i(t)$ is given by

$$g_i(t) \sim N\left(\log[-\log(f_i^*(t))], \sigma[\frac{1}{f_i^*(t)\log(f_i^*(t))}]^2\right), i = 1,2,3,4.$$

We denote the standard error of $\hat{g}_i(t)$ (estimator of $g_i(t)$) as SE. A $100(1-\alpha)$% prediction interval for $g_i(t)$ is given by

$$\log[-\log(\hat{f}_i^*(t))] \pm z_{\alpha/2}SE.$$

After transforming back, a $100(1-\alpha)$% prediction interval for $f_i(t)$ is

$$[\hat{f}_i^*(t)]^{e^{(\pm z_{\alpha/2}SE)}}.$$

In the independent model, we have

$$SE = \frac{\hat{\sigma}_i\{1+e^{-\hat{\beta}_i(t-\hat{\gamma}_i)}\}}{\hat{\alpha}_i\left\{\log(\hat{\alpha}_i) - \log\{1+e^{-\hat{\beta}_i(t-\hat{\gamma}_i)}\}\right\}}.$$

In the competitive model, we have

$$SE = \begin{cases} \frac{\hat{\sigma}\{1+\sum_{j=1}^{3} e^{\hat{b}_j t+c_j}\}}{e^{\hat{b}_i t+c_i}\left\{\hat{b}_i t+c_i - \log\left\{1+\sum_{j=1}^{3} e^{\hat{b}_j t+c_j}\right\}\right\}}, & i = 1,2,3 \\ \frac{\hat{\sigma}\{1+\sum_{j=1}^{3} e^{\hat{b}_j t+c_j}\}}{-\log\left\{1+\sum_{j=1}^{3} e^{\hat{b}_j t+c_j}\right\}}, & i = 4. \end{cases}$$

3. Simulation

In this section, we validate the performance of our estimation methods via simulation in order to demonstrate the accuracy of each method. We generate 1000 samples for each year, following a multinomial distribution, and apply our proposed methods to estimate the distribution of cumulative shares. We consider the scenario in which the true values of the parameters in the multinomial distribution are the true proportions for each year. We evaluate the performance of each method using coverage probability, that is, the probability that an interval will cover the true value. There are six steps in our simulation study.

1. Calculate true values of parameters in a multinomial distribution $(p_1(t), p_2(t), p_3(t), p_4(t), n(t))$, where $p_i(t) = \frac{n_i(t)}{n(t)}$, and $i = 1, 2, 3, 4$. Here, $n_i(t)$ denotes the number of Nobel Prizes awarded to group i by time t, where $1901 \leq t \leq 2022$. Subscripts $i = 1, 2, 3, 4$ correspond to the four groups, North America, Europe, Asia, and Other, respectively.
2. Generate the number of Nobel Prizes awarded to each group for every year according to the multinomial distribution specified in Step 1.
3. Convert the generated data to cumulative shares for each group.
4. Fit the proposed independent and competitive models to generate 90% prediction intervals for four groups in each year.
5. Record 1 if the prediction interval contains the true cumulative share; otherwise, record 0.
6. Repeat Step 2 through Step 5 1000 times to calculate the coverage probability as the proportion of prediction intervals that cover the true cumulative shares.

In Figure 3, we compare the simulated data to the true number of Nobel Prizes awarded to each group from 1901 to 2022. The gray bars give the total number of Nobel Prizes awarded to each group during the period 1901 to 2022. The box plot displays the distribution for the 1000-sample simulated dataset for the total number of Nobel Prizes in each group. It is clear that the simulated data are consistent with the total true number of Nobel Prizes awarded for each group. This validates the simulated data because the median in the box plots is in line with total true number of Nobel Prizes.

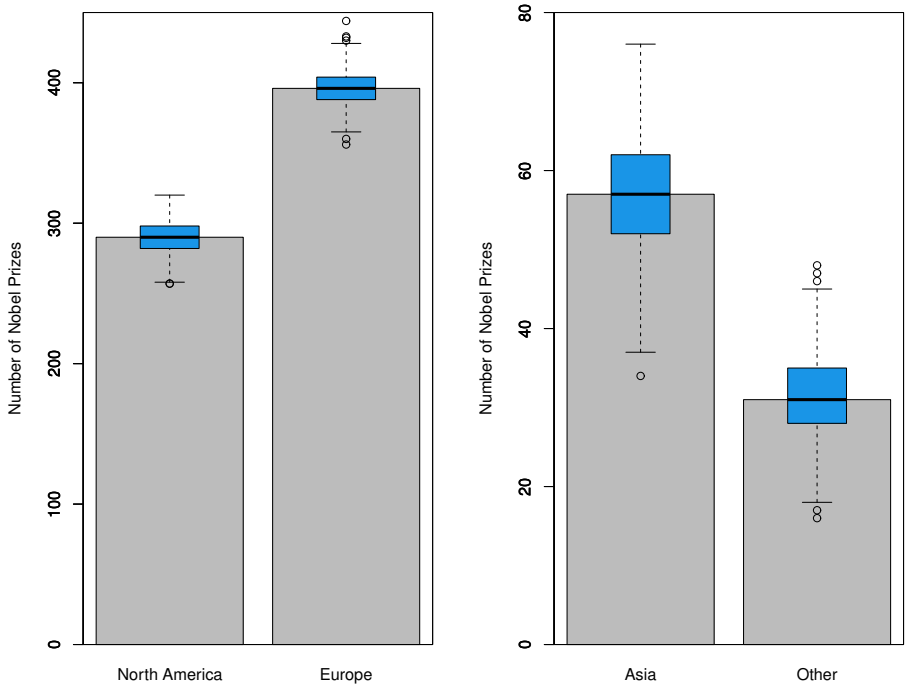

Figure 3. Box plots (in blue) for simulated data and the true total number of Nobel Prizes (gray bars) awarded to each group from 1901 to 2022.

In Figure 4, we show the coverage probability using the proposed competitive and independent models. After the prediction interval is generated using our two models in each simulation, we obtain the coverage probability as the proportion of the time that the interval contains the true value of the cumulative share at time t for each group. For example, the coverage probability is 0.96 in 1922 for North America in the top left graph in Figure 4. This indicates that out of 1000 intervals in the simulation, 960 prediction intervals calculated from the competitive model cover the true cumulative share of North America in 1922. After using a log–log transformation, the prediction interval will not contain a

cumulative share of 0. Therefore, in Figure 4, we exclude the first few years, where some groups did not have a Nobel Prize recipient. For example, most of prizes are awarded to recipients in Europe before 1923.

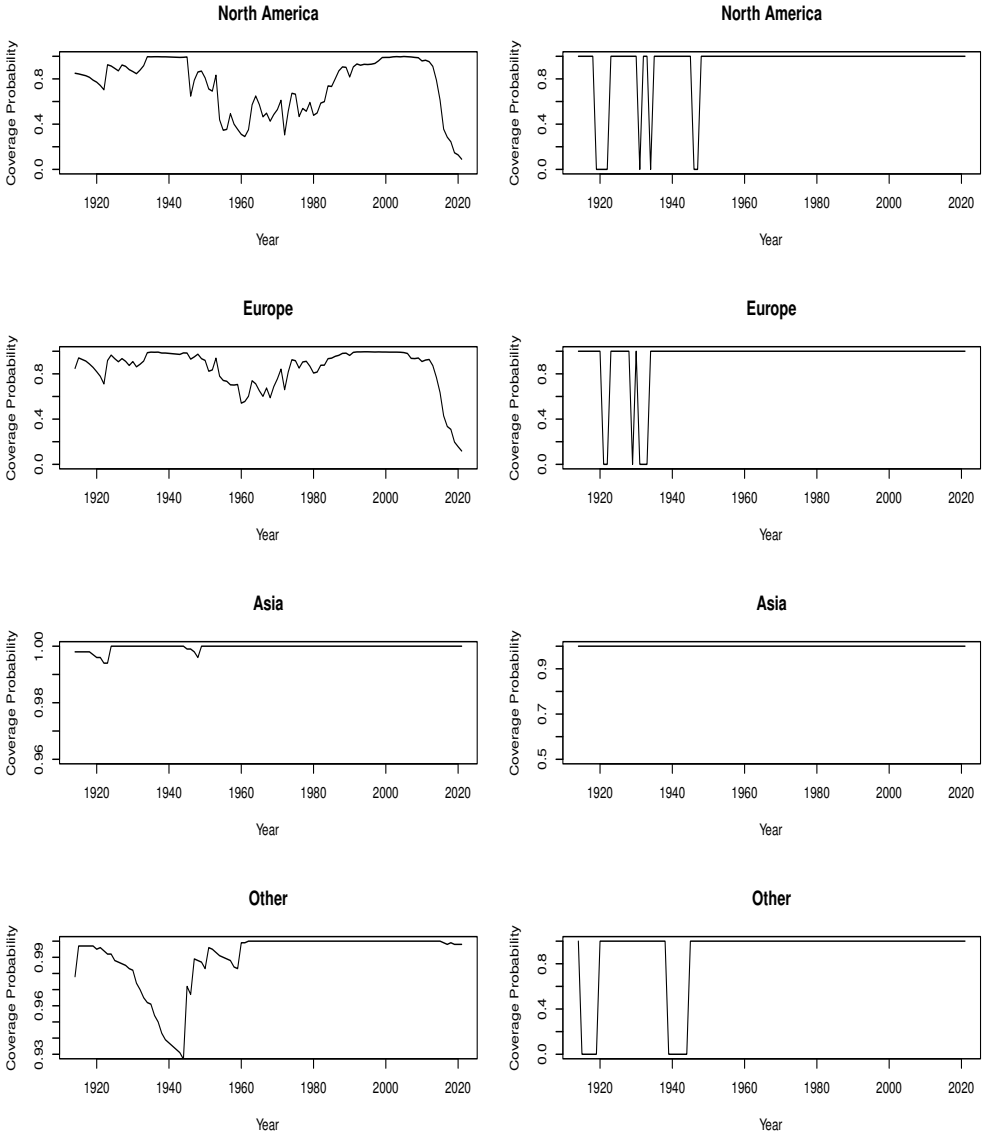

Figure 4. Coverage probability calculated by competitive (**left**) and independent (**right**) models in our simulation study using the multinomial distribution.

The four groups are presented using the competitive and independent models on the left and right sides of Figure 4, respectively. We observe that the competitive model performs consistently well with a high coverage probability for the Asia and Other groups. However, it tends to cover less true cumulative shares for the North America and Europe groups in more recent years. We suspect that the competitive model could be affected by other groups because it takes interactions among groups into consideration. Although there are exceptions in a few years, the independent model consistently contains the most true cumulative shares of the four groups, validating the robustness of its prediction interval generated by the independent model.

In addition to visualizing the coverage probability in Figure 4, we examine the performance of each method using average coverage probability. In Table 1, we calculate the average coverage probability for 1901 to 2022. For example, 0.92 indicates the roughly 92% prediction interval generated by the independent model, covering the true cumulative

shares in North America over all years. As we see from Table 1, the overall result of the proposed models is robust as most of the entries are above 0.9.

Table 1. Average coverage probability for the independent and competitive models for four groups using the multinomial distribution.

Model	North America	Europe	Asia	Other
Independent	0.92	0.94	1	0.93
Competitive	0.73	0.84	0.99	0.99

In order to validate the performance robustness of our estimation methods, we run an additional simulation. In this simulation study, we use a different method to generate the data and analyze the simulated data using proposed independent and competitive models. Similarly, as in the previous simulation study, we have six steps. In the first step, in contrast to generating 1000 samples for each year following a multinomial distribution, we fit the data with a Poisson distribution. We consider the case in which the true values of parameters in the Poisson distribution are the true number of Nobel Prizes awarded to each group for every year. Therefore, we have six steps as follows:

1. Calculate true values of parameters $(\lambda_1(t), \lambda_2(t), \lambda_3(t), \lambda_4(t))$ in a Poisson distribution and $n_i(t) \sim Poisson(\lambda_i(t))$, and $i = 1, 2, 3, 4$. Here, $n_i(t)$ denotes the number of Nobel Prizes awarded to group i by time t where $1901 \leq t \leq 2022$. Subscripts i = 1, 2, 3, 4 correspond to the four groups: North America, Europe, Asia, and Other, respectively.
2. Generate the number of Nobel Prizes awarded to each group for every year according to the Poisson distribution described in Step 1.
3. Convert the generated data to cumulative shares for each group.
4. Fit the proposed independent and competitive models to generate 90% prediction intervals for four groups in each year.
5. Record 1 if the prediction interval contains the true cumulative share; otherwise, record 0.
6. Repeat Step 2 through Step 5 1000 times to calculate the coverage probability as the proportion of prediction intervals that cover the true cumulative shares.

In Figure 5, we present the four groups using the competitive and independent models. Overall, the coverage probabilities over years in each group using either competitive or independent models are similar to the probabilities in Figure 4. In Table 2, we calculate the average coverage probability over years for the independent and competitive models for each group. The numbers are similar to those in Table 1. Therefore, by comparing the simulation results generated from different distributions, we observe that the proposed models including the competitive or independent models exhibit robust performance.

Table 2. Average coverage probability for the independent and competitive models for four groups using the Poisson distribution.

Model	North America	Europe	Asia	Other
Independent	0.92	0.94	1	0.93
Competitive	0.76	0.84	0.99	0.99

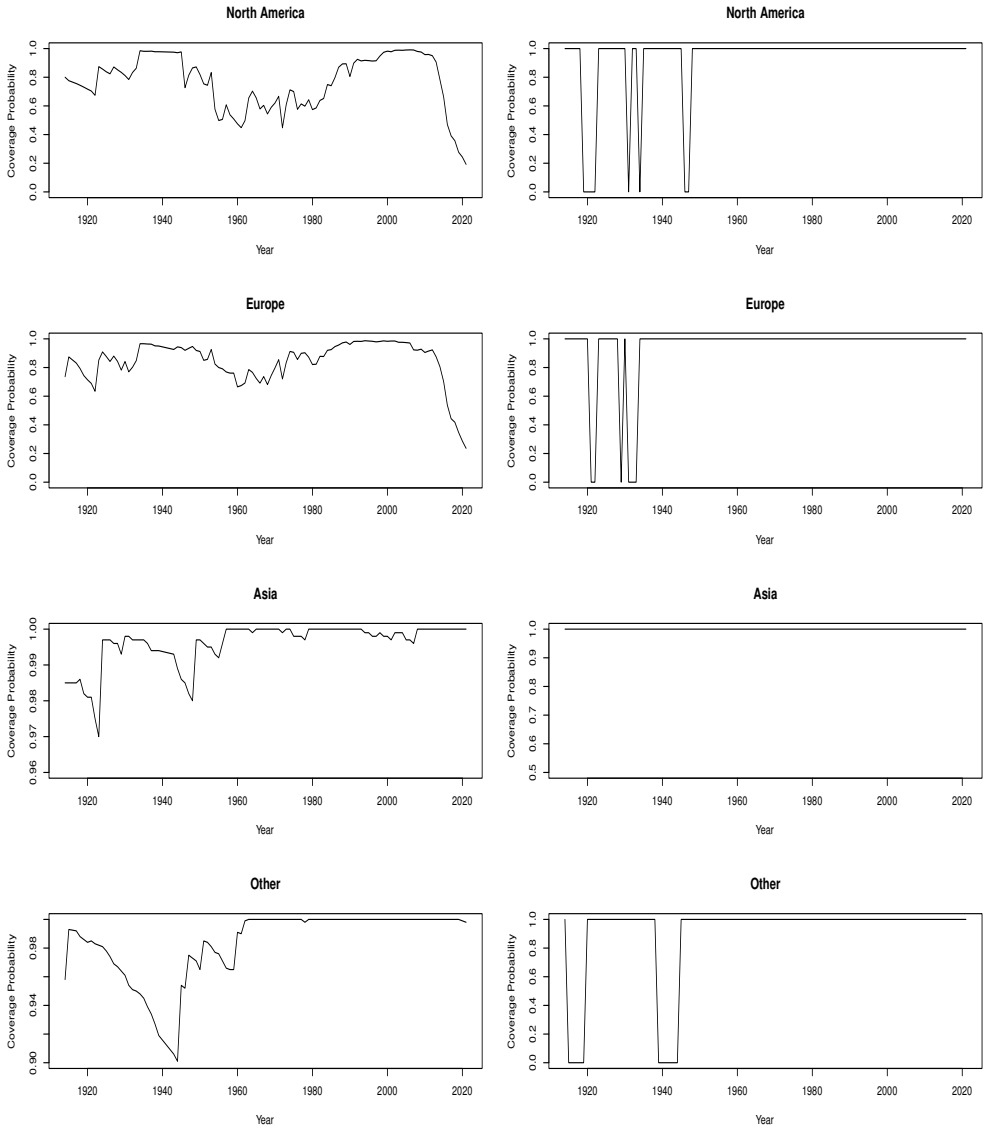

Figure 5. Coverage probability calculated by competitive (**left**) and independent (**right**) models in our simulation study using the Poisson distribution.

4. Performance of Models on Nobel Prize Results from 1901 to 2022

In this section, we evaluate our models using Nobel Prize results from 1901 to 2022. We apply our methods to data from four geographical groups. We only use data for Nobel Prizes in the fields of physics, chemistry, medicine, and economic sciences. In previous articles on this topic, data from all six fields were included. We have decided to focus on the "scientific" prizes, since we view these as less political than the other two categories. We show that our models are useful for estimating the cumulative shares across all groups. We compare estimated curve and prediction intervals of cumulative shares with the true cumulative shares in each group.

In Figures 6–9, we show our 90% estimated prediction intervals as dashed lines and the true cumulative shares as dotted lines using the competitive and independent models. Solid lines are estimated cumulative shares generated by the proposed models. In Figure 6, the curves of the estimated cumulative shares oscillate around the true cumulative shares over time. These curves are included in the estimated prediction intervals most of the time when we use the independent model. We observe similar patterns in Figures 7–9. Using the competitive model, the estimation performances are similar to the independent model with slightly wider prediction intervals in the Asia and Other groups. In North America

and Europe, the cumulative shares estimated by the competitive model tend to be convex or concave in the middle, thus leading to less coverage of the true cumulative shares. These results are consistent with our observations in the simulation study.

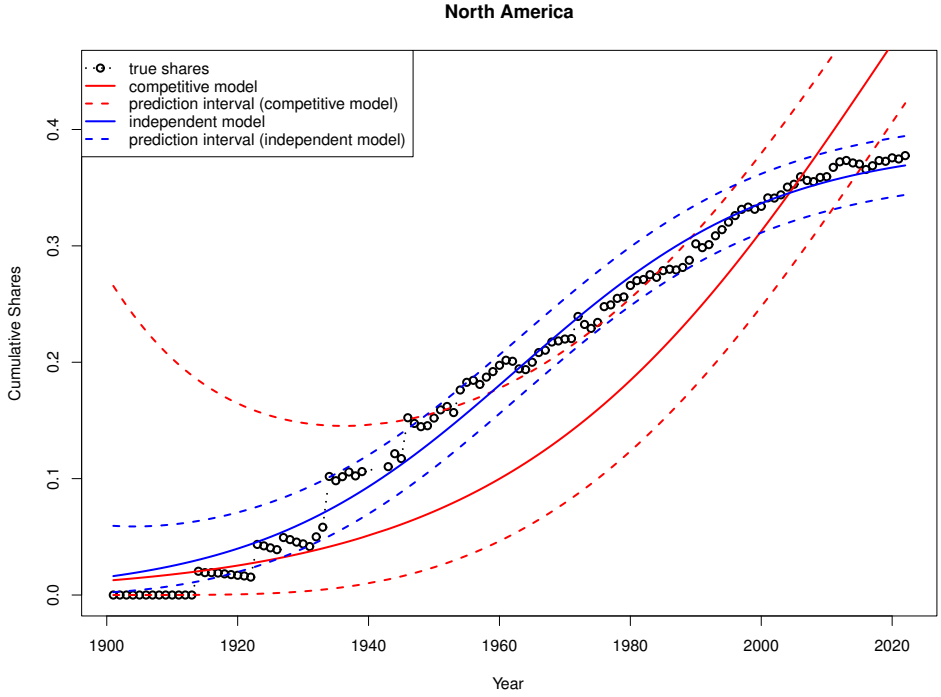

Figure 6. True and estimated cumulative shares in the North America group.

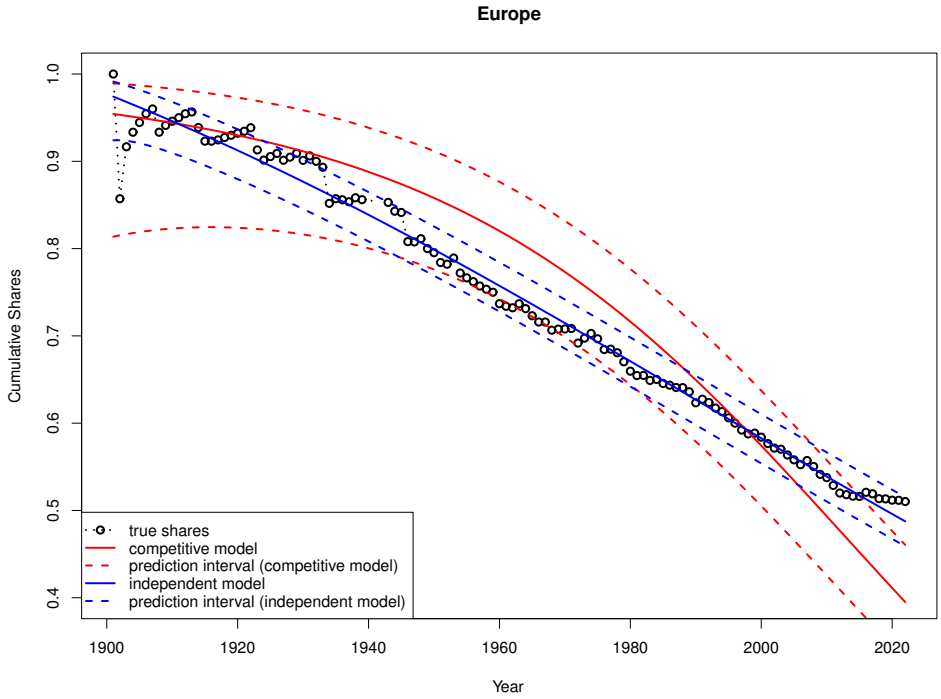

Figure 7. True and estimated cumulative shares in the Europe group.

Overall, these models are robust and accurate in estimating the cumulative shares for all groups. In particular, using the Nobel Prize data from 1901 to 2022, the independent model performed the best on the four groups. In both models, we see an increasing trend in cumulative shares for North America. Previous studies are based on deterministic models. These models are useful, but they suffer from several drawbacks; for example, they cannot

handle variation caused by unknown or unobserved factors. Our stochastic models not only provide accurate fit, but they are able to capture the nature of randomness and provide prediction intervals to cover true cumulative shares with a certain probability.

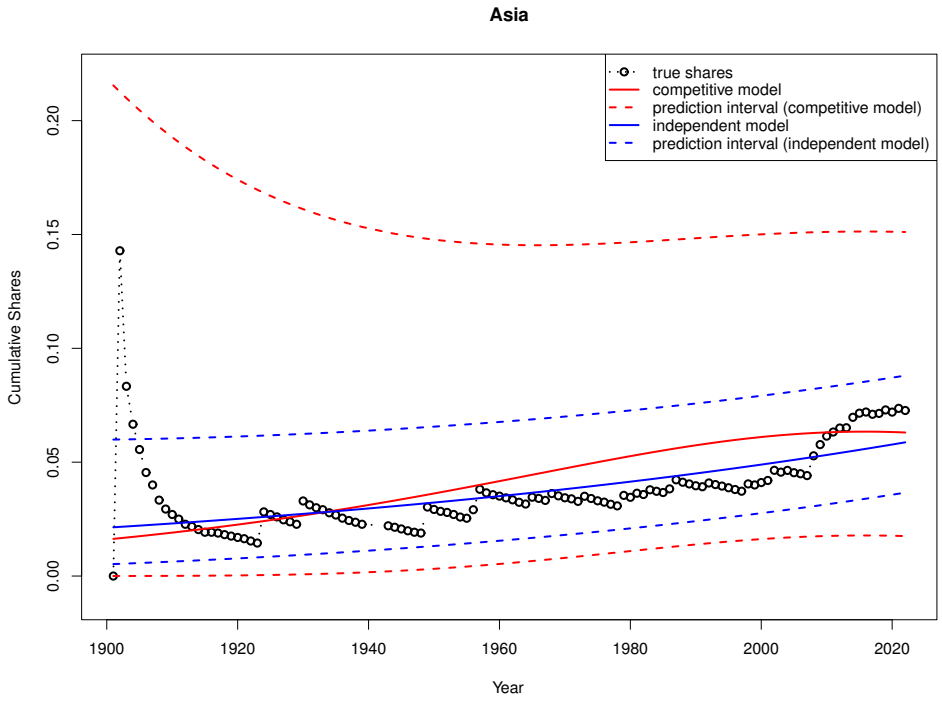

Figure 8. True and estimated cumulative shares in the Asia group.

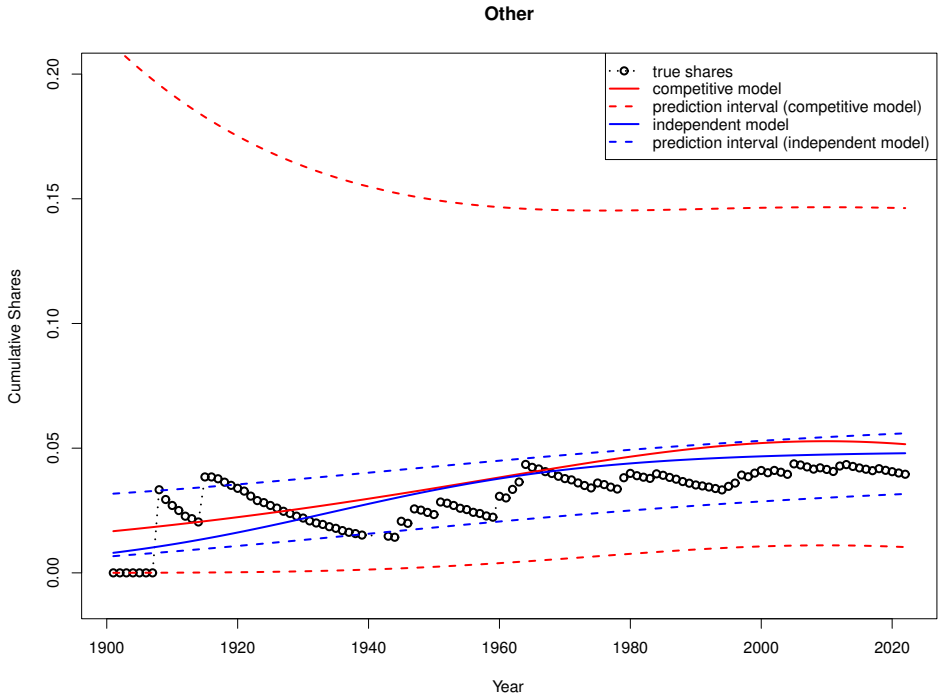

Figure 9. True and estimated cumulative shares in the Other group.

5. Discussion

In this paper, we developed an independent model and a competitive model to estimate the distribution of cumulative shares of Nobel Prizes awarded to recipients in four geographical groups from 1901 to 2022. After we estimated the parameters in both models,

we used a log–log transformation to derive a prediction interval to estimate the cumulative shares where the cumulative shares of $f_i(t)$ are contained in the interval $[0, 1]$.

As we observed in the simulation study and in the analysis of Nobel Prize results, the independent model consistently performed better in terms of fit for all four groups from 1901 to 2022. However, using this model could possibly lead to a large variance problem when predictions are generated due to overfitting. The competitive model has a possibly larger difference between true and predicted values due to better fitting performance in the independent model. One difference between the independent model and the competitive model is that the competitive model takes into account constraints of unity, that is, $\sum_{i=1}^{4} f_i^b(t) = 1$ for any t. This difference possibly helps the independent model have a better fit than the competitive model.

Although there are limitations to our methods, it would be interesting to use both models to forecast future Nobel Prize recipient trends. In Figure 10, we show the cumulative shares for each group for the next decade using both models. In the Asia and Other groups, both models indicate that the changes in cumulative shares remain very small. In the Europe group, both models show a much greater decline in cumulative shares. In North America, we observe a large increase predicted by the competitive model, whereas the independent model shows a small increase in cumulative shares from 2023 to 2033. As we have discussed in this paper, there is a bias–variance tradeoff between the two models. Therefore, we should examine prediction results from both models, although they likely have similar trends.

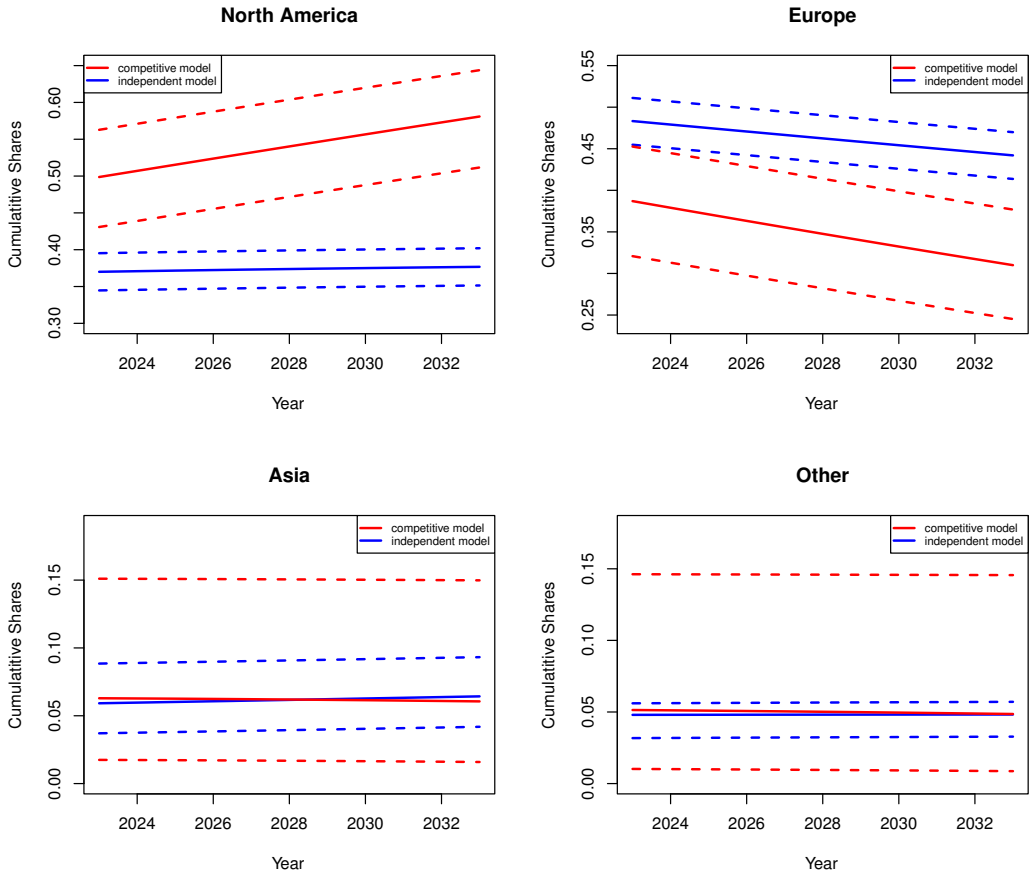

Figure 10. Predicted cumulative shares with prediction intervals (dotted lines) in four groups from 2023 to 2033.

In future work, we could extend our modeling effort in two ways. First, we could remove the assumption of normality and examine different parametric and nonparametric models. Second, we might formulate a covariate-adjusted model so it considers important events related to Nobel Prize fields. We note that it is a major challenge to identify or

collect such covariates. However, if such covariates could be identified, methods such as classification and regression models, including linear and nonlinear models, could be explored in order to understand what variables may significantly relate to Nobel Prize fields. Nonlinear models such as tree-based methods, random forests, support vector machines, and neural networks may provide interesting insights in contrast to linear models.

Author Contributions: All authors have contributed equally to conceptualization, methodology, and writing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

Appendix A

We show how to solve for $f_i^b(t)$ in the competitive model. We start with the following:

$$\begin{cases} [f_i^b(t)]' = f_i^b(t)[b_i - \sum_{j=1}^{3} b_j f_j^b(t)] \quad \text{where} \quad i = 1,2,3 & \text{(A1)} \\ f_i^b(0) = x_i. & \text{(A2)} \end{cases}$$

From Equation (A1), we have

$$b_i - \frac{[f_i^b(t)]'}{f_i^b(t)} = \sum_{j=1}^{3} b_j f_j^b(t).$$

After differentiating both sides, it follows that

$$\int_0^t b_i - \frac{(f_i^b(t))'}{f_i^b(t)} dt = \int_0^t \sum_{j=1}^{3} b_j f_j^b(t) dt. \tag{A3}$$

By setting Equation (A3) to $F(t)$ on both sides, we have

$$\begin{cases} \int_0^t b_i - \frac{(f_i^b(t))'}{f_i^b(t)} dt = F(t) & \text{(A4)} \\ \int_0^t \sum_{j=1}^{3} b_j f_j^b(t) dt = F(t). & \text{(A5)} \end{cases}$$

In Equation (A4), we solve for $f_i^b(t)$. Then we have $f_i^b(t) = e^{b_i t + \log(x_i) - F(t)}$. Solving for $e^{F(t)}$ in Equation (A5), we have $e^{F(t)} = 1 + \sum_{j=1}^{3} x_j(e^{b_j t} - 1)$. Therefore,

$$f_i^b(t) = \frac{x_i e^{b_i t}}{1 + \sum_{j=1}^{3} x_j(e^{b_j t} - 1)}.$$

By applying the transformation $e^{c_i} = \frac{x_i}{1 - \sum_{i=1}^{3} x_i}$, the result follows.

Appendix B

We show how to solve for c_i, $i = 1, 2, 3$. By setting $t = 0$, we have

$$\begin{cases} \dfrac{e^{c_1}}{1+\sum_{i=1}^{3} e^{c_i}} = x_1 & \text{(A6)} \\ \dfrac{e^{c_2}}{1+\sum_{i=1}^{3} e^{c_i}} = x_2 & \text{(A7)} \\ \dfrac{e^{c_3}}{1+\sum_{i=1}^{3} e^{c_i}} = x_3. & \text{(A8)} \end{cases}$$

After summing Equations (A6)–(A8), it follows that $1+\sum_{i=1}^{3} e^{c_i} = 1/(1-\sum_{i=1}^{3} x_i) = 1/x_4$. Since $1+\sum_{i=1}^{3} e^{c_i} = \frac{e^{c_1}}{x_1} = \frac{e^{c_2}}{x_2} = \frac{e^{c_3}}{x_3}$, we have

$$\begin{cases} e^{c_1} = x_1/x_4 \\ e^{c_2} = x_2/x_4 \\ e^{c_3} = x_3/x_4 \end{cases}$$

and the result follows.

References

1. Levinovitz, A.W.; Ringertz, N. *The Nobel Prize: The First 100 Years*; World Scientific: Singapore, 2001.
2. Modis, T. Competition and forecasts for Nobel Prize awards. *Technol. Forecast. Soc. Chang.* **1988**, *34*, 95–102. [CrossRef]
3. Golden, B.; Zantek, P. Inaccurate forecasts of the logistic growth model for Nobel Prizes. *Technol. Forecast. Soc. Chang.* **2004**, *71*, 417–422. [CrossRef]
4. Jank, W.; Golden, B.; Zantek, P. Old world vs. new world: Evolution of Nobel Prize shares. *INFOR Inf. Syst. Oper. Res.* **2005**, *43*, 41–49. [CrossRef]
5. Hartley, H.O. The modified Gauss-Newton method for the fitting of non-linear regression functions by least squares. *Technometrics* **1961**, *3*, 269–280. [CrossRef]
6. Levenberg, K. A method for the solution of certain non-linear problems in least squares. *Q. Appl. Math.* **1944**, *2*, 164–168. [CrossRef]
7. Marquardt, D.W. An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.* **1963**, *11*, 431–441. [CrossRef]
8. Modis, T. US Nobel laureates: Logistic growth versus Volterra–Lotka. *Technol. Forecast. Soc. Chang.* **2011**, *78*, 559–564. [CrossRef]
9. Marasco, A.; Picucci, A.; Romano, A. Market share dynamics using Lotka–Volterra models. *Technol. Forecast. Soc. Chang.* **2016**, *105*, 49–62. [CrossRef]
10. Casella, G.; Berger, R.L. *Statistical Inference*; Cengage Learning: Boston, MA, USA, 2021.