Burgos-Simon, C.; Cortés, J.; Martínez-Rodríguez, D.; Villanueva Micó, RJ. (07-2). Computational modelling with uncertainty of frequent users of e-commerce in Spain using an age-group dynamic nonlinear model with varying size population. Advances in Complex Systems. 22(4):1950009-1-1950009-17. https://doi.org/10.1142/S0219525919500097

Computational Modelling with uncertainty of frequent users of e-commerce in Spain using an age-group dynamic nonlinear model with varying size population

C. Burgos, J.-C. Cortés, D. Martínez-Rodríguez, R.-J. Villanueva

Abstract

Electronic commerce has numerous advantages. It allows saving time when we purchase an item, offers the possibility of review without depending on the schedules of traditional stores, access to a wider variety and quantity of articles, in many cases, with lower prices, etc. Based upon mathematical epidemiology tenets strongly related to social behavior able to describe the influence of peers, in this paper we propose an age-group dynamic model with population varying size based on a system of difference equations to study the evolution of the frequent users of electronic commerce over time in Spain. Using data from surveys retrieved from the Spanish National Statistics Institute, we use and design computational algorithms to perform a probabilistic estimation of the model parameters that allow the model output to capture the data uncertainty. Then, we will be able to perform a precise prediction with uncertainty.

Keywords: Electronic Commerce; Real-World Mathematical Model; Nonlinear System of Difference Equations; Uncertainty Quantification.

1 Introduction

Electronic commerce (in the following EC) gathers all the possibilities of purchasing or selling via the Internet. It includes electronic marketing, safe payments via the Internet, automatic systems to control the inventories, the supply chain management, etc.

The diffusion of EC has increased its business volume much more than the traditional commerce in the recent years. Several factors have fostered this growth: as it is not necessary to go physically to the shop, it saves time; it allows to compare quality, features and prices without moving from home; there are a lot of free apps for shopping using the smartphone; it is possible to read the opinion and comments of other customers who bought the same item you want to purchase; the purchasing platforms allow the connection to social networks [15].

Furthermore, other fields of science have been involved in the evolution of EC, for instance, artificial intelligence, smart commerce, analytics and big data. Also, EC is transforming the traditional business models in travels, banking, fashion, transportation, etc. [15].

EC has become a part of our lives, transforming the way of making economic transactions. Although the number of individuals who buy via the Internet is increasing over the time, there are still many people who do not. To carry out the present study, we will use available data from the National Statistic Institute of Spain, where the amount of people who use frequently EC for their purchases has been measured at different time instants using sampling statistical techniques.

As any human activity, EC can be considered a practice susceptible to be transmitted by peers with whom we are related with, that is, our social network [6]. Thus, the study of the evolution of the EC users may be approached using a proper model that considers the influence of peers. The people influenced are usually called imitators [4]. However, in the economic activities also arises the profile of innovator, an individual who makes his/her own decisions regardless of the decisions of others [4]. Both, the influence of the innovators and the imitators should be considered in the building of our proposed mathematical model. This will allow us to investigate and quantify the influence of each one of them.

Dynamic mathematical models are powerful tools to explain and predict the process of adoption of an innovation over the time. The formulation of a reliable mathematical model must consider the particular features of the technology as well as its users. In spite of several authors have developed interesting mathematical diffusion models for study the dynamics of some technologies using different approaches [14, 16, 13, 12], they do not consider aspects like the different habits among people depending on their age and a major impact of innovations on certain age groups. This is particularly relevant because the use of EC requires a certain knowledge and technological skills that are more common in younger people.

In [7], the age, the innovators and the imitators have been considered in the model. Nevertheless the uncertainty of the data coming from a survey, and consequently, with an intrinsic error that may influence the study, has not been considered and treated.

In this paper, the goal is to focus on the study of the potential users of EC, as a fundamental part in the expansion of this increasing business. This way, the target population will be divided into different age groups which allow us to distinguish the different population behavior with respect to this technology. Then, we will build a model that considers the effect of the innovators, the imitators and the data uncertainty measurement error in the survey. In the model, we will assume that the population varies in size over time, which is not usually considered.

In the mathematical quantification of the uncertainty, we will use a new technique for probabilistic estimation of the model parameters that best capture the data uncertainty, different to the one developed in [8]. The new technique is

computationally cheaper than the one introduced in [8]. The model parameters estimated probabilistically will give us tools to analyze how the dynamics occurs. Finally, the proposed model provides a useful tool for forecasting the short-term trends of EC in Spain and therefore to enable strategic decision-making marketing for ensuring efficiency in production, sales and advertising campaigns.

The paper is organized as follows. In Section 2, we state the underlying age-structured demographic model and, onto it, the model that describes the dynamics of the frequent users of EC. In Section 3, we describe the probabilistic estimation technique to fit the model to the data of frequent users capturing the uncertainty involved in the sampled data. In Section 4, we present the results of the probabilistic estimation, the probabilistic prediction and the discussion of the obtained results. Finally, conclusions are drawn in Section 5.

2 Model Building

In this section, first, we recall the data available in the Spanish INE: birth rates, death rates and percentages of frequent users of EC. Frequent users of EC are those who have bought by the Internet in the last three months. Then, we introduce a demographic model to know the general dynamics of the population. Finally, using the parameters of the demographic model, we will be able to describe the dynamics of the frequent EC users.

2.1 Available data

In Spanish INE [1], we can find the data corresponding to the percentage of EC users from 15 to 74 years old divided into the following age groups: 15-24, 25-34, 35-44, 45-54, 55-64 and 65-74, every year in the period 2009-2017.

As we have pointed out in the Introduction, in our previous contribution [7] we proposed a deterministic mathematical model to describe the dynamics of EC in Spain, where the above six age groups were considered. In the present paper, our aim is to treat the uncertainty included in the sampled data. Thus, the idea of considering all the age groups could make the random model very complex, as it occurs in [7]. Therefore, taking into account the significant difference between the percentage of frequent users of EC if they are younger or older than 45 years old, we are going to consider the age groups 15-44 and 45-74. We assume that this difference will be reduced as young people are getting older, but it will take longer than the prediction time of 4 years we will propose.

In Table 1, we can see the aggregated percentages of people who have purchased by the Internet in the last 3 months (frequent users of EC) and the people who do not (non-frequent users of EC), per age groups 15-44 and 45-74 in the period 2009-2017.

2.2 Demographical model

As we discussed in the above section, we consider two age groups, that is,

- Group 1 (G_1): Population aged between 15 and 44 years old.
- Group 2 (G_2): Population aged between 45 and 74 years old.

Then, following [10, p.623-624], taking into account that the time step is going to be fixed in a month and the demographic data retrieved from [3] are in years, the demographic model is given by the following system of difference equations,

$$\begin{aligned} G_1(t+1) &= G_1(t) + \frac{\mu}{12} - \frac{c_1}{12}G_1(t) - \frac{d_1}{12}G_1(t), \\ G_2(t+1) &= G_2(t) + \frac{c_1}{12}G_1(t) - \frac{d_2}{12}G_2(t), \end{aligned} \tag{1}$$

where

- μ is the yearly birth rate (assuming that almost nobody dies between 0 and 14 years),
- d_1 is the yearly death rate in the age group G_1,
- c_1 is the yearly growth rate from G_1 to G_2.
- d_2 is the yearly rate of people who leave age group G_2, by death or because they turn 75.

The total population $P_T(t) = G_1(t)+G_2(t)$ is not constant. At this point, we could use the yearly demographic data retrieved from [3] and calculate the corresponding constant values of $G_1(t)$ and $G_2(t)$ for each year $t = 2009, \ldots, 2017$. Nevertheless, demographic data beyond 2017 to perform predictions of frequent EC users will not be available. Consequently, we are going to consider that the demographic parameters will be determined later by probabilistic estimations, being these values between the maximum and minimum values retrieved from [3]. Thus, these *average* fitted demographic parameter values will be used for predictions.

Then, from [3], we have that $\mu \in [0.0035, 0.0251]$, $d_1 \in [3.5279\times10^{-5}, 0.0032]$, c_1 is about $\frac{1}{45-15} = \frac{1}{30} = 0.0333$ because the length of the age group $15-44$ is 30 years and then, approximately 1/30 of people in this age group grow from 44 to 45 years old, moving then to age group G_2. Also, we should note that d_2 gathers the death rate of people in G_2, in the interval $[0.001269, 0.02432]$ plus the rate of leaving the system, approximately $\frac{1}{75-45} = \frac{1}{30} = 0.0333$, again, because the length of the age group $45-74$ is 30 years.

2.3 Electronic commerce dynamic model

Considering the demographic model previously introduced, we are going to build a discrete model able to describe the transmission dynamics of the habit of the frequent use of EC over time. First, taking the time t in months, we introduce the following subpopulations

- $N_i(t)$, $i = 1, 2$, denotes the number of individuals in the age group G_i who have not purchased by the Internet at least in the last three months, at the month t.

- $Y_i(t)$, $i = 1, 2$, denotes the number of individuals in the age group G_i who have purchased by the Internet in the last three months, at the month t.

A first consequence of this division is that $N_1(t)+Y_1(t) = G_1(t)$ and $N_2(t)+Y_2(t) = G_2(t)$. Therefore, $P_T(t) = N_1(t) + Y_1(t) + N_2(t) + Y_2(t)$ and as we mentioned before, the total population is not constant over the time.

Furthermore, the frequent users of EC can be classified into two main groups: the innovators, who make the decision of using EC because of advertising or marketing strategies through the media or other external factors regardless of the decisions of others; and the imitators, who will use EC due to the influence received from social interaction with frequent users [4]. The effect of both, the innovators and the imitators, is going to be taken into account in the modeling process.

The diffusion of the frequent use of the EC will be represented by the transition of people from the population $N_i(t)$ to $Y_i(t)$ $(i = 1, 2)$ through the coefficients of innovation and imitation described by:

- p_i, $i = 1, 2$, are the coefficients of innovators for the i-th group, and the transitions due to innovators from $N_i(t)$ to $Y_i(t)$ are modeled by $p_i N_i(t)$, $i = 1, 2$;

- $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are the coefficients of the imitators, that is, when someone is a frequent user of EC and influences another person who has not used it yet or does it less frequently, that is, it has not used EC in the last three months. These transitions are modeled by the terms [5, p. 14]

$$\alpha_1 N_1(t)\frac{Y_1(t)}{P_T(t)} + \alpha_2 N_1(t)\frac{Y_2(t)}{P_T(t)},$$

and

$$\alpha_3 N_2(t)\frac{Y_1(t)}{P_T(t)} + \alpha_4 N_2(t)\frac{Y_2(t)}{P_T(t)};$$

- γ_i, $i = 1, 2$, are the transition parameters of those who have not purchased by the Internet in the last three months and the transitions from $Y_i(t)$ to $N_i(t)$ are modeled by the terms $\gamma_i Y_i(t)$, $i = 1, 2$.

Then, assuming that the individuals that turn 15 years old are not frequent users of EC, the following age-structured mathematical diffusion model based on the nonlinear system of difference equations given by expressions (2)–(5) describes the evolution of the frequent users of EC in Spain over the time.

$$N_1(t+1) = N_1(t) - \frac{d_1}{12}N_1(t) + \gamma_1 Y_1(t) - \frac{c_1}{12}N_1(t) - N_1(t)\frac{\alpha_1 Y_1(t) + \alpha_2 Y_2(t)}{P_T(t)} - p_1 N_1(t) + \frac{\mu}{12}P_T(t), \tag{2}$$

$$Y_1(t+1) = Y_1(t) - \frac{d_1}{12}Y_1(t) - \gamma_1 Y_1(t) - \frac{c_1}{12}Y_1(t) + N_1(t)\frac{\alpha_1 Y_1(t) + \alpha_2 Y_2(t)}{P_T(t)} + p_1 N_1(t), \tag{3}$$

$$N_2(t+1) = N_2(t) - \frac{d_2}{12}N_2(t) + \gamma_2 Y_2(t) + \frac{c_1}{12}N_1(t) - N_2(t)\frac{\alpha_3 Y_1(t) + \alpha_4 Y_2(t)}{P_T(t)} - p_2 N_2(t), \tag{4}$$

$$Y_2(t+1) = Y_2(t) - \frac{d_2}{12}N_2(t) - \gamma_2 Y_2(t) + \frac{c_1}{12}Y_1(t) + N_2(t)\frac{\alpha_3 Y_1(t) + \alpha_4 Y_2(t)}{P_T(t)} + p_2 N_2(t). \tag{5}$$

Figure 1 shows the compartmental representation of the system of difference equations (2)–(5).

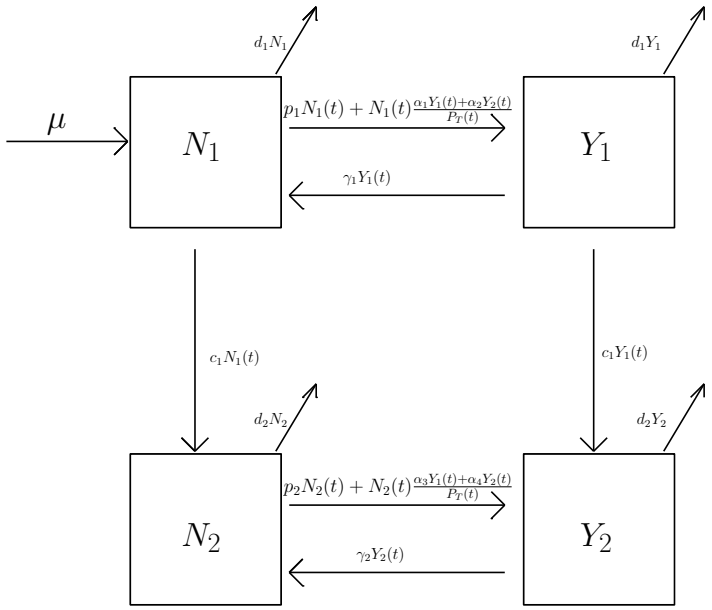

Figure 1: Compartmental model corresponding to the system of nonlinear of difference equations (2)–(5). The boxes represent the subpopulations and the arrows the transitions between them.

As our data are given in percentages, we need to scale the model to match

the magnitudes. To do so, first, we need to establish a relationship between $P_T(t+1)$ and $P_T(t)$. Summing up all the expressions in (2)–(5), we obtain that,

$$P_T(t+1) = P_T(t) + \frac{\mu}{12}P_T(t) - \frac{d_1}{12}(N_1(t) + Y_1(t)) - \frac{d_2}{12}(N_2(t) + Y_2(t)). \quad (6)$$

Now, if we define

$$n_1(t) := \frac{N_1(t)}{P_T(t)},\ y_1(t) := \frac{Y_1(t)}{P_T(t)},\ n_2(t) := \frac{N_2(t)}{P_T(t)},\ y_2(t) := \frac{Y_2(t)}{P_T(t)}, \quad (7)$$

and we divide the expression (2) by $P_T(t+1)$ given by (6), we obtain

$$\frac{N_1(t+1)}{P_T(t+1)} = \frac{N_1(t) - \frac{d_1}{12}N_1(t) + \gamma_1 Y_1(t) - \frac{c_1}{12}N_1(t) - N_1(t)\frac{\alpha_1 Y_1(t) + \alpha_2 Y_2(t)}{P_T(t)} - p_1 N_1(t) + \frac{\mu}{12}P_T(t)}{P_T(t) + \frac{\mu}{12}P_T(t) - \frac{d_1}{12}(N_1(t) + Y_1(t)) - \frac{d_2}{12}(N_2(t) + Y_2(t))}. \quad (8)$$

Then, if we divide numerator and denominator of equation (8) by $P_T(t)$, taking into account (6) and (7), we have

$$n_1(t+1) = \frac{n_1(t) - \frac{d_1}{12}n_1(t) + \gamma_1 y_1(t) - \frac{c_1}{12}n_1(t) - n_1(t)(\alpha_1 y_1(t) + \alpha_2 y_2(t)) - p_1 n_1(t) + \frac{\mu}{12}}{1 + \frac{\mu}{12} - \frac{d_1}{12}(n_1(t) + y_1(t)) - \frac{d_2}{12}(n_2(t) + y_2(t))}, \quad (9)$$

and then, all the terms in (9) are scaled. Using the same procedure in (3)-(5), we obtain the following scaled system of nonlinear difference equations

$$n_1(t+1) = \frac{n_1(t) - \frac{d_1}{12}n_1(t) + \gamma_1 y_1(t) - \frac{c_1}{12}n_1(t) - n_1(t)(\alpha_1 y_1(t) + \alpha_2 y_2(t)) - p_1 n_1(t) + \frac{\mu}{12}}{1 + \frac{\mu}{12} - \frac{d_1}{12}(n_1(t) + y_1(t)) - \frac{d_2}{12}(n_2(t) + y_2(t))}, \quad (10)$$

$$y_1(t+1) = \frac{y_1(t) - \frac{d_1}{12}y_1(t) - \gamma_1 y_1(t) - \frac{c_1}{12}y_1(t) + n_1(t)\left(\alpha_1 y_1(t) + \alpha_2 y_2(t)\right) + p_1 n_1(t)}{1 + \frac{\mu}{12} - \frac{d_1}{12}(n_1(t) + y_1(t)) - \frac{d_2}{12}(n_2(t) + y_2(t))}, \quad (11)$$

$$n_2(t+1) = \frac{n_2(t) - \frac{d_2}{12}n_2(t) + \gamma_2 y_2(t) + \frac{c_1}{12}n_1(t) - n_2(t)\left(\alpha_3 y_1(t) + \alpha_4 y_2(t)\right) - p_2 n_2(t)}{1 + \frac{\mu}{12} - \frac{d_1}{12}(n_1(t) + y_1(t)) - \frac{d_2}{12}(n_2(t) + y_2(t))}, \quad (12)$$

$$y_2(t+1) = \frac{y_2(t) - \frac{d_2}{12}y_2(t) - \gamma_2 y_2(t) + \frac{c_1}{12}y_1(t) + n_2(t)\left(\alpha_3 y_1(t) + \alpha_4 y_2(t)\right) + p_2 n_2(t)}{1 + \frac{\mu}{12} - \frac{d_1}{12}(n_1(t) + y_1(t)) - \frac{d_2}{12}(n_2(t) + y_2(t))}. \quad (13)$$

In this manner, it is clear that $n_1(t) + y_1(t) + n_2(t) + y_2(t) = 1$ for all t.

3 Probabilistic estimation

3.1 Data

Here, the next usual step would be to fit the model to data in Table 1, that is, to find the model parameter values that make the model output be as close as possible to the data collected in Table 1 in the time instants t_1= Dec 2009, ..., t_9= Dec 2017. However, the data come from surveys and, therefore, contain intrinsic uncertainties (survey error) that we want the model captures.

In order to quantify the uncertainty of the data, it would be interesting to have the complete data of the surveys. Nevertheless, this information does not use to be available. As an alternative to avoid this drawback, our goal is to assign reasonable probability distributions to data that allow us to simulate the real survey samples in a reliable way.

To do so, we need the sample sizes of the surveys, collected in Table 2 and available in [1]. We will assume that people interviewed each year is different and, consequently, the survey outputs are independent. For each one of the 9 available surveys, let us denote by $X^j = \left(X_1^j, X_2^j, X_3^j, X_4^j\right), 0 \leq X_i^j \leq n_j, i = 1, \ldots, 4, j = 1, \ldots, 9$ a random vector whose entries are:

- X_1^j = Number of individuals who purchase frequently by the Internet in the first age group G_1 at the time instant j.
- X_2^j = Number of individuals who do not purchase frequently by the Internet in the first age group G_1 at the time instant j.
- X_3^j = Number of individuals who purchase frequently by the Internet in second age group G_2 at the time instant j.
- X_4^j = Number of individuals who do not purchase frequently by the Internet in second age group G_2 at the time instant j.

These components represent exclusive selections (events) with probabilities:

$$\mathbb{P}^j(X_i^j = x_i) = \rho_i^j, \qquad i = 1, \ldots, 4 \qquad j = 1, \ldots, 9,$$

where $\rho_1^j, \rho_2^j, \rho_3^j$ and ρ_4^j are the percentages collected in Table 1 for each survey j, $j = 1, \ldots, 9$. Thus each random vector has a multinomial (tetranomial) probability distribution. Therefore, the probability that X_1^j occurs x_1 times, X_2^j occurs x_2 times, X_3^j occurs x_3 times and X_4^j occurs x_4 times is given by

$$\mathbb{P}_{n_j}^j(x_1, x_2, x_3, x_4) = \frac{n_j!}{x_1!x_2!x_3!x_4!}(\rho_1^j)^{x_1}(\rho_2^j)^{x_2}(\rho_3^j)^{x_3}(\rho_4^j)^{x_4},\ j = 1, \ldots, 9, \quad (14)$$

where $x_1 + x_2 + x_3 + x_4 = n_j$.

Now, let us scale the above tetranomial distributions in order to have the same magnitudes in the data and in the model. Then, we define the random vector

Year	Non-users $15-44$	Users $15-44$	Non-users $16-74$	Users $16-74$
t_1 = Dec 2009 ($j=1$)	0.4796	0.1008	0.392	0.0276
t_2 = Dec 2010 ($j=2$)	0.4588	0.1170	0.3885	0.0357
t_3 = Dec 2011 ($j=3$)	0.4371	0.1307	0.3929	0.0393
t_4 = Dec 2012 ($j=4$)	0.4204	0.1393	0.3927	0.0476
t_5 = Dec 2013 ($j=5$)	0.3885	0.1627	0.3934	0.0554
t_6 = Dec 2014 ($j=6$)	0.3725	0.1684	0.3979	0.0612
t_7 = Dec 2015 ($j=7$)	0.3338	0.1958	0.3897	0.0807
t_8 = Dec 2016 ($j=8$)	0.2976	0.2213	0.3806	0.1005
t_9 = Dec 2017 ($j=9$)	0.2709	0.2401	0.3785	0.1105

Table 1: Proportion of people who have purchased by the Internet in the last 3 months (frequent users of EC) and the people who do not (non-frequent users of EC), for the age groups 15-44 and 45-74 in the period 2009-2017 in Spain [1].

Year	Sample Size (=: n_j)
t_1= Dec 2009 (j=1)	24935
t_2= Dec 2010 (j=2)	24877
t_3= Dec 2011 (j=3)	24972
t_4= Dec 2012 (j=4)	20647
t_5= Dec 2013 (j=5)	20484
t_6= Dec 2014 (j=6)	20815
t_7= Dec 2015 (j=7)	20786
t_8= Dec 2016 (j=8)	23877
t_9= Dec 2017 (j=9)	24132

Table 2: Sample size for each survey, [1].

$$Z^j = \left(Z_1^j, Z_2^j, Z_3^j, Z_4^j\right), \qquad Z_i^j = \frac{X_i^j}{n_j}, \qquad i = 1, \ldots, 4, \qquad j = 1, \ldots, 9.$$

Note that the random variables $Z_i^j \in [0, 1]$ and the join probability mass function of the scaled tetranomial distribution Z^j is the same as the one of X^j shown in expression (14).

Now, sampling a hundred thousand times the 9 scaled tetranomial probability distributions of Z^j, substituting ρ_1^j, ρ_2^j, ρ_3^j and ρ_4^j by their corresponding values in Table 1, for $j = 1, \ldots, 9$, we can obtain the quantiles 2.5 and 97.5, (95% confidence interval) of these scaled tetranomial probability distributions. These quantiles are collected in Table 3 and capture most of the data uncertainty.

Year	95% CI of $n_1(t)$ $15-44$	95% CI of $y_1(t)$ $15-44$	95% CI of $n_2(t)$ $45-74$	95% CI of $y_2(t)$ $45-74$
t_1= Dec 2009 (j=1)	$[0.4733, 0.4857]$	$[0.0971, 0.0104]$	$[0.3859, 0.3980]$	$[0.0255, 0.0296]$
t_2= Dec 2010 (j=2)	$[0.4526, 0.4650]$	$[0.1130, 0.1210]$	$[0.2834, 0.3944]$	$[0.0334, 0.0380]$
t_3= Dec 2011 (j=3)	$[0.4308, 0.4432]$	$[0.1265, 0.1348]$	$[0.3868, 0.3990]$	$[0.0368, 0.0416]$
t_4= Dec 2012 (j=4)	$[0.4135, 0.4270]$	$[0.1344, 0.1439]$	$[0.3861, 0.3994]$	$[0.0448, 0.0506]$
t_5= Dec 2013 (j=5)	$[0.3818, 0.3951]$	$[0.1576, 0.1677]$	$[0.3867, 0.4001]$	$[0.0522, 0.0585]$
t_6= Dec 2014 (j=6)	$[0.3659, 0.3790]$	$[0.1632, 0.1735]$	$[0.3913, 0.4046]$	$[0.0578, 0.0643]$
t_7= Dec 2015 (j=7)	$[0.3272, 0.3401]$	$[0.1905, 0.2013]$	$[0.3830, 0.3962]$	$[0.0770, 0.0844]$
t_8= Dec 2016 (j=8)	$[0.2918, 0.3034]$	$[0.2160, 0.2266]$	$[0.3744, 0.3868]$	$[0.0966, 0.1042]$
t_9= Dec 2017 (j=9)	$[0.2652, 0.2765]$	$[0.2345, 0.2454]$	$[0.3732, 0.3846]$	$[0.1066, 0.1145]$

Table 3: Calculated 95% confidence intervals, for each subpopulation, $n_1(t)$, $y_1(t)$, $n_2(t)$, $y_2(t)$, for the time instants Dec 2009, Dec 2010, ..., Dec 2017, of the data.

3.2 Probabilistic estimation

The goal of this section is to determine sets of model parameters, which substituted into the model and calculating the model output, the means and the 95% confidence intervals of the model output in the time instants t_1= Dec 2009, ..., t_9= Dec 2017 approximate as much as possible the data means and the data 95% confidence intervals given in Tables 1 and 3, respectively.

In order to determine the appropriate model parameters, we are going to apply a computational technique other than the probabilistic fitting technique presented in the paper [8], with the aim of saving computation time with similar or even better results.

Thus, in this section, we are going to determine an appropriate fitness function that measures if the model output lies inside or is close the data 95% confidence intervals. Then, we will run several times the optimization Particle Swarm Optimization algorithm (PSO) to minimize the defined fitness function with the aim to calibrate the model. However, we are going to store all the model evaluations and their errors (fitnesses). Now, among these performed evaluations, we need to select those that allow us to capture the data uncertainty as well as possible. To select the suitable evaluations and, therefore, their model parameters, we have to propose a new selection algorithm. This algorithm has been inspired in the PSO algorithm with an appropriate fitness function that measures the closeness between the means and the 95% confidence intervals of the model output of the evaluations and the data.

For the sake of clarity in the explanation of the processes properly, first, we need to establish certain definitions and notations. Let us denote as $\mathbb{M}(t;\mathcal{P})$ a short representation of model (10)–(13), where

$$\mathcal{P} = \{\mu, d_1, c_1, d_2, p_1, \alpha_1, \alpha_2, \gamma_1, p_2, \alpha_3, \alpha_4, \gamma_2\},$$

are the model parameters and t is the time instant (in months). Now, given a set of model parameters, say $\mathcal{P}^*$, the model output $\mathbb{M}(t_i;\mathcal{P}^*) = (o_{t_i1}, o_{t_i2}, o_{t_i3}, o_{t_i4})$ is a vector with 4 elements corresponding to subpopulations n_1, y_1, n_2 and y_2, for the time instants t_1= Dec 2009, ..., t_9= Dec 2017. Also, we denote by I_{kl}, $k = 1, \ldots, 9$, $l = 1, 2, 3, 4$, the 95% confidence intervals given in Table 3. For example, $I_{34} = [0.0368, 0.0416]$. Furthermore, we define the distance from a point $p \in \mathbb{R}$ to an interval $[a, b]$ as follows:

$$D(p,[a,b]) = \begin{cases} 0 & \text{if } a \le p \le b, \\ \min\{|a-p|, |b-p|\} & \text{otherwise.} \end{cases} \tag{15}$$

Thus, we define the fitness function F in the Algorithm 1, where we consider that the fitness is zero if the model output values lie inside the 95% confidence intervals of the data.

Algorithm 1: Fitness function F used for model calibration.

Input : Model parameters
$\mathcal{P} = \{\mu, d_1, c_1, d_2, p_1, \alpha_1, \alpha_2, \gamma_1, p_2, \alpha_3, \alpha_4, \gamma_2\}$.
Output: Fitness $F(\mathcal{P})$

1 Substitute the model parameters $\mathcal{P}$ into the model;
2 **for** t_1= *Dec 2009, ...,* t_9= *Dec 2017* **do**
3 | Calculate the model outputs $\mathbb{M}(t_i;\mathcal{P}) = (o_{t_i1}, o_{t_i2}, o_{t_i3}, o_{t_i4})$;
4 **end**
5 Calculate the fitness of $\mathcal{P}$ as $F(\mathcal{P}) = \sum_{k=1}^{9} \sum_{l=1}^{4} D(o_{t_kl}, I_{kl})$;

In the following, the calculation performed in the loop 2-4 of the Algorithm 1 will be called realization. Once the fitness function has been determined, we

will use the rPSO algorithm introduced in [11] to calibrate the model. In fact, we are going to run it several times, storing the sets of model parameters and their fitnesses used by rPSO in the set $\mathcal{P}$. Later, we have to select the model parameters of $\mathcal{P}$ whose output capture the best as possible the data uncertainty.

To perform the selection of these model parameters, we are going to introduce an adapted version of the rPSO algorithm. Then, we define the fitness function G given by the Algorithm 2.

Algorithm 2: Fitness function G.

Input : $\mathbb{A} = \{\mathcal{P}_{i_1}, \mathcal{P}_{i_2}, \ldots, \mathcal{P}_{i_n}\}$, $\mathcal{P}_{i_j}$ sets of model parameter values $i = 1, \ldots, n$.

Output: Fitness $G(\mathbb{A})$

1 **foreach** $\mathcal{P}_{i_j}$ *set of parameters in* $\mathbb{A}$ **do**
2 | Substitute the model parameters $\mathcal{P}_{i_j}$ into the model;
3 | Perform the realization of the model with the parameters $\mathbb{A}_i$;
4 **end**
5 Calculate the mean, percentile 2.5 and percentile 97.5 of all the model outputs in the time instants t_1= Dec 2009, ..., t_9= Dec 2017;
6 Calculate $G(\mathbb{A})$, the 1-norm [9] of the difference between the mean, percentile 2.5 and percentile 97.5 calculated in Step 5 with the corresponding mean, percentile 2.5 and percentile 97.5 of the data in Table 1 and Table 3, respectively.

Now, we propose the Algorithm 3, a rPSO-inspired selection algorithm to select the sets of model parameters whose model-outputs best capture the data uncertainty.

The selection Algorithm 3, with a probability of 10%, rejects the current updated particle and generates randomly a new one. Also, with a probability of 10%, the current updated particle is mutated, where the mutation consists of changing some sets of model parameters in the current particle by others randomly chosen, avoiding repetitions. These features allow a deeper exploration of the space of parameters.

At this point, we should remark that, computationally, this probabilistic estimation procedure is less expensive than the one in [8], because the time headed to fit the model using rPSO will be smaller than the number of times we have to sample and fit the model in the probabilistic fitting technique developed in [8]. Also, the selection procedure is better because it allows us more flexible combinations to capture the data uncertainty.

Also, we must say that the random selection of the elements in the loop 10-25 of the Algorithm 3 uses to lead to a reduction in the number of elements of P_i, making difficult to reach a good fitting. Therefore, we suggest to take the same fixed given value every time the loop 10-25 is executed.

Algorithm 3: rPSO-inspired selection algorithm.

Input : $\mathcal{P}$, set of model parameters obtained by applying several times the rPSO algorithm; N, number of particles; $ITMAX$ the maximum number of iterations; T number of elements of the particles, $T \leq card(\mathcal{P})$.

Output: The sets of model parameters S^{best}_{global} that best capture the data uncertainty.

Define $S^{best}_{global} = \emptyset$ and $G(S^{best}_{global}) = +\infty$;
for $i \leftarrow 1$ **to** N **do**
    Initialize $S_i \subseteq \mathcal{P}$ with T elements chosen randomly without repetitions;
    Evaluate its fitness $G(S_i)$;
    Define its individual best fitness as $S^{best}_i = S_i$;
    if $G(S_i) < G(S^{best}_{global})$ **then**
        $S^{best}_{global} = S_i$
    end
end
for $i \leftarrow 1$ **to** $ITMAX$ **do**
    for $j \leftarrow 1$ **to** N **do**
        Build the new set $P_i = S_i \cup S^{best}_i \cup S_{global}$, that is, joining the current particle, its individual best and the global best;
        Remove the repeated elements;
        Build the new particle S_i as the random selection without repetition of T elements of P_i;
        With a probability of 10%, rejects the current S_i and generates randomly a new one;
        With a probability of 10%, the current S_i is mutated;
        Evaluate the fitness of the new S_i, $G(S_i)$;
        if $G(S_i) < G(S^{best}_i)$ **then**
            $S^{best}_i = S_i$
        end
        if $G(S_i) < G(S^{best}_{global})$ **then**
            $S^{best}_{global} = S_i$
        end
    end
end

4 Results

We have performed 30 different calibrations using rPSO, 10 using $N = 30$ particles, 10 using $N = 45$ particles and 10 using $N = 60$ particles, everyone with $N \times ITMAX = 3000$ evaluations, therefore, a total of $90,000$ evaluations of the model were performed. Some of them were withdrawn because their model output presented unrealistic oscillations, leaving $44,853$.

Our goal, now, is to find among the $44,853$ realizations of the model those such that the means and the 95% confidence intervals of these realizations be as much close as possible of the corresponding means and the 95% confidence intervals of the data in Tables 1 and 3.

Nevertheless, it would be interesting to reduce the number of eligible realizations to much less than $44,853$. There are 66 realizations with error less than 0.02, 359 with error less than 0.025, 755 with error less than 0.03 and 1590 with error than 0.04.

Then, we have performed $10,000$ evaluations of the Algorithm 3 with the realizations with error less than 0.025, 0.03 and 0.04, with $N = 30$, $N = 45$ and $N = 60$ particles, and selecting $T = 100$, $T = 150$ and $T = 200$ elements. Among all of them, the lowest error has been 0.07056 for the set of realizations with error less than 0.04 performed with $N = 45$ particles and selecting $T = 100$ elements. Substituting the $T = 100$ chosen model parameters values into the model and obtaining the model output, in Figure 2, we can assess visually the goodness-of-fit the probabilistic estimation.

Furthermore, the Figure 2 shows the probabilistic prediction over the next four years, on the right of the black dotted-dashed vertical line. The prediction preserves the trends drawn by the data, increasing for the frequent users of EC and decreasing otherwise. In the Tables 4 and 5, we can see the means and their 95% confidence intervals for the predictions from Dec 2018 to Dec 2021.

Date	Mean $n_1(t)$ $15-44$	Mean $y_1(t)$ $15-44$	Mean $n_2(t)$ $45-74$	Mean $y_2(t)$ $45-74$
Dec 2018	0.2454	0.2637	0.3643	0.1266
Dec 2019	0.2159	0.2867	0.3522	0.1452
Dec 2020	0.1874	0.3092	0.3382	0.1652
Dec 2021	0.1604	0.3304	0.3226	0.1866

Table 4: Mean of the probabilistic prediction from Dec 2018 to Dec 2021 for every subpopulation.

Now, taking the 100 sets of parameters selected, we can calculate, for each parameter, the mean and the 95% confidence interval, and the results are collected in Table 6.

Looking at Table 6, we can see that the model parameters p_1 and p_2 related to the people who make their own decisions to use frequently EC (innovators) are zero. Therefore, according to the results of our model, the use of the EC in Spain depends mainly on the transmission by peers (imitators) rather than the

Date	95% CI of $n_1(t)$ $15-44$	95% CI of $y_1(t)$ $15-44$	95% CI of $n_2(t)$ $45-74$	95% CI of $y_2(t)$ $45-74$
Dec 2018	$[0.2342, 0.2529]$	$[0.2544, 0.2721]$	$[0.3573, 0.3737]$	$[0.1207, 0.1345]$
Dec 2019	$[0.2035, 0.2255]$	$[0.2757, 0.2965]$	$[0.3439, 0.3623]$	$[0.1383, 0.1548]$
Dec 2020	$[0.1736, 0.2002]$	$[0.2954, 0.3207]$	$[0.3280, 0.3513]$	$[0.1568, 0.1770]$
Dec 2021	$[0.1453, 0.1766]$	$[0.3133, 0.3435]$	$[0.3108, 0.3371]$	$[0.1765, 0.2010]$

Table 5: 95% confidence interval of the probabilistic prediction from Dec 2018 to Dec 2021 for every subpopulation.

Parameters	Mean	95% CI
μ	0.00351	$[0.00351, 0.00355]$
c_1	0.04164	$[0.03172, 0.05430]$
d_1	0.00139	$[0.00033, 0.00269]$
d_2	0.05333	$[0.02882, 0.07809]$
p_1	0	$[0, 0]$
α_1	0.00376	$[0, 0.01726]$
α_2	0.06905	$[0.02934, 0.08545]$
γ_1	0.00011	$[0, 0.00182]$
p_2	0	$[0, 0]$
α_3	0.00011	$[0, 0.00065]$
α_4	0.019124	$[0.014733, 0.02262]$
γ_2	0.00011	$[0, 0.00231]$

Table 6: Mean and 95% confidence interval of the selected model parameters whose model outputs best capture the data uncertainty.

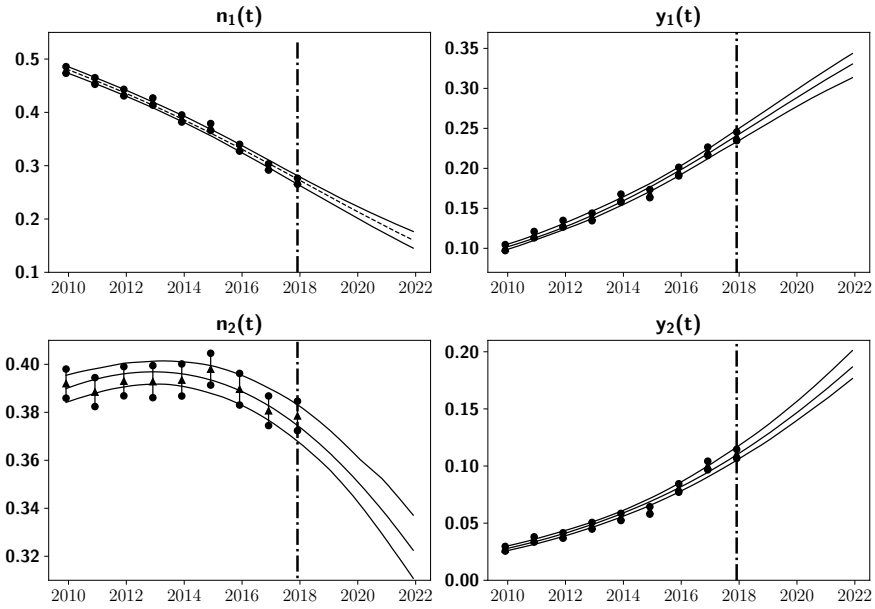

Figure 2: Probabilistic estimation and probabilistic prediction. Solid lines represent the model output 95% confidence bands and means, respectively. The dots are the mean of the data and the data 95% confidence intervals, respectively. As we can see, the model output band captures most of the data uncertainty represented by the dots. The dotted-dashed vertical lines separate the estimation (on the left) and the prediction (on the right) from Jan 2018 to Dec 2021.

own decisions (innovators).

Also, we can see that the model parameters γ_1 and γ_2 are very small, that is, when an individual uses frequently EC, it is not usual he/she gives it up for more than 3 months and moves to the non-user state.

Furthermore, for the transmission parameters, we have that α_2 and α_4 are greater than α_1 and α_3, respectively. This means that the group $y_2(t)$, frequent users of EC in the age group $45-74$, even being less people than $y_1(t)$, they influence more effectively to make the others to become frequent users of EC, regardless the age. Thus, elder frequent EC users are less people but more convincing.

Recently, the Spanish INE has released the data of EC frequent users corresponding to year 2018, [2]. These data with their CI 95% are collected in Table 7. Comparing with the model probabilistic prediction from Dec 2018, given in Table 5, we can see that our prediction is in agreement with the real data to Dec 2018, because the CI 95% for each subpopulation have intersection or are very close.

5 Conclusion

In this paper, we propose a mathematical epidemiological-type model with varying size population, based on a scaled system of difference equations, to study the dynamics of the frequent users of EC in Spain using real data retrieved from the INE (Spanish Statistical Institute).

Then, we propose a new technique to estimate probabilistically the model parameters in such a way the model is able to capture the data uncertainty. With the estimated model parameters we can perform a probabilistic monthly prediction over the next four years via the mean and the 95% confidence intervals each month from Jan 2018 to Dec 2021.

A deeper look to the model parameters selected to capture the best the data uncertainty, shows us some hidden behavior of the frequent and non frequent users of EC and how the habit of the frequent use EC transmission occurs. For instance: it is not usual the individuals get frequent users of EC by own decisions; when an individual uses frequently EC, he/she does not use to give it up; the elder users of EC are more convincing to make the others to change their mind and use EC.

Furthermore, the probabilistic prediction shows a sustained increasing in the subpopulations of frequent user of EC, reaching in Dec 2021 mean values around 33% and 18.5% of the total population for y_1 and y_2, respectively.

In comparison with the technique proposed in [8], this new technique is much less expensive and more accurate, taking advantage of all the evaluations performed during the rPSO fitting. However, we can not provide estimations of the model parameters, only the means and the 95% confidence interval.

Acknowledgments

This work has been partially supported by the Ministerio de Economía y Competitividad grant MTM2017-89664-P and by the European Union through the Operational Program of the European Regional Development Fund (ERDF) / European Social Fund(ESF) of the Valencian Community 2014-2020, grants GJIDI/2018/A/009 and GJIDI/2018/A/010.

References

[1] Spanish INE. Encuesta sobre equipamiento y uso de tecnologías de información y comunicación en los hogares (Survey on equipment and use of the information technologies and communication in the household, `http://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176741&menu=resultados&idp=1254735576692`.

[2] Spanish INE. Encuesta sobre equipamiento y uso de tecnologías de información y comunicación en los hogares (Survey on equipment and use of the information technologies and communication in

the household, https://www.ine.es/jaxi/Tabla.htm?path=/t25/p450/base_2011/a2018/l0/&file=02002.px&L=0.

[3] Spanish INE. Indicadores demográficos básicos (Basic demographic indicators), http://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177003&menu=resultados&idp=1254735573002.

[4] Bettencourt, L., Customer voluntary performance: Customers as partners in service delivery, *Journal of Retailing* **73** (1997) 383–406.

[5] Brauer, F. and Castillo-Chávez, C., *Mathematical Models in Population Biology and Epidemiology* (Springer New York, 2001).

[6] Christakis, N. A. and Fowler, J. H., *Conected. The surprising power of our social networks and how they shape our lives* (Little, Brown & Company, 2009).

[7] Cortés, J.-C., Lombana, I.-C., and Villanueva, R.-J., Age-structured mathematical modeling approach to short-term diffusion of electronic commerce in spain, *Mathematical and Computer Modelling* **52** (2010) 1045–1051.

[8] Cortés, J.-C., Santonja, F.-J., Tarazona, A.-C., Villanueva, R.-J., and Villanueva-Oller, J., A probabilistic estimation and prediction technique for dynamic continuous social science models: The evolution of the attitude of the basque country population towards ETA as a case study, *Applied Mathematics and Computation* **264** (2015) 13–20.

[9] Golub, G. and Van Loan, C., *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences (Johns Hopkins University Press, 1996).

[10] Hethcote, H. W., The mathematics of infectious diseases, *SIAM Review* **42** (2000) 599–653.

[11] Khemka, N. and Jacob, C., Exploratory toolkit for evolutionary and swarm-based optimization, *The Mathematica Journal* **11** (2010) 376–391.

[12] Li, Y. and Siming, Z., Competitive dynamics of e-commerce web sites, *Applied Mathematical Modelling* **31(5)** (2007) 912–919.

[13] Li, Y. and Zhu, S., Global analysis to a kind of competition model of e-commerce sites, *Annals of Differential Equations 3* (2003) 325–333.

[14] Mahajan, V., Muller, E., and Bass, F. M., New product diffusion models in marketing: A review and directions for research, in *Diffusion of Technologies and Social Behavior* (Springer Berlin Heidelberg, 1991), pp. 125–177.

[15] Turban, E., Outland, J., King, D., Lee, J., Liang, T.-P., and Turban, D., *Electronic Commerce 2018*, Series: Springer Texts in Business and Economics (Springer, 2018).

[16] Zhang, D., Ntoko, A., and Dong, J., Mathematical model of technology diffusion in developing countries, in *Applied Optimization* (Springer US, 2002), pp. 525–539.

t_{10} = Dec 2018 ($j = 10$)	$n_1(t_{10})$ $15 - 44$	$y_1(t_{10})$ $15 - 44$	$n_2(t_{10})$ $45 - 74$	$y_2(t_{10})$ $45 - 74$
Mean CI 95%	0.2279 $[0.2230, 0.0.2326]$	0.2570 $[0.2512, 0.2622]$	0.3746 $[0.3680, 0.3805]$	0.1405 $[0.1366, 0.1447]$

Table 7: Data released for 2018. Mean and CI 95% of people who have used EC in the last 3 months (frequent users of EC) and the people who do not (non-frequent users of EC), for the age groups 15-44 and 45-74 during the year 2018 in Spain [2].