Uncertainty quantification in dynamical models. An application to cocaine consumption in Spain.

PhD Dissertation

Presented by: María Rubio Monzó

Supervisor: Dr. Francisco José Santonja Gómez
Dr. Rafael J. Villanueva Micó

Department of Applied Mathematics
Universidad Politécnica de Valencia **July 2015**

Francisco José Santonja Gómez, professor at the Universidad de Valencia, and Rafael Jacinto Villanueva Micó, professor at the Universidad Politécnica de Valencia,

Certify that the present thesis *Uncertainty quantification in dynamical models. An application to cocaine consumption in Spain* has been directed under our supervision in the Department of Applied Mathematics of the Universidad Politécnica de Valencia by María Rubio Monzó and makes up her thesis to obtain the doctorate in Mathematics.

As stated in the report, in compliance with the current legislation, we authorize the presentation of the above PhD. Dissertation before the doctoral commission of the Universidad Politécnica de Valencia, signing the present certificate

Valencia, July 2015.

Francisco José Santonja Gómez Rafael Jacinto Villanueva Micó

Acknowledgments

Firstly, I am grateful to my advisers, Francisco J. Santonja Gómez and Rafael J. Villanueva Micó. I want to thank them not only for their supervision, but also for their valuable contributions concerning my work and their advising during years to develop my ph.D. Thesis. Sincere thanks are due to Francisco J. Santonja Gómez, by his dedication and implication in this work.

I greatly appreciate the technical support of Emilio Sánchez Hervás, clinical psychologist of the Unidad de Conductas Adictivas (Conselleria de Sanitat) of Catarroja, Valencia, Spain. His technical assistance about cocaine consumption has been invaluable.

Finally, I owe a very special debt of gratitude to my fiancé Juanjo, to my parents, Julia and Enrique, and to my brother Quique. Without their constant encouragement and patience this work would not have been completed.

Abstract

The present Ph.D. Thesis considers epidemiological mathematical models based on ordinary differential equations and shows its application to understand the cocaine consumption epidemic in Spain. Three mathematical models are presented to predict the evolution of the epidemic in the near future in order to select the model that best reflects the data. By the results obtained for the selected model, if there are not changes in cocaine consumption policies or in the economic environment, the cocaine consumption will increase in Spain over the next few years. Furthermore, we use different techniques to estimate 95% confidence intervals and, consequently, quantify the uncertainty in the predictions. In addition, using several techniques, we conducted a model sensitivity analysis to determine which parameters are those that most influence the cocaine consumption in Spain. These analysis reveal that prevention actions on cocaine consumer population can be the most effective strategy to control this trend.

Resumen

La presente Tesis considera modelos matemáticos epidemiológicos basados en ecuaciones diferenciales ordinarias y muestra su aplicación para entender la epidemia del consumo de cocaína en España. Se presentan tres modelos matemáticos para predecir la evolución de dicha epidemia en un futuro próximo, con el objetivo de seleccionar el modelo que mejor refleja los datos. Por los resultados obtenidos para el modelo seleccionado, si no hay cambios en las políticas de consumo de cocaína ni en el ámbito económico, el consumo de cocaína aumentará en los próximos años. Además, utilizamos diferentes técnicas para estimar los intervalos de confianza al 95% y, de esta forma, cuantificar la incertidumbre en las predicciones. Finalmente, utilizando diferentes técnicas, hemos realizado un análisis de sensibilidad para determinar qué parámetros son los que más influyen en el consumo de cocaína. Estos análisis revelan que las acciones de prevención sobre la población de consumidores de cocaína pueden ser la estrategia más efectiva para controlar esta tendencia.

Resum

La present Tesi considera models matemàtics epidemiològics basats en equacions diferencials ordinàries i mostra la seua aplicació per a entendre l'epidèmia del consum de cocaïna en Espanya. Es presenten tres models matemàtics per a predir l'evolució d'aquesta epidèmia en un futur pròxim, amb l'objectiu de seleccionar el model que millor reflecteix les dades. Pels resultats obtinguts per al model seleccionat, si no hi ha canvis en les polítiques de consum de cocaïna ni en l'àmbit econòmic, el consum de cocaïna augmentarà en els pròxims anys. A més, utilitzem diferents tècniques per a estimar els intervals de confiança al 95% i, d'aquesta manera, quantificar la incertesa en les prediccions. Finalment, utilitzant diferents tècniques, hem realitzat un anàlisi de sensibilitat per a determinar quins paràmetres són els que més influencien el consum de cocaïna. Aquestos anàlisis revelen que les accions de prevenció en la població de consumidors de cocaïna poden ser l'estratègia més efectiva per a controlar aquesta tendència.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Cocaine consumption is growing at a worrying rate in developed and developing countries [23, 80]. In Spain it is becoming a serious problem not only from an individual health point of view but also from the public socioeconomic one [61, 62].

Thus, it is in the interest of public health to study the dynamics of cocaine consumption. In this dissertation, we analyse the evolution of people with habitual cocaine consumption in Spain and simulate some health policy proposals and their effect in reducing this population.

Spanish Government strategy on drug abuse appears in the National Plan on Drugs [9, 61], issued by the Spanish Health Ministry. The objectives mentioned in this document are:

1. The prevention of drug consumption, pointing out the health concerns produced by their consumption, delaying the age of the first contact with drugs, education programmes and legal fight against drugs dealing.

2. To improve quantitative and qualitative research, to implement new treatments, evaluate current therapy programmes and training to increase professional competence of the people who work with drug abusers.

In this dissertation, we take cocaine consumption to be as socially transmitted epidemic disease. We treat cocaine consumption as a disease that spreads through social peer pressure or social contact. These social contacts have an influence on the probability of transmission of cocaine consumption. These facts lead us to propose an epidemiological-type model to study the evolution of this consumption. This type of mathematical models have also been used in the study of other drug addictions, such as alcohol, tobacco, ecstasy or heroin addiction [25, 75, 77, 82] and in the approach to other sociological topics that are spread by social contact as obesity or extreme ideological behaviour [37,71].

1.1 Epidemiological models

The spread of infectious diseases has always been a potential public health concern and it has influenced the economic and social development of the human society. Thus, its prevention and control become extremely important.

Epidemiological models are an interesting approach to understand the transmission dynamics of infectious diseases. These models and their numerical simulations allow us to make reliable predictions, identify the most important and sensitive parameters and help to improve prevention and control strategies. Understanding the dynamics of the spread of infectious diseases can lead to better approaches to decrease the transmission of these diseases [7,46].

Investigating and controlling infectious diseases is a complex task that has long been carried out by mathematical modeling. Although it can go back to 1760 when Bernoulli used mathematical models for smallpox, the research of infectious diseases, using deterministic mathematical models, actually began in the 20th century. In 1906 a discrete time model for the spread of measles was proposed by Hamer, which may have been the first model which assumed that the incidence (number of new cases per unit time) depends on the product of the densities of the susceptibles (individuals who might become infected if they are exposed) and infectives [31]. In 1911 a differential equation model was used by the physician Dr. Ross to describe the transmissions of malaria. In 1926 a compartmental model proposed by Kermack and McKendrick established the foundations of the theory of epidemic dynamics: the *SIR* model [31,40]. In the Kermack and McKendrick's work, epidemics such as the plague (London 1665 - 1666, Bombay 1906) and cholera (London 1865) were studied. Additionally, they obtained the epidemic threshold such that if the density of susceptibles exceeds this threshold, an epidemic outbreak occurs.

Epidemic models based on differential equations are mostly based on dividing the population into different groups or compartments, depending on their status with respect to the disease. In the Kermack and McKendrick SIR model, the population is divided into three compartments: susceptibles (S), individuals who might become infected if they are exposed; infected (I), individuals who are infected and can transmit the disease; and removed (R), individuals who are removed or recovered from the infection and they cannot transmit the disease.

In this standard model, it is assumed that N=S+I+R, where N is the total population. The transmission rate, β, is proportional to the total number of susceptibles. Therefore, we model this contact rate with the non-lineal term $\beta SI/N$. Additionally, the recovery rate, γ, is proportional to the size of subpopulation I. All the individuals in R remain immunes.

Based on these assumptions, the flow diagram of the SIR model is shown in Figure 1.1.

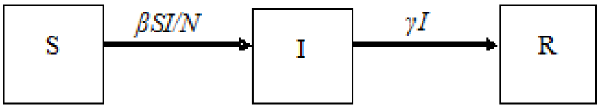


Figure 1.1: Flow diagram of the SIR mathematical model.

The corresponding model equations are given by the following system of differential equations:

$$\begin{aligned} S'(t) &= -\beta S(t)I(t)/N(t) \\ I'(t) &= \beta S(t)I(t)/N(t) - \gamma I(t) \\ R'(t) &= \gamma I(t), \end{aligned}$$

where the non-lineal term $\beta S(t)I(t)/N(t)$ models the disease transmission.

This celebrated SIR model has been used and extended to model infectious diseases, for example, exposed periods, vaccinations, isolations, quarantines, contact rates, vertical and vectors transmissions. In addition, other more complex models with ages, gender or spatial structure have also been studied [7, 46].

Mathematical epidemiology has grown exponentially in the last years and multitude of mathematical models have been formulated and developed to study several infectious diseases, such as measles, malaria, tuberculosis, sexually transmitted diseases (STDs), AID/HIV, etc. Recently, systems of ordinary differential equations are being used to study social epidemics, as alcoholism, smoking and drug abuse, and in the approach of other social topics that are spread by social contact, such as obesity [72], suicide or extreme ideological behaviour [71].

1.2 Epidemiological models of drug abuse

Mathematical models, simpler than the reality, allow us to understand the global dynamic behaviour of the drug consumption in the population and to help the policymakers in targeting drug prevention and treatment.

Considering drug consumption as a disease that spreads through social peer or social contact, some epidemiological-type models have been used to study the evolution of drugs consumption, such as alcohol, heroin, tobacco, ecstasy or cocaine. While social problems, such as alcohol and drug use, have been referred to in terms of epidemics, only few have been published on the application of mathematical modeling methods to such problems. In this dissertation, we will focus on the models defined by ordinary differential equations (ODE).

Agent-based modeling and other computer-based simulations have been used broadly in the Social Sciences since the 1990s as a means of understanding social processes and dynamics. In 2006, Gorman et al. developed a preliminary agent-based simulation model to explore both the social dynamics and the environmental influences that affect drinking behaviour [25]. Specifically, this model examined the iterations of three types of agents defined according to their current status (susceptible nondrinkers, current drinkers and formal drinkers) as well as what happens to these interactions when a new bar was introduced into the environment. Within this framework, the basic model shows that even a single current drinker introduced into a population of susceptibles could convert the population into drinkers over time. However, including a new bar in the model both enhanced and soften the rate of propagation, changing the dynamics.

In 2008, Sánchez et al. proposed an epidemiological model to study the dynamics of drinking behaviour. Building on similarities between drinking and "infection", a simple SDR (Susceptibles- Drinkers- Removed) model was described within the context of the classic SIR epidemiological model [70]. From the analysis of the "drinking-free" equilibrium, that is, the state where drinking is not part of the culture, the model's basic reproductive number, R_0, and the basic reproductive number with recovery (R_ϕ) and relapse (R_ρ) , were computed and an uncertainty and sensitivity analysis was carried out. The basic reproductive number tells us how many secondary infections will result from the introduction of one infected individual into a susceptible population. The threshold value of R_0 indicates under what circumstances an epidemic will be avoided and if this does not happen, an endemic equilibrium of drug-users may be established in the population and eradication of drug-use could become more difficult. For more details related to R_0 and R_ϕ, see [59, 70, 82]. In this work [70], conditions for the “successful” invasion of a “non-drinking” culture were established. Model results were used to highlight the impact of nonlinear interactions on the dynamics of the alcohol use at the population level. The public health and social policy implications were discussed. It was shown, for example, that if a drinking culture is established, it is difficult to bring it to a level low enough to eliminate it completely. Moreover, this model revealed that the basic reproductive number R_ϕ (as a function of treatment) is not always the key. In fact, it may be more effective to try to limit the average residence times of susceptible individuals in drinking environments until treatments with more sustained effects are identified and widely implemented.

Recently, drinking has also been modeled as a socially contagious process in low-and high-risk connected environments [57]. A simple compartmental model with two distinct drinking environments (low- and high- risk) and three classes of drinkers (susceptible, moderate and heavy drinkers) was employed to examine the effects of residence time upon the differential development and persistence of heavy drinking. A threshold parameter, R_d (referred to as the drinking or basic reproduction number) was computed. The parameter R_d gives the average number of individual transitions from light to moderate drinking that result from the introduction of a moderate drinker in a population of light drinkers. This model shown that the increase on individual residence times in the community or increase in the rate of progression to heavy drinking can lead to increase in the proportion of heavy drinkers. It pays attention to the need to focus on reducing the risk of progression in the high-risk environment and/or in placing or creating structures that limit the “contagion” facilitated in these settings.

Furthermore, a stochastic model was derived from the deterministic formulation described in [70] by Cintrón-Arias et al. [15]. The goal of this work was to quantify the variability on drinking dynamics due to stochastic effects; intending to highlight some of the differences and similarities between the deterministic and stochastic approaches. Numerical simulations, which were obtained using known parameter values, were carried out on both models (deterministic and stochastic model). From sensitivity analysis, they concluded that the increase in the recovery rate, as well as reductions in the relapse rate, have a positive effect. Therefore, they suggest promoting reductions in the number of problematic drinkers. It was concluded it is in the best interest of treatment programmes to concentrate efforts preventing temporarily recovered individuals from relapse into drinking [15].

Another interesting contribution was the system of differential equations proposed in 2006 by Song et al. to model the peer-driven dynamics of ecstasy use. After an analytical study of the model, the most influence factors on ecstasy use were identified. Through analysis of threshold conditions and estimation of parameters, predictions for the future of ecstasy use in the United States were made. Finally, all the parameters were varied with the aim to predict the most efficient manner of decreasing ecstasy use by means of education [77]. This model shows that, once a considerable number of people use ecstasy, decrease this number is extremely difficult. In other words, a peer-driven drug epidemic should be avoided at all costs.

Moreover, mathematical studies for assessing the dynamics of smoking have also been conducted. In 1997 Castillo-Garsow et al. presented a general epidemiological model (SDR model) to describe the dynamics of drug use among adolescents, specifically tobacco use. Specific models are derived by considering other factors that have been identified to have an effect on the growing trend of tobacco use. The factors considered are peer pressure, relapse, counseling and treatment [11]. In this work, the parameters of the models were estimated and a rough approximation of the basic reproductive number R_0 was determined. Based on these parameters, some simulations were performed. It clearly points out the importance of educational (preventive) measures against drug abuse.

Later, a slight refinement of this model was presented to account for variability in smoking frequency, by introducing two classes of mild and chain smokers, as well as the public health impact of smoking-related illnesses [75]. This study shows that smoking and smoking-related illnesses can be effectively controlled in a community if public health related to a threshold value known as the smokers generation number, less than unity, are implemented. Recently, another epidemic-type mathematical model has been developed to study the evolution of tobacco use in Spain, with the aim to quantify the effect of the smoke-free law [27].

One of the first ODE models to opiate addiction, based on the principles of mathematical epidemiology, was presented by White and Comiskey in 2007 [82]. Following standard methods, the population was divided into three classes, namely susceptibles, heroin users and heroin users undergoing treatment. In this work they identified parameters of interest in the drug abuse dynamics and proposed an epidemic threshold value, R_0, the basic reproductive number. Sensitivity analysis was performed on R_0 and it was then used to examine the stability of the system. A key result arising from this model is that prevention is indeed better than cure; efforts to increase prevention are more effective in controlling the spread of habitual drug use than efforts to increase the numbers of individuals accessing treatment.

Furthermore, this ODE model was revisited by Mulone and Straughan, who proved that the positive equilibrium of the White and Comiskey model of heroin epidemics is stable under the realistic condition that the relapse rate of those in treatment returning to untreated drug use is greater than the prevalence rate of susceptibles becoming drug users [58], that is, an epidemic in heroin use will never occur, under this condition.

Moreover, Battista introduced a new mathematical model with more than three classes of people and compared it with the White and Comiskey's model [2], showing the existence of stable equilibrium for both of them (the White and Comiskey model and the Battista model), suggesting a situation where heroin use can be eradicated (existence of ideal equilibrium) and a situation where drug use remains in society (existence of endemic steady-states), depending on the values of parameters.

More recently, the White and Comiskey model has been modified to develop a heroin epidemic model with distributed time delays [43, 68]. In [43] the restriction where the total population is constant was deleted, a delay effect in those returning to untreated drug taking from a treatment programme was included, and finally a delay model was developed. In [68], time dependent parameters, time dependent total population size and distributed time delay to become heroin user have been introduced. In both works, parameters of interest and the basic reproduction number R_0 for the model and its threshold property (conditions for permanence and extinction of the heroin use) have been identified.

In addition, the White and Comiskey model of heroin epidemics has been modified to model the dynamics of methamphetamine use in a South African province [63]. The stabilities of the model equilibria have been ascertained and persistence conditions established. Furthermore, numerical simulations were performed and the implications of the results in drug policy, treatment and prevention were discussed.

Dynamical models have also been used to study the prevalence for the cocaine epidemic in the US. Toward the goal of designing effective drug control policy, in 1994 Everingham and Rydell created a model of how cocaine demand changes over time using available data from the NHSDA (National Household Survey of Drug Abuse) and other sources. They set up a Markovian model of population flows that differentiates between light and heavy users [24]. It is referred to as the LH model. Even though this model is not well suited for modeling epidemics on the macro scale, this approach is useful for short- and intermediate- range prevalence estimation. Although this model by itself is not predictive for the future evolution of cocaine epidemic, it can project the evolution of epidemic given any hypothetical scenario. In 2004 Everingham and Rydell's Markov chain model of cocaine demand was modified and updated in light of recent data [12]. The time-continuous version of the time-discrete Everingham and Rydell's model for the evolution of cocaine consumption was developed by Behrens et al. (1999) and extended by introducing an endogenous function of prevalence as initial condition. They carried out a sensitivity analysis to understand how variations in the parameters affect both the system and the levels of drug consumption and obtained significant results regarding the optimum allocation of resources for treatment and prevention, where the objective is to minimize social and control costs [3]. Focused on the LH model, in 2004 Kaya studied how to bring down the prevalence of drug use to a target as soon as possible [39].

More recently, instead of using the LH model, Caulkins et al. have used a SA model. The SA model is much like the classic SIR models of infectious diseases, with the difference that recovered users (R) are not modeled explicitly. It has been parameterized for two different countries and used to simulate what might happen if the policy was changed from use reduction to harm reduction or vice versa [13].

1.3 Uncertainty models

The translation from a real problem to a set of differential equations is a complex task. Not only because of the difficulty of design a mathematical model, but by the combination of the uncertainties involved.

Traditionally, the formal modeling of systems has been done via mathematical models, which attempt to find analytical solutions enabling the prediction of the behaviour of the system from a set of parameters and initial conditions. Computer simulation is often used as an adjunct to, or substitution for, modeling systems for which simple closed form analytic solutions are not possible.

Scientific computing plays an ever-growing role in predicting the behaviour of natural and engineered systems. However, whereas the simulations are generally deterministic in nature, applications are steeped in uncertainty arising from a number of sources such as those due to manufacturing processes, natural material variability, initial conditions, conditions of the system, and the system surroundings. Furthermore, the modeling and simulation process itself introduces uncertainty related to the form of the model as well as the numerical approximations employed in the simulations. Each of these different sources of uncertainty must be estimated and included in order to estimate the total uncertainty in a simulation prediction. In addition, an understanding of the sources of the uncertainty can provide guidance on how to reduce uncertainty in the prediction in the most efficient and cost-effective way.

The flourishing of simulation-based scientific discovery has also resulted in the emergence of the verification and validation (V&V) and uncertainty quantification (UQ) disciplines. The goal of these emerging disciplines is to enable scientists to make precise statements about the degree of confidence they have in their simulation-based predictions. Here we focus on the UQ discipline which is essential for validating and verifying computer models.

The main task of uncertainty quantification is to define and quantitatively describe these uncertainties. It can be defined as the

- identification (Where are the uncertainties?),
- characterization (Which form they have?),
- propagation (How do they evolve during the simulation?), analysis (How do they influence?), and reduction

of all uncertainties in simulation models.

Identification

For a complete uncertainty quantification framework, all of the possible sources of uncertainty must be identified and characterized. Sources of uncertainty can be broadly categorized as occurring in model inputs, numerical approximations or in the form of the mathematical model.

A. Model Inputs. Model inputs include not only parameters used in the model of the system, but also data from the surroundings. Model input data includes things such as geometry, constitutive model parameters, initial conditions, and can come from a range of sources including experimental measurement, theory, other supporting simulations or even expert opinions. Data from the surrounding includes boundary conditions and system excitation (mechanical forces or moments acting on the system, forcing fields such as gravity and electromagnetism, etc.).

B. Numerical Approximation. Since complex differential equation-based models rarely admit exact solutions for practical problems, approximate numerical solutions must be used. The characterization of the numerical approximation errors associated with a simulation is called verification. It includes discretization error, iterative convergence error, round-off error and also errors due to coding mistakes.

C. Model Form. The model results come from all assumptions, conceptualizations, abstractions and mathematical formulations on which the model relies such as ignored physics or physics coupling in the model. Model form uncertainties are quantified using model validation procedures which include a comparison of model predictions to experimental data and the extrapolation of this uncertainty structure to points in the application domain where experimental data do not exist.

Characterization

While there are many different ways to classify uncertainty, we will use the taxonomy prevalent in the risk assessment community which categorizes uncertainties according to their fundamental essence [16, 55, 81]. Thus, uncertainty is classified as either aleatory or epistemic. All of these sources of uncertainty can be classified as either purely aleatory, purely epistemic, or a mixture of aleatory and epistemic uncertainty.

A. Aleatory Uncertainty is due to inherent variation or randomness and can occur among members of a population or due to spatial or temporal variations. Aleatory uncertainty is generally characterized by a probability distribution.

B. Epistemic Uncertainty arises due to a lack of knowledge on the part of the analyst conducting the modeling and simulation. Epistemic uncertainty is traditionally represented as either an interval with no associated probability distribution or a probability distribution which represents degree of belief of the analyst, as opposed to frequency of occurrence discussed in aleatory uncertainty.

Propagation

Propagation of uncertainties can be performed forward and backward involving:

A. Analysis and quantification of the overall uncertainty in model outputs. At the beginning this involves taking into account all sources of uncertainty, although it has often been interpreted in the past more narrowly as quantifying the uncertainty in outputs due to input uncertainty (Uncertainty Analysis).

B. The finding of the major sources of uncertainties (sensitivity analysis), i.e., identifying which parameters are the most relevant in contributing to uncertainty in the prediction.

C. The determination of parameter posterior distributions based on data (calibration/data fusion).

D. The exploration of “interesting” regions in the parameter space (model exploration).

Several theories address the definition of uncertainty. These theories include probability theory [26, 30], fuzzy set theory [84] and evidence theory [1, 56]. In this dissertation, we work under the framework of probability theory, which provides a solid and comprehensive theoretical foundation and offers the most versatile statistical tools.

The methods involved in this step include statistical analysis, experimental error analysis and often, expert judgment [22]. Although how to quantify model uncertainties and numerical uncertainties is still a topic of current research [19, 83], there are successful examples of quantifying the uncertainty sources in very complex engineering systems [5, 67].

Once the sources of uncertainties are quantified, we need to calculate how these uncertainties propagate through the simulation to the quantities of interest. Many techniques exist for propagating input uncertainties through the mathematical model to obtain uncertainties in the system response quantities (SRQs). Sampling techniques (e.g., Monte Carlo sampling or Latin hypercube sampling) are the most common techniques to propagate input uncertainties through the model.

1.4 Overview of this dissertation

In the present dissertation we will present a mathematical epidemiological model to study the dynamics of cocaine consumption in Spain. This model is based on differential equations, within the context on the classic SIRS epidemiological model, considering different levels in the drug abuse. Moreover, we predict the consumption trends over the next few years. First, we will obtain a deterministic prediction of the mathematical model and then, using different methods, we will introduce uncertainty in the model. This fact allows us to evaluate how uncertainty in the parameters influences uncertainty in the model and to obtain credible intervals for the cocaine consumption prevalence in Spain over the next few years. Furthermore, we carry out a sensitivity analysis of the mathematical model to determine which parameters are those that most influence the model in order to propose some public health strategies to reduce cocaine consumption.

The methods used in order to evaluate how uncertainty in the parameters influences uncertainty in the model output are the following:

- *Monte Carlo method.* This method provides approximate solutions to a variety of mathematical problems by performing statistical sampling experiments.

  A Monte Carlo simulation is based on performing multiple model evaluations using random or pseudo-random numbers to sample from probability distributions of model inputs, depending on a priori information. The results of these evaluations can be used to both evaluate the uncertainty in model output and perform a sensitivity analysis to identify critical inputs of the model.

  This method proceeds as follows:

  1. Generate sample values of the inputs of the model (parameters and initial conditions) from their assumed probability density function.

2. Solve the deterministic system of differential equations corresponding to each value (or group of values).
3. Calculate mean and variance of the solutions set.

To perform uncertainty analysis we implemented the LHS algorithm. LHS belongs to the Monte Carlo stratified sampling methods. It is a sophisticated and efficient method for achieving equitable sampling of all input parameters simultaneously. For each parameter a probability density function is defined and stratified into N equal probability intervals. A single value is then selected randomly from every interval and this is done for every parameter. In this way, each interval for each parameter is sampled exactly once (without replacement). Thus, the entire range for each parameter is explored. Distributions of the outcome variables can then be derived directly by running the model N times with each of the sampled parameters set [32, 49].

- *Approximate Bayesian computation (ABC method).* This method has been conceived to infer posterior distribution in case where likelihood functions are computationally intractable or too costly to evaluate. In ABC methods, the evaluation of the likelihood is replaced by a simulation-based procedure.

  The ABC methods have the following generic form:

  1. Sample a parameter vector from some proposal distribution.
  2. Simulate a dataset from the model described by a conditional probability distribution.
  3. Compare the simulated dataset with the experimental data, using a distance function and tolerance (that is the desired level of agreement between simulated dataset and experimental data).

- *Bootstrap method.* The goal of bootstrap confidence interval theory is to calculate confidence limits for the parameters from their distribution (density function). Bootstrapping can handle violations of the maintained assumptions of traditional methods. The main idea behind bootstrapping is resampling. Roughly the idea is as follows: to construct many “new” data sets by resampling the original data set, and estimate the parameter value(s) for each of these “new” data sets, generating a distribution of parameter estimates. Using the resulting empirical distribution of parameters, estimate the confidence intervals.

The residual method used to study error terms for the estimated parameters is implemented as follows:

1. Fit the model to the actual data by optimizing the parameter values.
2. Compute the error terms for the optimum parameter values. An error term is actual data minus model output.
3. Resample the error terms using a parametric technique and obtain new error values.
4. Construct new perturbed data sets by adding the resampled error terms to the model output.
5. For each new set of perturbed data, compute the parameters which best fit the model with the new set of perturbed data.

   Repeat steps 3, 4 and 5 many times to obtain a sufficiently large bootstrap sample for the parameters of the model.

In this dissertation, we model the dynamics of cocaine consumption in Spain with differential equations considering uncertainty in the parameters. Note that we only consider uncertainty in parameters and not in the initial condition because the initial condition of the model is determined using a representative sample of the Spanish population. Therefore, we assume that it is known without uncertainty and it is not considered as a random variable. To quantify uncertainties, first we identify all of them, considering uncertainties occurring in model inputs (parameters estimations), numerical approximations (characterization of the numerical approximation error) and model form (using model selection). We characterize all the uncertainties considered as epistemic. Then, we will focus on the propagation of uncertainties through the model, using the methods described above (LHS, ABC and Bootstrap). The proposed UQ framework is used to predict an average behaviour and obtain a confidence interval for cocaine consumption in Spain over the next few years.

Chapter 2

Predicting cocaine consumption in Spain: A mathematical modeling approach

In this Chapter [1], we present an epidemiological-type mathematical model to analyse the evolution of cocaine consumption in Spain. Also we predict the consumption trends over the next few years. First, we will obtain a deterministic prediction of the mathematical model and then, using the LHS method, we will introduce uncertainty in the model. This fact allows us to obtain credible intervals for the cocaine consumption prevalence in Spain. Furthermore, we carry out a sensitivity analysis of the mathematical model in order to propose some public health strategies to reduce cocaine consumption.

2.1 Introduction

Cocaine consumption is growing at a worrying rate in developed and developing countries [23, 80]. In Spain, it is becoming a serious problem not only from an individual health point of view but also from the public socioeconomic one [61, 62]. Table 2.1 shows the prevalence rates for the last years. We notice that cocaine consumption is increasing from 1995 to 2007.

[1] The contents of this chapter have been published in E. Sánchez, R. J. Villanueva, F.J. Santonja, and M. Rubio: Predicting cocaine consumption in Spain: a mathematical modeling approach. *Drugs: Education, Prevention and Policy*, 18(2), 108-115, 2011.

Table 2.1: Evolution of the proportion of non-consumers (N), occasional consumers (C_o), regular consumers (C_r) and habitual consumer subpopulations (C_b) for different years. The data have been obtained from the Drug National Observatory Reports [53, 54, 61, 62].

Year	N	C_o	C_r	C_b
1995	0.944	0.034	0.018	0.004
1997	0.948	0.032	0.015	0.005
1999	0.948	0.031	0.015	0.006
2001	0.911	0.049	0.026	0.014
2003	0.903	0.059	0.027	0.011
2005	0.884	0.070	0.030	0.016
2007	0.874	0.080	0.030	0.016

2.2 Methods

Recently, it has been shown how our social contacts shape our life. Everything we feel, do or say may spread through the people we know and conversely, the people we know influence the way we are [14]. Following this idea, we are going to consider that cocaine consumption is a behaviour or habit that may be transmitted socially. Thus, it can be treat as a disease that spreads through social peer pressure or social contact. These social contacts have an influence on the probability of transmission of cocaine consumption. These facts lead us to propose an epidemiological-type mathematical model to study the evolution of this consumption. This type of mathematical models have also been used in the study of other drug addictions (alcohol, tobacco, ecstasy, etc.) and other social topics (obesity, extreme ideological behaviour, etc.), as we discussed in the Section 1.2.

2.2.1 Mathematical model

2.2.1.1 Building the model

In order to build the mathematical model, the 15-64-year-old Spanish population is considered and divided into four subpopulations, according to the classification defined by the Spanish Health Ministry [53, 54, 61, 62]:

- *N(t): Non-consumers*, individuals who have never consumed cocaine.
- $C_o(t)$*: Occasional consumers*, individuals who have consumed sometimes in their life.
- $C_r(t)$*: Regular consumers*, individuals who have consumed in the last year.
- $C_b(t)$*: Habitual consumers*, individuals who have consumed in the last month.

Furthermore, we consider the following assumptions:

1. Let us assume population homogeneous mixing. That is, each individual may transmit the consumption habit to any other one [59].
2. The transitions between the different subpopulations can be modeled as follows:
   (a) Let us consider that the newly recruited 15-year-old individuals become members of the *N(t)* subpopulation, i.e., we consider that they have never consumed cocaine before.
   (b) Once an individual begins cocaine consumption, he/she becomes an occasional consumer, $C_o(t)$. If this person increases the cocaine consumption he/she may become a regular consumer, $C_r(t)$. If this individual continues with his/her consumption, he/she may become a habitual consumer, $C_b(t)$.
   (c) An individual of subpopulation $C_b(t)$ becomes a member of subpopulation *N(t)*, non-consumer subpopulation, if he/she decides to give up the cocaine consumption and go into therapy. This reasoning is based on the assumption given by expert clinicians after patients go into therapy. In this work we have considered therapy as the only way to reduce cocaine consumption once patient became aware of its problem. This detail was proposed by the expert, the clinical psychologist.
   (d) An individual in *N(t)* transits to $C_o(t)$ because people in $C_o(t)$, $C_r(t)$ or $C_b(t)$ transmit cocaine consumption habit by social contact at rate β. Therefore, the contagion is a non-linear term modeled by $\beta N(t)(C_o(t) + C_r(t) + C_b(t))$. The remainder transits are governed by terms proportional to the sizes of the subpopulations:
      i. $\gamma C_o(t)$ to transit from $C_o(t)$ to $C_r(t)$,

ii. $\sigma C_r(t)$ to transit from $C_r(t)$ to $C_b(t)$,

iii. $\varepsilon C_b(t)$ to transit from $C_b(t)$ to N(t).

Using the above assumptions, a dynamic cocaine consumption model for Spanish population is given by the following non-linear system of ordinary differential equations (t, time in years):

$$\begin{aligned}
N'(t) &= \mu P(t) - d_N N(t) - \beta \frac{N(t)(C_o(t)+C_r(t)+C_b(t))}{P(t)} + \varepsilon C_b(t) && (2.1)\\
C_o'(t) &= \beta \frac{N(t)(C_o(t)+C_r(t)+C_b(t))}{P(t)} - d_C C_o(t) - \gamma C_o(t) && (2.2)\\
C_r'(t) &= \gamma C_o(t) - d_C C_r(t) - \sigma C_r(t) && (2.3)\\
C_b'(t) &= \sigma C_r(t) - d_C C_b(t) - \varepsilon C_b(t) && (2.4)\\
P(t) &= N(t) + C_o(t) + C_r(t) + C_b(t)
\end{aligned}$$

where the parameters of the model are:

- μ, rate of people entering into the system. Taking into account the low mortality rates of people younger than 15, we are going to suppose that this rate is the birth rate in Spain.

- d_N, death rate in Spain.

- d_C, death rate for cocaine consumers. Obviously, this death rate is higher than death rate for non-consumers.

- β, transmission rate due to social pressure to consume cocaine.

- γ, rate at which an occasional consumer transits to the regular consumption subpopulation.

- σ, rate at which a regular consumer transits to the habitual consumption subpopulation.

- ε, rate at which a habitual consumer goes into therapy and becomes a non-consumer.

Figure 2.1 shows the diagram for the evolution dynamics of cocaine consumption in Spain. The boxes represent the subpopulations and the arrows represent the transitions between the subpopulations. Arrows are labeled by their corresponding model transition terms.

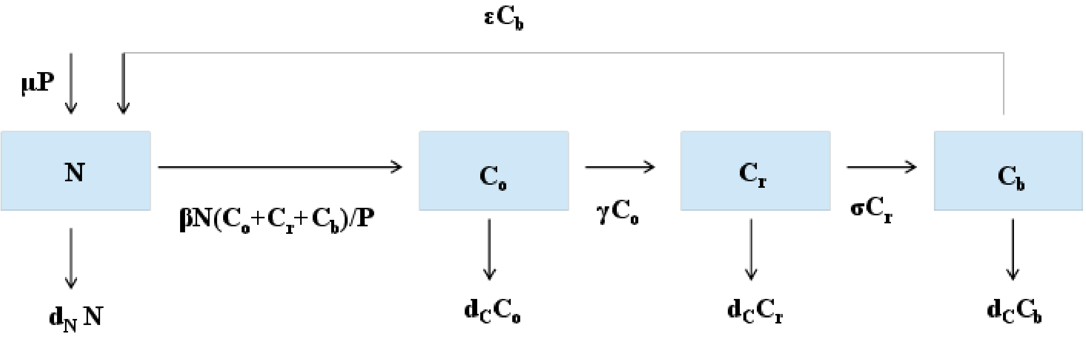


Figure 2.1: Flow diagram of the mathematical model for the dynamics of cocaine consumption in Spain.

2.2.1.2 Scaling the model

Data shown in Table 2.1 are related to the percentages of population, but the equations shown above are related to the number of individuals. It leads us to transform the model into the same units as data. Hence, we scale the model in order to estimate the unknown parameters by fitting it with data in Table 2.1. Parameters estimation is shown in the next section.

Following the ideas developed in [33, 36, 50] about how to scale models where the size of the population depends on the time, we are going to obtain the equations of the scaled model.

If we add the four equations of the model, eqs. (2.1)–(2.4), we get:

$$P'(t) = \mu P(t) - d_N N(t) - d_C C_o(t) - d_C C_r(t) - d_C C_b(t). \tag{2.5}$$

Dividing both members by $P(t)$ we have that:

$$\frac{P'(t)}{P(t)} = \mu - d_N \frac{N(t)}{P(t)} - d_C \frac{C_o(t)}{P(t)} - d_C \frac{C_r(t)}{P(t)} - d_C \frac{C_b(t)}{P(t)}. \tag{2.6}$$

If we define the rates (depending on time) as:

$$n = \frac{N}{P}, \quad c_o = \frac{C_o}{P}, \quad c_r = \frac{C_r}{P}, \quad c_b = \frac{C_b}{P}, \tag{2.7}$$

then, equation (2.6) can be transformed into:

$$\frac{P'}{P} = \mu - d_N n - d_C c_o - d_C c_r - d_C c_b \tag{2.8}$$

where

$$n + c_o + c_r + c_b = 1 \tag{2.9}$$

According to this, we can compute the derivative of n defined in eq. (2.7).

$$n' = \frac{N'P - NP^2}{P^2} = \frac{N'}{P} - n\frac{P'}{P} \tag{2.10}$$

Multiplying (2.1) by 1/P and substituting by the corresponding rates defined in (2.7), using (2.8) and (2.9), (2.10) can be transformed into:

$$n' = \mu - \beta n(c_o + c_r + c_b) + \varepsilon c_b - n\mu + n(d_C - d_N)(c_o + c_r + c_b) \tag{2.11}$$

In an analogous way, we also obtain:

$$\begin{aligned} c_o' &= \beta n(c_o + c_r + c_b) - \gamma c_o - \mu c_o + n c_o(d_N - d_C) & (2.12)\\ c_r' &= \gamma c_o - \sigma c_r - \mu c_r + n c_r(d_N - d_C) & (2.13)\\ c_b' &= \sigma c_r - \varepsilon c_b - \mu c_b + n c_b(d_N - d_C) & (2.14) \end{aligned}$$

These are the scaled equations. Now, model and data are in the same units, so we will be able to compare them directly. Note than $n + c_o + c_r + c_b = 1$.

We will focus in system (2.11)–(2.14) in the rest of the chapter.

2.2.2 Parameters estimation

We have estimated the parameters μ, ε, d_N and d_C using available data in literature. The other parameters of the model (β, γ and σ) have been estimated, by least squares, by fitting the model with data from Table 2.1.

Using sources of literature we have obtained the following estimations:

- $\mu = 0.01\ years^{-1}$. We consider the average Spanish birth rate between years 1995 and 2007 [35].

- $d_N = 0.008388\ years^{-1}$ is the average Spanish death rate between years 1995 and 2007 [35].

- $d_C = 0.01636\ years^{-1}$ is the death rate for cocaine consumers. This value has been estimated taking into account that, in Spain, approximately 6.8% of mortality is due to drugs consumption [53] and the value of d_N.

- $\varepsilon = 0.0000456\ years^{-1}$. To estimate the rate at which a habitual consumer goes into therapy and becomes a non-consumer, we define $\varepsilon{=}\varepsilon_1{\times}\varepsilon_2{\times}\varepsilon_3{\times}\varepsilon_4{\times}\varepsilon_5$, where:

  - ε_1 is related to the average percentage of the subpopulation of habitual consumers. Using data from Table 2.1 corresponding to the National Drug Observatory Reports, the average value of population with habitual consumption is 0.93%, i.e., $\varepsilon_1 = 0.0093$.
  - ε_2 is the percentage of habitual consumers in therapy. From official data [53], 4.25% of habitual consumers begin a therapy programme every year in the Comunidad Valenciana. We assume the same rate for the whole of Spain. Then, $\varepsilon_2 = 0.0425$.
  - ε_3 is the time a habitual consumer takes before going into therapy. Moreover, using the average value presented in literature [8,10,20,52], a habitual consumer takes about 9 years before going to therapy. That means than $\varepsilon_3 = 1/9$.

    Therefore, the percentage of habitual consumers in therapy per year is 0.00439%. To be precise, 0.0093×0.0425×1/9=0.0000439.

    Additionally, around 52% of the individuals on therapy recover with an average of 6 months [8, 20, 38, 42, 52, 73, 78]. Therefore:
  - $\varepsilon_4 = 0.52$ is the average percentage of success for therapy programmes.
  - $\varepsilon_5 = 1/0.5$ indicates that the success in therapy programmes is reached after half a year.

  Then, we obtained:

  $\varepsilon = \varepsilon_1{\times}\varepsilon_2{\times}\varepsilon_3{\times}\varepsilon_4{\times}\varepsilon_5 = 0.0093{\times}0.0425{\times}1/9{\times}0.52{\times}1/0.5 = 0.0000456.$

On the other hand, taking as the initial conditions of the model (year 1995, i.e., t=0), $N(t = 0) = 0.944$, $C_o(t = 0) = 0.034$, $C_r(t = 0) = 0.018$ and $C_b(t = 0) = 0.004$, the parameters β, γ and σ have been estimated by fitting the scaled model with data from Table 2.1.

In order to compute the best fitting, we carried out computations with *Mathematica* [51]. To estimate the parameters mentioned (β, γ and σ), we implemented a function

$$\begin{aligned} \mathbb{F}\colon \mathbb{R}^3 &\to \mathbb{R} \\ (\beta,\, \gamma,\, \sigma) &\to \mathbb{F}(\beta,\, \gamma,\, \sigma) \end{aligned}$$

whose variables are β, γ and σ and such that:

1. Solve numerically (NDSolve[ ]) the scaled system of differential equations with initial values $N(t = 0) = 0.944$, $C_o(t = 0) = 0.034$, $C_r(t = 0) = 0.018$ and $C_b(t = 0) = 0.004$ and the parameter values defined above ($\mu = 0.01$, $d_N = 0.008388$, $d_C = 0.01636$, $\varepsilon = 0.0000456$).
2. For t = year 1995, year 1997, year 1999, year 2001, year 2003 and year 2005, corresponding to the biannual drugs use surveys, evaluate the computed numerical solution for each subpopulation $N(t)$, $C_o(t)$, $C_r(t)$ and $C_b(t)$.
3. Compute the mean square error between the values obtained in Step 2 and the real data presented in Table 2.1. This is the definition of function $\mathbb{F}$.

Function $\mathbb{F}$ takes values in $\mathbb{R}^3$ (β, γ and σ) and returns real values. Hence, we minimize this function using the Nelder-Mead algorithm that does not need the computation of any derivate or gradient, impossible to know in this case [60, 66].

In order to find a global minimum the feasible chosen domain is

$$\mathrm{D} = [0,\, 1]\ \times[0, 1]\ \times[0, 1] \subset \mathbb{R}^3,$$

and it is divided in disjoint subdomains where, in each one, Nelder-Mead algorithm is applied. We stored all the minima obtained and, among them, the values of β, γ and σ that minimize the function $\mathbb{F}$ are $\beta = 0.09614$, $\gamma = 0.0596$ and $\sigma = 0.0579$.

Table 2.2 summarizes all the estimated parameters of the model.

As we commented previously, the modeling and simulation itself introduces uncertainty related to the form of the model as well as the numerical approximations employed. Each of these sources of uncertainty must be estimated and included in order to estimate the total uncertainty in the prediction with the aim to report on how to reduce it in the most efficient and effective manner. In the next section we will explain the technique used to incorporate uncertainty in the model.

Table 2.2: Parameters of the mathematical model.

Parameter	Value ($years^{-1}$)
μ	0.01
d_N	0.008388
d_C	0.01636
β	0.09614
γ	0.0596
σ	0.0579
ε	0.0000456

2.2.3 LHS method

Latin Hypercube Sampling (LHS) method is a type of stratified Monte Carlo sampling method [4, 45, 64]. LHS allows an un-biased estimate of the average model output, with the advantage that it requires fewer samples than simple random sampling to achieve the same accuracy. It is a sophisticated and efficient method for achieving equitable sampling of all input parameters simultaneously.

In LHS the estimation uncertainty for each input parameter is modeled by treating each input parameter as a random variable. For each parameter a probability density function is defined and stratified into N equal probability intervals, which are then sampled. N represents the sample size. It is usual to use the uniform distribution centered at deterministic parameter estimators in the absence of data to inform on the distribution for a given parameter. A single value is then selected randomly from every interval and this is done for every parameter. In this way, each interval for each parameter is sampled exactly once (without replacement). Thus, the entire range for each parameter is explored. Distributions of the outcome variables can then be derived directly by running the model N times with each of the sampled parameters set [32, 49].

In our model for predicting cocaine consumption in Spain, we assume that all the parameters follow a uniform probability distribution. The ranges of variation of the parameters have been as follows:

- Birth and death rates (μ, d_N and d_C), have remained fixed because in the years that the model is defined these parameters do not vary significantly (median 0.010109, 0.008797 and 0.016201, standard deviation 0.0007030, 0.0002339 and 0.0003029, respectively). Their values are shown in Table 2.2.

- The intervals for the parameters β, γ and σ, are chosen assuming that the value of the parameter presented in Table 2.2 may have a perturbation not greater than 100% (Unbiased maximum likelihood estimation).
- The parameter ε is moved according to the intervals detailed in Table 2.3:
  - The intervals for ε_2 (percentage of habitual consumers in therapy) is chosen assuming that the value of the parameter may have a perturbation not greater than 100% (unbiased maximum likelihood estimation).
  - The interval for ε_3 (years of cocaine consumption before therapy) is chosen assuming the years of cocaine use before therapy presented in Dutra study [20], 5-15 years.
  - The interval for ε_4 (success rate of therapy programmes) is chosen taking into account all of the programmes analysed in Dutra study [20]. This percentage takes into account all of the programmes analysed in Budney et al. [8], Dutra et al. [20], Mercer and Woody [52], Johnson et al. [38], Levin et al. [42], Schmitz et al. [73] and Stotts et al [78].
  - The interval for ε_5 (time of treatment) is chosen assuming a perturbation not greater than 100% to consider all of the treatments studied in Dutra study [20].

LHS was used to generate 5000 different values of the parameters. Then we used these samples to run 5000 evaluations of the model. The results of these evaluations allow us to determinate the 95% confidence intervals to the consumption predictions for each year.

2.3 Results

2.3.1 Predictions

The graphical representation of the model fitting and the predictions in the next few years can be seen in Figure 2.2. Points represent data from Table 2.1. The green line corresponds to the deterministic solution and the red lines correspond to the 95% confidence intervals obtained by LHS method.

Table 2.3: Parameters values

Parameter	Deterministic value	Interval
β	0.09614	[0,0.19]
γ	0.0596	[0,0.12]
σ	0.0579	[0,0.12]
ε_2	0.0425	[0,0.085]
ε_3	1/9	[0.06,0.2]
ε_4	0.52	[0.32,0.72]
ε_5	1/0.5	[0,4]

We noted a decreasing trend in non-consumer subpopulation, $N(t)$. Also, there is an increasing trend of cocaine consumption in Spain, that is, of occasional consumers subpopulation, $C_o(t)$, regular consumers subpopulation, $C_r(t)$, and habitual consumers subpopulation, $C_b(t)$. In Table 2.4, some of the numerical values depicted in Figure 2.2 are presented.

Table 2.4: Model predictions for years 2007, 2010, 2012 and 2015 of percentage of non-consumers (N), occasional consumers (C_o), regular consumers (C_r) and habitual consumers (C_b).

Year	N	C_o	C_r	C_b
2007	0.869	0.078	0.034	0.019
2010	0.842	0.093	0.040	0.025
2012	0.821	0.105	0.046	0.028
2015	0.785	0.125	0.055	0.035

If there are not changes in current cocaine consumption policies in the next few years, the model predicts that 78.5%, 12.5%, 5.5% and 3.5% of 15-64-year-old individuals in Spain will be, by year 2015, non-consumer, occasional consumer, regular consumer and habitual consumer, respectively.

Additionally, the LHS method allows us to obtain a prediction by intervals for each subpopulation over the next few years. The ranges of variation of the parameters have been shown previously in Table 2.3. The intervals for the different

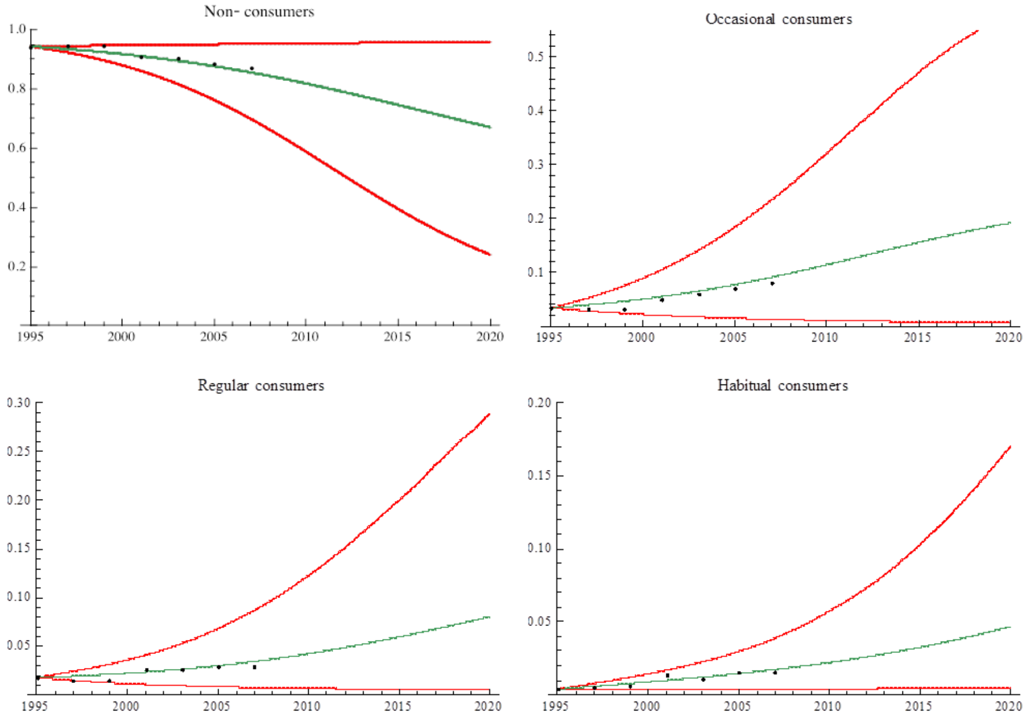


Figure 2.2: Numerical simulations of the fitted mathematical model. Points are data from Table 2.1. The green line corresponds to the deterministic solution and the red lines correspond to the 95% confidence intervals obtained by LHS method. Moreover, the predictions for the following years until 2020 are included.

subpopulations for 2011 to 2015 are shown in Table 2.5. We can observe that data and the deterministic solution falls within the interval of prediction for all the years and for all the subpopulations. However, as time goes on, the intervals are greater. With the aim to improve this result, we will apply other techniques in the following chapters in order to control the range of the prediction intervals.

2.3.2 Sensitivity analysis

In order to propose some strategies to control the epidemic, we carry out a sensitivity analysis. We performed several simulations varying the parameters of the model in order to find out what the influence of the changes on the final solution (cocaine consumption) is.

The objectives of the Spanish Government strategy on drug abuse [9, 61] are:

Table 2.5: 95% credible intervals (CI) for the period 2011–2015 of percentage of non-consumers (N), occasional consumers (C_o), regular consumers (C_r) and habitual consumers (C_b). Credible intervals are calculated considering the 2.5% and 97.5% percentile for each year.

	N	C_o	C_r	C_b
Year 2011				
Median	0.8335	0.0997	0.0359	0.0216
95% CI	[0.5485,0.9546]	[0.0102,0.3517]	[0.0068,0.1365]	[0.0043,0.0648]
Year 2012				
Median	0.8229	0.1060	0.0377	0.0228
95% CI	[0.5091,0.9552]	[0.0096,0.3818]	[0.0065,0.1510]	[0.0044,0.0729]
Year 2013				
Median	0.8117	0.1124	0.0399	0.0242
95% CI	[0.4699,0.9558]	[0.0091,0.4123]	[0.0063,0.1673]	[0.0044,0.0819]
Year 2014				
Median	0.8000	0.1188	0.0418	0.0257
95% CI	[0.4315,0.9563]	[0.0087,0.4426]	[0.0061,0.1835]	[0.0045,0.0920]
Year 2015				
Median	0.7879	0.1256	0.0439	0.0271
95% CI	[0.3944,0.9569]	[0.0082,0.4707]	[0.0060,0.2001]	[0.0045,0.1029]

1. The prevention of drug consumption, pointing out the health concerns produced by their consumption, delaying the age of the first contact with drugs, education programmes and the legal fight against drugs dealing.

2. To improve quantitative and qualitative research, to implement new treatments, evaluate current therapy programmes and training to increase the professional competence of the people who work with drug abusers.

These policies, related to prevention and treatment, involve focused efforts on controlling parameters β (contagion rate) and ε (rate at which a habitual consumer goes into therapy and becomes a non-consumer), respectively. The parameter ε is associated with the implementation of new treatments, evaluation of current therapy programmes, training plans to increase professional competence of the people who work with drug consumers, etc. That is, the parameter ε is associated with treatment policies, while the parameter β is associated with prevention policies. Thus, in order to analyse the strategies of Spanish Government against drug abuse, the six health policies simulated here, related with the ones mentioned above, are:

1. Variation on the percentage of habitual consumers in therapy. It involves a variation in parameter ε_2.

2. Variation on the time that a habitual consumer takes before going into therapy. It involves a variation in parameter ε_3.

3. Variation on the success rate of therapy programmes. It involves a variation in parameter ε_4.

4. Variation on the duration of therapy programmes. It involves a variation in parameter ε_5.

5. Variation on the transition rate from non-consumers to occasional consumers. It involves a variation of parameter β.

6. Variation of all these parameters (ε_2, ε_3, ε_4, ε_5 and β) together.

Note that the first four policies involves a variation in parameter ε, i.e., they are associated with treatment policies, while the fifth one involves a variation in parameter β, i.e., it is associated with prevention policies. In the last one all the parameters are modified. Therefore, it combines both policies.

To vary ε_2, ε_3, ε_4, ε_5 and β, we assume that all of them follow a uniform probability distribution with support on the intervals [0, 0.085], [0.06, 0.2], [0.32, 0.72], [0, 4] and [0, 0.19], respectively, according to the intervals shown in Table 2.3. It is usual to use the uniform distribution in the absence of data.

LHS technique described above was used to generate 5000 different values of the parameters ε_2, ε_3, ε_4, ε_5 and β (input). Then we used these samples to run 5000 evaluations of the model. The results of these evaluations allow us to determinate the 95% confidence intervals to the consumption predictions. The obtained predictions (regular consumption and habitual consumption) for year 2011 and year 2015 after the variation of the parameters can be observed in Table 2.6 and Table 2.7.

Note that the variation of the epsilons produces a variation in the 95% confidence interval smaller than 10^{-2}. Moreover the perturbation of parameter β leads to larger variations in output than the rest of the perturbed parameters. This fact allows us to say that health prevention policies (the ones related with parameter β) may have a noticeable effect on the reduction of drug consumption. Alternatively, if prevention policies are disregarded, cocaine consumption will increase.

Table 2.6: Sensitivity analysis: regular consumption

Strategy 1	*Variation of the regular consumers in therapy (ε_2)*	
Regular consumers(%)	*year 2011*	*year 2015*
95% confidence interval	[4.34,4.34]	[5.54,5.54]
Mean 5000 realizations	4.34	5.54
Model estimation	4.3	5.5
Strategy 2	*Variation of the time before going into therapy (ε_3)*	
Regular consumers(%)	*year 2011*	*year 2015*
95% confidence interval	[4.34,4.34]	[5.54,5.54]
Mean 5000 realizations	4.34	5.54
Model estimation	4.3	5.5
Strategy 3	*Variation of the rate of therapy success (ε_4)*	
Regular consumers(%)	*year 2011*	*year 2015*
95% confidence interval	[4.34,4.34]	[5.54,5.54]
Mean 5000 realizations	4.34	5.54
Model estimation	4.3	5.5
Strategy 4	*Variation of the duration of therapy programmes (ε_5)*	
Regular consumers(%)	*year 2011*	*year 2015*
95% confidence interval	[4.34,4.34]	[5.54,5.54]
Mean 5000 realizations	4.34	5.54
Model estimation	4.3	5.5
Strategy 5	*Variation of the transition rate to occasional consumers (β)*	
Regular consumers(%)	*year 2011*	*year 2015*
95% confidence interval	[1.66,9.80]	[1.49,14.36]
Mean 5000 realizations	4.89	6.51
Model estimation	4.3	5.5
Strategy 6	*Variation of ε_2, ε_3, ε_4, ε_5 and β (all together)*	
Regular consumers(%)	*year 2011*	*year 2015*
95% confidence interval	[1.66,9.80]	[1.49,14.36]
Mean 5000 realizations	4.89	6.51
Model estimation	4.3	5.5

Table 2.7: Sensitivity analysis: habitual consumption

Strategy 1	*Variation of the habitual consumers in therapy (ε_2)*	
Habitual consumers(%)	*year 2011*	*year 2015*
95% confidence interval	[2.62,2.62]	[3.56,3.56]
Mean 5000 realizations	2.62	3.56
Model estimation	2.6	3.5
Strategy 2	*Variation of the time before going into therapy (ε_3)*	
Habitual consumers(%)	*year 2011*	*year 2015*
95% confidence interval	[2.62,2.62]	[3.56,3.56]
Mean 5000 realizations	2.62	3.56
Model estimation	2.6	3.5
Strategy 3	*Variation of the rate of therapy success (ε_4)*	
Habitual consumers(%)	*year 2011*	*year 2015*
95% confidence interval	[2.62,2.62]	[3.56,3.56]
Mean 5000 realizations	2.62	3.56
Model estimation	2.6	3.5
Strategy 4	*Variation of the duration of therapy programmes (ε_5)*	
Habitual consumers(%)	*year 2011*	*year 2015*
95% confidence interval	[2.62,2.62]	[3.56,3.56]
Mean 5000 realizations	2.62	3.56
Model estimation	2.6	3.5
Strategy 5	*Variation of the transition rate to occasional consumers (β)*	
Habitual consumers(%)	*year 2011*	*year 2015*
95% confidence interval	[1.80,4.09]	[2.03,6.58]
Mean 5000 realizations	2.75	3.85
Model estimation	2.6	3.5
Strategy 6	*Variation of ε_2, ε_3, ε_4, ε_5 and β (all together)*	
Habitual consumers(%)	*year 2011*	*year 2015*
95% confidence interval	[1.80,4.09]	[2.03,6.58]
Mean 500 realizations	2.75	3.85
Model estimation	2.6	3.5

2.4 Conclusion

In this chapter, we propose a type-epidemiological mathematical model applied to cocaine consumption in Spain. If there are not changes in current cocaine consumption policies over the next few years, the model predicts a decreasing trend in non-consumer subpopulation, while there is an increasing trend in all the consumer subpopulations, that is, occasional, regular and habitual consumer subpopulations. Specifically, the model predicts that 78.5%, 12.5%, 5.5% and 3.5% of 15-64-year-old individuals in Spain will be, by the year 2015, non-consumer, occasional consumer, regular consumer and habitual consumer, respectively. In addition, we have introduced uncertainty in the model, considering uncertainty in the parameters estimation. Then, the 95% credible intervals for cocaine consumption prediction for the period 2011–2015 have been estimated, for the different subpopulations. Using the LHS method, the model predicts the following intervals for cocaine consumption in 2015: $[39.44\%, 95.69\%]$, $[0.82\%, 47.07\%]$, $[0.60\%, 20.01\%]$ and $[0.45\%, 10.29\%]$ for non-consumer, occasional consumer, regular consumer and habitual consumer, respectively.

Furthermore, we can associate the parameters of the model with policies of the Spanish Health Ministry. Parameter β is associated with prevention policies and parameter ε with treatment policies. After the simulation of different hypothetical scenarios where different health policies are performed, we can conclude that prevention policies seems to be the best effective strategy to reduce the population of regular and habitual consumers. Similar conclusions have been obtained in other works [43, 68, 77, 82]. We note that taking into account random perturbations on β, the 95% confidence interval prediction presents the most important variability, i.e., modifications in prevention programmes (variations of β) are the best option to modify the levels of consumption (confidence interval).

In the other cases, the ones related to parameters ε_2, ε_3, ε_4 and ε_5, the variations on the parameters do not produce noticeable variations on the confidence intervals (cocaine consumption prediction). This happens because the percentage of habitual consumers that begin therapy every year is around 4.25% as we mentioned in the section “Sensitivity analysis”, and it is a very small amount of the total population.

Obviously this is not the only epidemiological model designed to study the spread of cocaine use, and it is not the only one capable of simulating scenarios, with a view to informing and assisting policy-makers in targeting prevention and treatment resources for maximum effectiveness. However, to the best of our knowledge, the model presented is the first one applied to Spain with real data.

In the following chapter we will present other mathematical models to study the dynamics of cocaine consumption in Spain and we will check that the model defined in this chapter is the one that best explains the Spanish situation. Moreover, with the aim to improve and to control the uncertainty in estimation, in the following chapters we will use other different methods for confidence interval estimation, the ABC and the Bootstrap methods. These methods will be applied to the model defined in this chapter. Our objective will be to reduce the range of the prediction intervals, whenever the real data fall within them. Estimate the parameters assuming uncertainty allows us to predict considering the parameters uncertainty.

Chapter 3

Approximate Bayesian Computation (ABC) method for model selection and confidence interval parameters estimation

In this chapter, three possible scenarios (three mathematical models) are presented to study the evolution of cocaine consumption in Spain and using an Approximate Bayesian Computation (ABC) technique we will select the model that best matches the Spanish situation. We will check that the model defined in the previous chapter is the one that best explains this situation. Then, the ABC algorithm is applied to this cocaine consumption model, for which parameters and credible intervals are inferred. Moreover, ABC provides information about model sensitivity to parameter changes.

3.1 Introduction

Population dynamic models present unknown parameter values or impossible to measure directly. For that reason, assessing the uncertainty about their estimates and model predictions is a key point. Population dynamical models, commonly used in the study of epidemics and other complex population processes, have traditionally been considered deterministic, i.e., with constant coefficients in these equations. However, in many situations, equations with random coefficients are better suited in describing the real behaviour of the quantities of interest than their counterparts with deterministic coefficients. It is the case of the social epidemics. Predicting in social epidemics is an exercise that involves uncertainties. A number of interesting issues, such as sampling, rounding errors and lack of information in the parameter estimation processes, need to be addressed.

In this chapter we will present the use of Approximate Bayesian Computation (ABC) approach to study the evolution of cocaine consumption epidemic in Spain by a mathematical model based on ordinary differential equations with randomness in the parameters. The ABC method allows us to consider that the parameters are random variables and to obtain the evolution of the solution of the mathematical model considering the effects of the randomness on the predictions. A first deterministic version of the cocaine consumption mathematical model considered was presented in [69] and in the previous chapter. Additionally, ABC methods, also known as likelihood-free techniques, have been conceived to infer posterior distribution in case where likelihood functions are computationally intractable or too costly. In these methods, the evaluation of the likelihood is replaced by a simulation-based procedure. Although there is a wide variety of tools available for parameter estimation and, to a lesser extent, model selection, the ABC SMC (Sequential Monte Carlo) method yields reliable parameter estimates with credible intervals, can be applied to different types of models (e.g. deterministic or stochastic models), is computationally efficient, allows discrimination among a set of candidate models and gives us an assessment of parameter sensitivity.

3.2 Methods

3.2.1 Mathematical models

In this section, we propose three feasible models, based on differential equations, to study the transmission dynamics of cocaine consumption in Spain and using Approximate Bayesian Computation (ABC) technique we will select the model that best describes this situation. We will check that the model that best explains it, that is, the model that fits the data better, is the one described in the previous chapter. Then, we will work with it for the rest of the dissertation. There are obviously other alternatives to model this dynamical process (delay models, stochastic models, etc.), but in this work we restrict ourselves to the three models described.

3.2.1.1 Model 1

In order to build the mathematical model, the 15-64-year-old Spanish population is considered and divided into four subpopulations, following the division proposed in the previous chapter:

- $N(t)$: Non-consumers, individuals who have never consumed cocaine.
- $C_o(t)$: Occasional consumers, individuals who have consumed sometimes in their life.
- $C_r(t)$: Regular consumers, individuals who have consumed in the last year.
- $C_b(t)$: Habitual consumers, individuals who have consumed in the last month.

The evolution of the different subpopulations for the last years has been shown in Table 2.1. However, in this case, we update the database incorporating more recent available data (year 2009). In the following table (Table 3.1) we can see the evolution of the different subpopulations for the last years.

Table 3.1: Evolution of the proportion of non-consumers (N), occasional consumers (C_o), regular consumers (C_r) and habitual consumer subpopulations (C_b) for different years. The data have been obtained from the Drug National Observatory Reports (Spanish Ministry of Health) [61, 62].

Year	N	C_o	C_r	C_b
1995	0.944	0.034	0.018	0.004
1997	0.948	0.032	0.015	0.005
1999	0.948	0.031	0.015	0.006
2001	0.911	0.049	0.026	0.014
2003	0.903	0.059	0.027	0.011
2005	0.884	0.070	0.030	0.016
2007	0.874	0.080	0.030	0.016
2009	0.860	0.102	0.026	0.012

Model 1 (m_1) is the model defined in Chapter 2, with some improvement. In this case, the same assumptions are considered and the same definitions for the parameters of the model are used. However, as well as in the previous chapter we considered different mortality rates for cocaine consumers from non-consumers, and we admitted this mortality rate was the same to all the cocaine consumers, here we go beyond. In this new version of the model shown in the last chapter, we do not only consider different mortality rates between consumers and non-consumers. Furthermore, we suppose that this mortality rate is different for the different types of consumers. Also, we admit that this mortality is higher when the consumption increases. That is, if we denote:

- d_{c_1}, death rate for occasional cocaine consumers,
- d_{c_2}, death rate for regular cocaine consumers,
- d_{c_3}, death rate for habitual cocaine consumers,

then,

$$d_{c_1} \leq d_{c_2} \leq d_{c_3}.$$

The transitions between the subpopulations N, C_o, C_r and C_b according to the model 1 (m1) are shown in Figure 3.1 and are described by the equations (3.1)–(3.5).

$$\begin{aligned}
\frac{dN(t)}{dt} &= \mu P(t) - d_N N(t) - \beta \frac{N(t)(C_o(t)+C_r(t)+C_b(t))}{P(t)} + \\
&\quad + \varepsilon C_b(t), & (3.1)\\
\frac{dC_o(t)}{dt} &= \beta \frac{N(t)(C_o(t)+C_r(t)+C_b(t))}{P(t)} - d_{c_1} C_o(t) - \gamma C_o(t) & (3.2)\\
\frac{dC_r(t)}{dt} &= \gamma C_o(t) - d_{c_2} C_r(t) - \sigma C_r(t) & (3.3)\\
\frac{dC_b(t)}{dt} &= \sigma C_r(t) - d_{c_3} C_b(t) - \varepsilon C_b(t) & (3.4)\\
P(t) &= N(t) + C_o(t) + C_r(t) + C_b(t) & (3.5)
\end{aligned}$$

In this first model, we admit that the only possibility to decrease cocaine consumption is by therapy if the consumers are in C_b. That is, we consider that an individual of subpopulation C_b becomes a member of subpopulation of non-consumers, N, if he/she decides to give up the cocaine consumption and go into therapy.

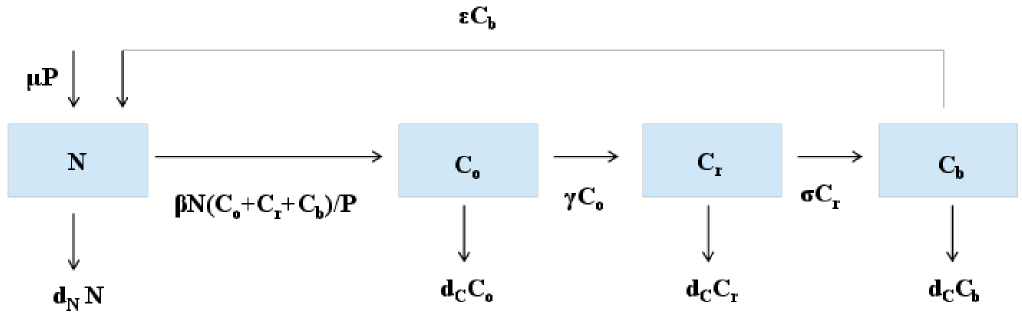


Figure 3.1: Flow diagram of the mathematical model 1 for the dynamics of cocaine consumption in Spain. The boxes represent the subpopulations and the arrows represent the transitions between the subpopulations. Arrows are labeled by their corresponding model transition terms.

3.2.1.2 Model 2

Model 2 (m_2) is defined taking into account the same subpopulations as model 1 but in this case we consider a new transition, C_r to C_o, modeled by αC_r. Then, this model has an additional parameter, α, which is the rate at which a regular consumer becomes an occasional consumer by decreasing his/her frequency of consumption. The introduction of this new transition is in order to test the hypothesis that a non-problematic consumption can be controlled.

In this second model, we consider the possibility that a regular consumer can decrease his/her cocaine consumption (without therapy) and he/she can become an occasional consumer. In the first one, we admit that the only possibility to decrease cocaine consumption is by therapy if the consumers present a monthly frequency of consumption (habitual consumers).

The transitions between the subpopulations N, C_o, C_r and C_b according to the model 2 (m_2) are shown in Figure 3.2 and are described by the equations (3.6)–(3.10) (in bold, the new term introduced).

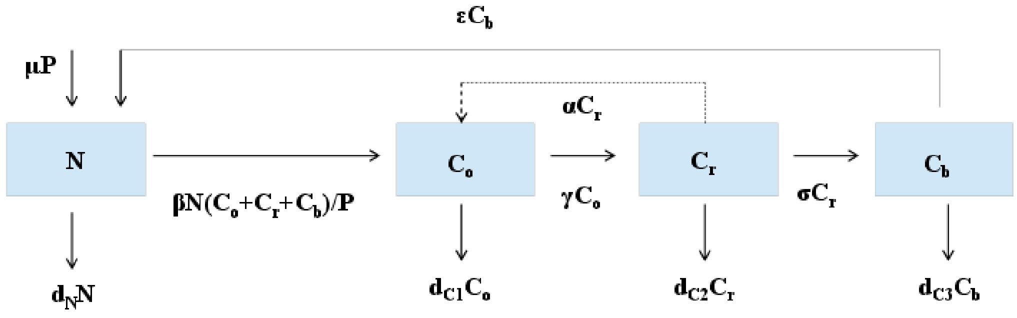


Figure 3.2: Flow diagram of the mathematical model 2 for the dynamics of cocaine consumption in Spain. The dashed arrow is the additional one corresponding to model 2.

$$
\begin{aligned}
\frac{dN(t)}{dt} &= \mu P(t) - d_N N(t) - \beta \frac{N(t)(C_o(t)+C_r(t)+C_b(t))}{P(t)} + \\
&\quad + \varepsilon C_b(t), && (3.6)\\
\frac{dC_o(t)}{dt} &= \beta \frac{N(t)(C_o(t)+C_r(t)+C_b(t))}{P(t)} - d_{c_1} C_o(t) - \gamma C_o(t) + \boldsymbol{\alpha} \mathbf{C_r(t)} && (3.7)\\
\frac{dC_r(t)}{dt} &= \gamma C_o(t) - d_{c_2} C_r(t) - \sigma C_r(t) - \boldsymbol{\alpha} \mathbf{C_r(t)} && (3.8)\\
\frac{dC_b(t)}{dt} &= \sigma C_r(t) - d_{c_3} C_b(t) - \varepsilon C_b(t) && (3.9)\\
P(t) &= N(t) + C_o(t) + C_r(t) + C_b(t) && (3.10)
\end{aligned}
$$

3.2.1.3 Model 3

To define model 3 (m_3), we consider a new subpopulation, T, made up of habitual cocaine consumers who decide to give up consumption and go into therapy. The new data are shown in Table 3.2.

Model 3 has two additional parameters and a different definition for ε as follows:

- ε, rate at which habitual consumers enter into therapy.
- ϕ, rate at which people in therapy, leave therapy and return to habitual consumption.
- d_{c_4}, death rate of people in therapy.

Table 3.2: Evolution of the proportions of the subpopulations defined for model 3 for different years.

	N	C_o	C_r	C_b	T
1997	0.947953	0.032	0.015	0.005	0.0000466
1999	0.947887	0.031	0.015	0.005	0.0001127
2001	0.910788	0.049	0.026	0.014	0.0002116
2003	0.902585	0.059	0.027	0.011	0.0004154
2005	0.883483	0.070	0.030	0.016	0.0005174
2007	0.873509	0.080	0.030	0.016	0.0004907
2009	0.859545	0.102	0.026	0.012	0.0004546

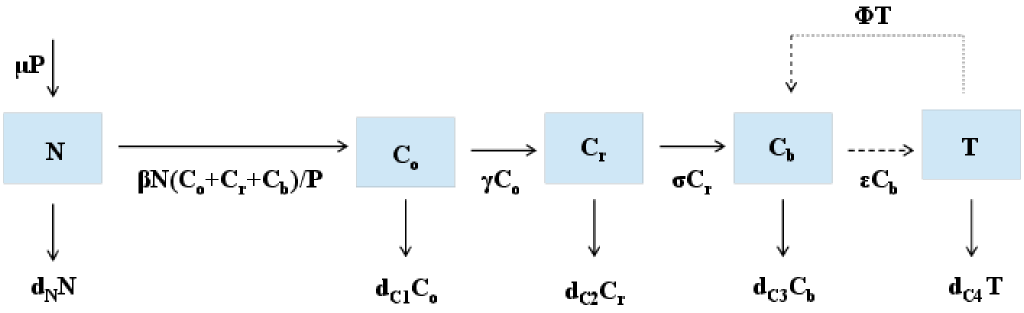


Figure 3.3: Flow diagram of the mathematical model 3 for the dynamics of cocaine consumption in Spain. The dashed arrows are the additional ones corresponding to model 3.

The transitions between the subpopulations N, C_o, C_r and C_b according to the model 3 (m_3) are shown in Figure 3.3 and are described by the equations (3.11)–(3.16) (in bold, the new terms introduced).

$$\begin{aligned}
\frac{dN(t)}{dt} &= \mu P(t) - d_N N(t) - \beta \frac{N(t)(C_o(t)+C_r(t)+C_b(t))}{P(t)} && (3.11)\\
\frac{dC_o(t)}{dt} &= \beta \frac{N(t)(C_o(t)+C_r(t)+C_b(t))}{P(t)} - d_{c_1} C_o(t) - \gamma C_o(t) && (3.12)\\
\frac{dC_r(t)}{dt} &= \gamma C_o(t) - d_{c_2} C_r(t) - \sigma C_r(t) && (3.13)\\
\frac{dC_b(t)}{dt} &= \sigma C_r(t) - d_{c_3} C_b(t) \boldsymbol{- \varepsilon \mathbf{C_b(t)} + \phi \mathbf{T(t)}} && (3.14)\\
\frac{dT(t)}{dt} &= \boldsymbol{\varepsilon \mathbf{C_b(t)} - \phi \mathbf{T(t)} - \mathbf{d_{c_4} T(t)}} && (3.15)\\
P(t) &= N(t) + C_o(t) + C_r(t) + C_b(t) + T(t) && (3.16)
\end{aligned}$$

Data in Table 3.1 and Table 3.2 are in percentages and models are defined considering individuals. Therefore, we also have to scale the three models presented in this chapter. In order to do it, techniques shown in [33, 36, 50] are used and the process is similar to the one described in the previous chapter. In this chapter, we are not going to show the process and the scaled models because it is a technical transformation and the resulting does not provide extra information about the models. Additionally, the scaled models have the same parameters as the non-scaled models with the same meaning and in order to avoid introducing new notation, we are going to consider that the subpopulations $N(t)$, $C_o(t)$, $C_r(t)$ and $C_b(t)$ correspond to the percentage of non-consumers, occasional consumer, regular consumers and habitual consumers. Note that the models used in ABC algorithm are the scaled models.

We have already presented three models and then we have to decide which one describes better the evolution of cocaine consumption in Spain. To do this, we are going to use the ABC SMC technique described by T. Toni et al. in [79]. Before the posterior distribution estimation is processed, prior probability distribution for the parameters needs to be defined as well as the initial condition of the model.

3.2.2 Parameters estimation

In order to specify a joint prior distribution for the parameters of the mathematical model,

$$\theta = (\mu, d_N, d_{c_1}, d_{c_2}, d_{c_3}, \beta, \gamma, \sigma, \varepsilon),$$

separate sources of information are used to specify each of these prior distributions. Therefore, we specify the model parameters to be independent a priori and having a uniform distribution. It is usual to use the uniform distribution centered at deterministic parameter estimators in the absence of data to inform on the distribution for a given parameter.

– Taking into account the information provided by the Spanish Statistical Office [35] for the period 1995–2009, minimum and maximum values for μ and d_N are defined (unbiased maximum likelihood estimation for the uniform distribution parameters).

– Let us consider the cases of γ and σ parameters. The outflow for C_o is γC_o therefore the condition $0 \leqslant \gamma C_o \leqslant C_o$ must be satisfied. Then dividing by C_o, we have $0 \leqslant \gamma \leqslant 1$. Analogously, σ must satisfy $0 \leqslant \sigma \leqslant 1$.

– For parameter β we consider that the outflow $\beta N(C_o + C_r + C_b)/P$ must be less than subpopulation N for each t. This leads to the following condition for β:

$$\beta \leqslant \frac{P}{C_o + C_r + C_b} = \frac{1}{c_o + c_r + c_b} = \frac{1}{n-1}, \tag{3.17}$$

where $n = N/P$.

Then, taking into account that $0 \leqslant n < 1$, the general condition for β is $0 \leqslant \beta \leqslant \infty$. However, since in our case n is always less that 0.95 (note that we are estimating the model for the period 1995–2009; see Table 3.2), then by (3.17), our range of variation for β will be $0 \leqslant \beta \leqslant 20$.

– As we have shown in the previous chapter and in [69], we know that $\varepsilon = 0.0000456$. The interval for ε is chosen assuming unbiased maximum likelihood estimation. For this parameter, the maximum likelihood estimation of Uniform(0, ε) is the maximum of the sample considered to estimate the parameter, i.e., the only value of the sample, the known value of the parameter. In this case, the expected value of parameter ε is a half of its known value. Therefore, if we consider the distribution defined by Uniform(0, $2\times\varepsilon$), we have that its expected value is the known value of the parameter.

– The intervals for d_{c_1}, d_{c_2} and d_{c_3} are chosen taking into account that mortality was four to eight times higher among cocaine users than age and sex peers in the general population [17]. Therefore, we can admit that this parameters have a range of variation defined by $[4 \times min\{d_N\}, 8 \times max\{d_N\}]$. Additionally, we consider that $d_{c_1} \leqslant d_{c_2} \leqslant d_{c_3}$.

These prior distributions are summarized in Table 3.3.

Table 3.3: Prior distribution for the parameters of the models. These values are for the period 1995–2009. We assume that all the parameters follow a uniform probability distribution. It is usual to use the uniform distribution in the absence of data to define a more informative prior distribution. Minimum and maximum values shown in the table are the ranges of variation of the parameters.

	Min	Max
μ	0.008343	0.009228
d_N	0.009227	0.010944
dc_1	0.0369	0.0875
dc_2	0.0369	0.0875
dc_3	0.0369	0.0875
β	0.0	20
γ	0.0	1
σ	0.0	1
ε	0.0	0.00009
α	0.0	1

We assume that the initial condition of the models, that is, the prevalence of cocaine consumption in 1995 (see Table 3.1) is known without uncertainty and it is not considered as a random variable. Note that the initial condition is determined using a representative sample of the Spanish population.

3.2.3 Approximate Bayesian Computation method

Taking into account Bayesian paradigm, Approximate Bayesian Computation (ABC) methods can be used to evaluate posterior distributions without having to calculate likelihoods, using a simulation-based procedure.

Let θ be a parameter vector of a differential equation system to be estimated and $\pi(\theta)$ its prior probability distribution. The objective is to estimate the parameters of the system. To do this, we try to obtain their posterior distribution $\pi(\theta|x)$, where

x are the observed data. Thus, we will obtain an approximation of it taking into account the Bayes' theorem where we know that $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$ and $f(x|\theta)$ is the likelihood of θ given the data x. Note that, unfortunately, the likelihood term $f(x|\theta)$ is usually unknown and it is expensive or impossible to calculate. In this context, ABC is an interesting alternative to estimate $\pi(\theta|x)$, that only requires being able to sample pseudo-observations from $f(\cdot|\theta)$.

In this work, we apply the ABC method based on Sequential Monte Carlo (ABC SMC) for model selection presented by T. Toni et al. in [79]. This Bayesian approach is based on the study of the evidence provided by the data (x) in favor of one model over the other. In our case, observed data, x, are shown in Table 3.1. Since we have presented three models, we are going to compare them in pairs. To be precise, the objective is to obtain a set of N parameter vectors of the model m, $\theta(m)$, divided between the two models, that satisfies the final condition that the prediction x^* given by model m with values of the parameters $\theta(m)$ has a distance less than a desired tolerance level ϵ_T from the observed data, that is, $d(x^*, x) \leq \epsilon_T$. At the end of the process, we select the model having the highest number of $\theta(m)$ that satisfy this condition. To obtain a final estimation of the parameters, we will have intermediate estimations, that is, populations of N parameter vectors $\theta(m)$, by refining the values of the maximum distance permitted in each iteration. In other words, we sample considering that the tolerance level decreases until ϵ_T.

This algorithm proceeds as follow:

Step 1. Initialize $\epsilon_1, \epsilon_2, \cdots, \epsilon_T$, where $\epsilon_1 > \epsilon_2 > \cdots > \epsilon_T$
Set the population indicator $t = 0$.

Step 2.

Step 2.0. Set the particle indicator $i = 1$ (i varies from 1 to N).

Step 2.1. Approach the model selection problem by including a 'model parameter' $m \in \{1, 2, 3\}$, where 3 is the number of models, as an additional discrete parameter and denote the model-specific parameters as $\theta(m)$. In our case, $m = 1$ for model 1, $m = 2$ for model 2 and $m = 3$ for model 3 and $\pi(m)$ is the same for the three models ($\pi(m)=\frac{1}{3}$). Sample a model indicator m from the prior distribution for each model $\pi(m)$. Denote this model indicator as m^*, where $m^* = 1, 2, 3$.

If $t = 0$, sample the parameter vectors, θ^{**}, from the previous distribution, $\pi(\theta(m^*))$. Thus we obtain a set of values for the parameters of model m from a uniform distribution for each of them.

Note that the parameter vectors from the previous distribution are denoted by a single asterisk, and after perturbation the vectors are denoted by a double asterisk.

If $t > 0$, sample a parameter vector θ^* from the previous population of the parameters $\{\theta(m^*)_{t-1}\}$ with weights $w(m^*)_{t-1}$ and perturb θ^* to obtain a new set of values $\theta^{**} \sim K_t(\theta|\theta^*)$, where the perturbation kernel K_t is a non-parametric way to estimate the probability density function of a random variable. A kernel is a non-negative real-valued integrable function K satisfying the following two requirements:$\int_{-\infty}^{+\infty} K(u)\,du = 1$ and $K(-u) = K(u)$ for all values of u. In our case, we have considered the uniform distribution to define it.

If $\pi(\theta^{**}) = 0$ return to Step 2.1.

Simulate a candidate dataset $x^* \sim M(x|\theta^{**}, m^*)$, where $M(x|\theta^{**}, m^*)$ is the dynamic model 1 (m_1) if $m^* = 1$, the dynamic model 2 (m_2) if $m^* = 2$ and the dynamic model 3 (m_3) if $m^* = 3$.

Verify that the prediction given by this set of values satisfies the condition for the distance $d(x^*, x) \leq \epsilon_t$, where x are the observed data:

If $d(x^*, x) \leq \epsilon_t$, go to Step 2.2.

If $d(x^*, x) \geq \epsilon_t$, return to Step 2.1.

Step 2.2. Assign weights for this set of parameters.

Set $m_t^{(i)} = m^*$ and add θ^{**} to the population of particles $\{\theta(m^*)_t\}$, and calculate its weight as:

$$w_t^{(i)} = \begin{cases} 1, & \text{if } t = 0 \\ \frac{\pi(\theta^{**})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\theta_{t-1}^{(j)}, \theta^{**})}, & \text{if } t > 0 \end{cases}$$

If $i < N$, set $i = i + 1$, go to Step 2.1.

Step 3. Normalize the weights for the N obtained vectors $\theta(m)$, with a set of parameters for each one, for every m.

If $t < T$, set $t = t + 1$ and go to Step 2.0.

Note that the procedure is:

(a) Select model m^* from the prior distribution $\pi(m)$.

(b) Select the parameter vector θ^{**}.

(c) Simulate output x^* from the model selected using the parameter vector θ^{**}.

(d) Compute the distance $d(x^*, x)$. If $d(x^*, x) \leqslant \epsilon$, accept and store θ^{**} for model m^* otherwise reject it.

(e) Return to step (a).

The output of the algorithm is a sample of parameters from the distribution $\pi(\theta | d(x^*, x) \leqslant \epsilon)$. If ϵ is small enough, the distribution $\pi(\theta | d(x^*, x) \leqslant \epsilon)$ will be an optimal approximation of the posterior distribution $\pi(\theta | x)$.

The parameter estimation for each model is calculated simultaneously with the model selection. The model with the highest posterior probability will have the greater number of particles. This ensures a good estimation of its parameters.

In the following section, we are going to use the ABC SMC approach to select the model that best describes the evolution of cocaine consumption in Spain. As we commented above, we are going to compare the three defined models in pairs. In addition, the ABC algorithm provides us an approximation to the posterior probability distribution of the parameters of the selected model. It will be used to predict the evolution of cocaine consumption in Spain over the next few years, by credible intervals.

3.3 Results

3.3.1 Model selection

In a first step, we are going to compare model 1 (m_1) with model 2 (m_2) and then, we will compare the best model obtained (between this two first models) with the third model.

3.3.1.1 Selection between model 1 and model 2

The values of ϵ_t that we have used to compare model 1 (m_1) with model 2 (m_2) are $\epsilon_1 = 0.0160$, $\epsilon_2 = 0.0090$, $\epsilon_3 = 0.0070$ and $\epsilon_4 = 0.0066$. These values are defined considering the deterministic fitting of the model 1 in the least square sense. The definition intervals shown in Table 3.3 are used in this deterministic fitting of the parameters of the model 1. The distance function $d(\cdot;\cdot)$ is defined by the root mean square error. Note that the distance between the observed prevalence data (Table 3.1) and the deterministic solution for model 1 is 0.006138. Then, the lowest distance to be reached is expected to be close to this number. Therefore, we choose the tolerance level ϵ_i accordingly. We take $T = 4$ and we have considered $N = 1000$ for the number of particles.

Figure 3.4 shows the distributions obtained for the parameter $\theta(m)$ for each iteration $t = 1, 2, 3, 4$ according to the four values of ϵ_t. We can see how the number of times that the algorithm selects the model 2 is decreasing as ϵ_t is decreasing. Finally, only model 1 is selected. Thus, we can conclude that model 1 is the one that best describes the evolution of the subpopulations.

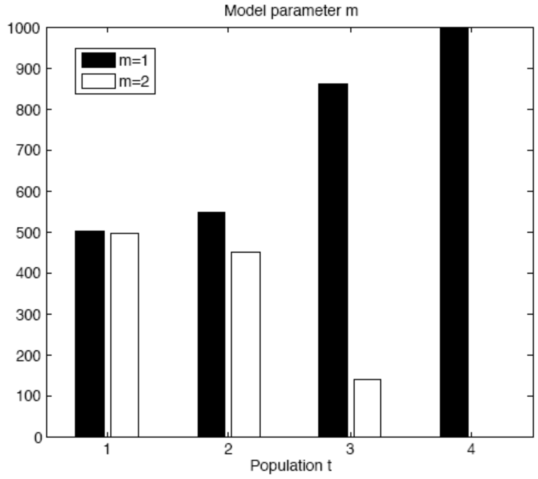


Figure 3.4: Evolution of the number of parameters vectors, $\theta(m)$, corresponding to model 1 and model 2 in each population $t = 1, 2, 3, 4$.

3.3.1.2 Selection between model 1 and model 3

Now we are going to compare model 1 (m_1) with model 3 (m_3) in an analogous way. In this case, the values of ϵ_t that we have used to ensure the transition from the prior distributions for the parameters to the posterior distributions are: $\epsilon_1 = 1.0$,

$\epsilon_2 = 0.5$, $\epsilon_3 = 0.3$ and $\epsilon_4 = 0.13$. Note that the selection of the values of ϵ_t is arbitrary and it does not matter whenever they satisfy the condition of decreasing order. In addition, we take $T = 4$ and $N = 1000$.

Since in data for model 3 the values for the subpopulation T are very small compared to N, C_o, C_r and C_b, we have to define the distance between our data and the deterministic solutions provided by the models by means of the relative error. Using the absolute error, the contribution to the error corresponding to the subpopulation T is negligible, and the algorithm does not take it into account resulting in a bad fit of the data. Then, since the selection of the distance does not influence the results (see [79]), we define it as follows:

$$d(x^*, x) = \sum_{i=1}^{n} \left| \frac{(x^*)_i - (x)_i}{(x)_i} \right|$$

where x^* are the values predicted by the model, x are the data values and n is the total number of data values for all the subpopulations together.

Figure 3.5 shows the distributions obtained for the parameter $\theta(m)$ for each iteration $t = 1, 2, 3, 4$ according to the four values of ϵ_t. We can see how the number of times that the algorithm selects the model 3 is decreasing as ϵ_t is decreasing. Finally, only model 1 survives. Thus, we can conclude that model 1 is the one that best describes the evolution of cocaine consumption in Spain. Therefore, we will work with this model in the rest of the dissertation. Note that this selected model is the one presented in the previous chapter, with slight improvements (more recent data and the definition of the mortality rates for cocaine consumers in a decreasing order).

In addition, the ABC algorithm provides us an approximation to the posterior distributions for the parameters of the selected model, in this case, model 1. This fact allows us to predict the evolution of cocaine consumption in Spain as we will see in the following section.

3.3.2 Posterior probability of the parameters of model 1

In order to know the posterior distribution of the parameters, we consider its prior distributions shown in Table 3.3 and the perturbation kernel definition for each one (see [79]). Then, we apply the ABC SMC approach. In each sample $\{1, \ldots, T\}$ the number of values for the parameters is $N = 1000$. Taking into

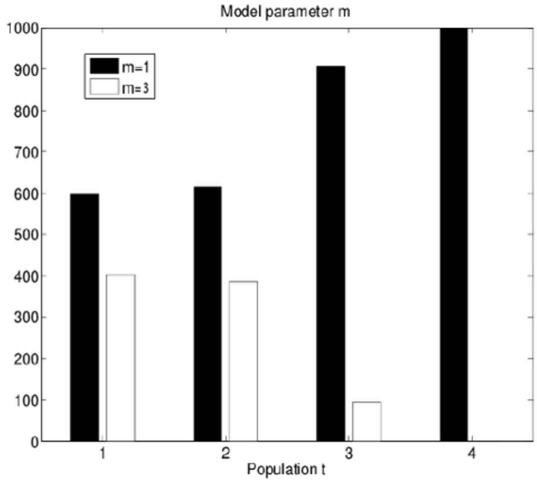


Figure 3.5: Evolution of the number of parameters vectors, $\theta(m)$, corresponding to model 1 and model 3 in each population $t = 1, 2, 3, 4$.

account the application on ordinary differential systems presented in [79], we take $T = 4$. The four ϵ_t values are $\epsilon_1 = 0.0223$, $\epsilon_2 = 0.0148$, $\epsilon_3 = 0.0099$ and $\epsilon_4 = 0.0066$. The distance function $d(\cdot;\cdot)$ is defined by the root mean square error. Note that the output of the ABC SMC algorithm consists of 1000 different vectors $\theta = (\mu, d_N, d_{c_1}, d_{c_2}, d_{c_3}, \beta, \gamma, \sigma, \epsilon)$. If we take any of these parameter vectors and we solve the dynamical system, Eqs. (3.1)–(3.5), the distance between the obtained solution (for years shown in Table 3.1) and the observed prevalences is less than ϵ_4 = 0.0066. This value is considered taking into account the deterministic estimation of the model in the mean square sense.

The distance between the observed prevalence data (Table 3.1) and the deterministic solution is 0.00645, therefore the lowest distance to be reached is expected to be close to this number and we choose the tolerance level ϵ_T accordingly. In our case we have defined ϵ_4 considering a difference of 2% with the mean square error (0.00645). Note that we choose ϵ_1, ϵ_2, ϵ_3 and ϵ_4 in decreasing order and we select the values to ensure that the distribution gradually evolves towards the posterior one, i.e., the distribution defined by ϵ_4. The others ϵ_t are chosen increasing 50% the value of ϵ_{t+1}, i.e. $\epsilon_3 = \epsilon_4 + 0.5 \times \epsilon_4$, $\epsilon_2 = \epsilon_3 + 0.5 \times \epsilon_3$ and $\epsilon_1 = \epsilon_2 + 0.5 \times \epsilon_2$. Note that we performed the computations with other values for ϵ_t and the results were similar.

Table 3.4 summarizes the posterior distribution of the final parameter sample.

For example, we obtain that the median time ($1/\gamma$) that an occasional consumer takes to become a habitual consumer is approximately 15 years ($1/0.067643 = 14.78$ years) –95% confidence interval: 13-18 years–. Also, as indicated by the median value of σ, 14 years (1/0.072904) are necessary for a regular consumer to be a habitual consumer.

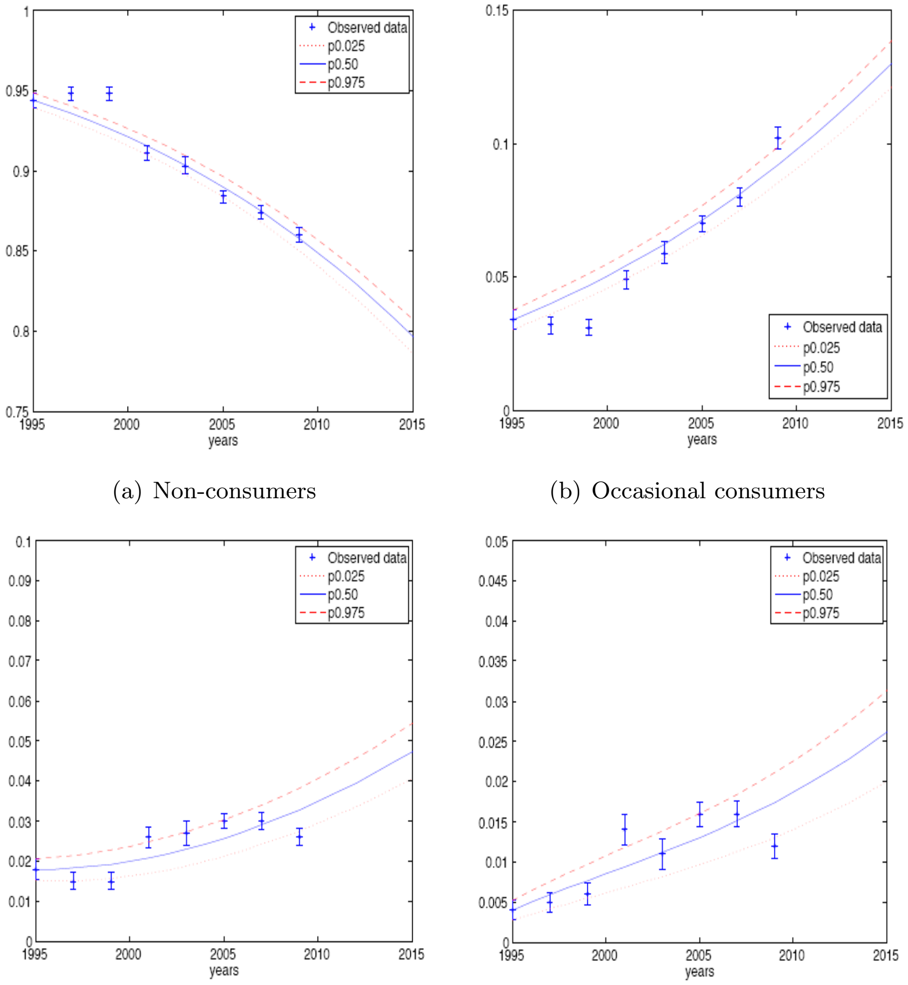
Table 3.4: Summary of the posterior (fitted) probability distributions for the model 1 parameters.

Parameter	2.5th percentile	25th percentile	50th percentile	75th percentile	97.5th percentile
μ	0.008392	0.008615	0.008791	0.008980	0.009200
d_N	0.009287	0.009730	0.010068	0.010455	0.010870
d_{c_1}	0.037407	0.041954	0.047139	0.053695	0.067030
d_{c_2}	0.045338	0.056014	0.062998	0.069565	0.079426
d_{c_3}	0.057828	0.070810	0.076774	0.082069	0.086483
β	0.116593	0.122941	0.127008	0.132320	0.143290
γ	0.057021	0.063710	0.067866	0.071813	0.078001
σ	0.055923	0.065737	0.072790	0.079982	0.089422
ε	3.32×10^{-6}	2.42×10^{-5}	4.36×10^{-5}	6.21×10^{-5}	8.52×10^{-5}

3.3.3 Predictions

To assess the predictive performance of our model, we ran 1000 times the dynamical model selected (Eqs.(3.1)–(3.5)) using the posterior parameter values inferred by ABC SMC. We obtain the posterior distribution of model predictions required by running the cocaine consumption model once for each parameter set stored (see Table 3.4). This means that we obtained 1000 values for the prediction of the prevalences at a given year. Then, we can define 95% credible intervals for each year for the proportions of non-consumer, occasional consumers, regular consumers and habitual consumers.

Figure 3.6 displays the observed cocaine consumption prevalence together with the corresponding model predictions intervals for each year. It shows predictions during the period 1995–2009 and the following six years. In this work, all the computations were performed using *Matlab* [34].



(a) Non-consumers

(b) Occasional consumers

(c) Regular consumers

(d) Habitual consumers

Figure 3.6: Probabilistic predictions for cocaine consumption population in Spain for the following years until 2015. The points represent the observed data, the dotted lines are the 2.5% and 97.5% percentiles and the continuous one is the median of the 1000 outputs of the model for each year. The error bars for cocaine consumption prevalence in period 1995-2009 are also shown.

Figure 3.6 shows that the model predicts a decreasing trend in non-consumer subpopulation, $N(t)$. Also, there is an increasing trend in all the populations of cocaine consumers, that is, occasional consumers, $C_o(t)$, regular consumers, $C_r(t)$, and habitual consumers, $C_b(t)$. However, the interval for cocaine consumption in 2009, according to the data from Table 3.1 shows a slight decrease in the percentage of regular consumers and habitual consumers. This decrease may be due to the effect of the cocaine consumption control plan proposed by the Spanish Health Ministry in 2007 [62]. Although our model predicts an increase in cocaine consumer subpopulations for the next years, we can observe that the interval for cocaine consumption in 2009 and the interval for model predictions in the same year have a non-zero intersection. Then, our predictions are in accordance with the interval for real data.

In Table 3.5, some of the numerical values depicted in Figure 3.6 are presented, for the period 2011–2015.

Table 3.5: 95% Credible intervals (CI) for the period 2011–2015 of percentage of non-consumers, occasional consumers, regular consumers and habitual consumers. Credible intervals are calculated considering the 2.5% and 97.5% percentile for each year.

	N	C_o	C_r	C_b
Year 2011				
Median	0.8393	0.1037	0.0371	0.0200
95% CI	[0.8306, 0.8478]	[0.0963, 0.1110]	[0.0315, 0.0431]	[0.0151, 0.0241]
Year 2012				
Median	0.8293	0.1099	0.0395	0.0214
95% CI	[0.8201, 0.8383]	[0.1022, 0.1175]	[0.0336, 0.0457]	[0.0162, 0.0258]
Year 2013				
Median	0.8189	0.1163	0.0420	0.0229
95% CI	[0.8091, 0.8283]	[0.1083, 0.1243]	[0.0359, 0.0485]	[0.0174, 0.0275]
Year 2014				
Median	0.8079	0.1231	0.0446	0.0245
95% CI	[0.7977, 0.8179]	[0.1146, 0.1313]	[0.0382, 0.0514]	[0.0187, 0.0294]
Year 2015				
Median	0.7965	0.1300	0.0474	0.0262
95% CI	[0.7858, 0.8070]	[0.1211, 0.1386]	[0.0408, 0.0545]	[0.0200, 0.0314]

Therefore, we can conclude that if there are no changes in current cocaine consumption habits over the next few years, the model predicts that 79.65% ([78.58%, 80.70%]), 13.0% ([12.11%, 13.86%]), 4.74% ([4.08%, 5.45%]) and 2.62% ([2.00%, 3.14%]) of 15-64-year-old individuals in Spain will be, in the year 2015, non-consumers, occasional consumers, regular consumers and habitual consumers, respectively.

Now, we are going to find out which model parameters (of model 1) are the most sensitive to the selected model. This sensitivity analysis allows us to propose some strategies to control cocaine consumption.

3.3.4 Sensitivity analysis

We use principal component analysis (PCA) to quantify the sensitivity of the system. Principal component analysis is a mathematical procedure that uses an orthogonal transformation to convert a set of initially correlated variables into a set of uncorrelated variables called *principal components.* The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each successive component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components.

Given the parameter vector $\theta = (\theta_1, ..., \theta_p)$, the principal components (PCs) are given by the linear combination of θ, $\chi_i = a_{i1}\theta_1 + ... + a_{ip}\theta_p$, for $i = 1, ...k$ and for $k \leq p$, where p denotes the dimension of the parameter vector.

In our case, the parameter vector is $\theta = (\mu, d_N, d_{c_1}, d_{c_2}, d_{c_3}, \beta, \gamma, \sigma, \varepsilon)$, and $p = 9$ is the number of parameters of our model and the dimension of θ.

The principal components (PCs) are the eigenvectors of the variance-covariance matrix, denoted by Σ. Σ is a square matrix whose main diagonal are the variances of each of the one-dimensional distributions and elements outside the diagonal are the corresponding covariances between two variables. In our case, the variables are the model parameters, specifically, the last population of N particles obtained from the ABC algorithm.

$a_i = (a_{i1}, ..., a_{ip})$ is the normalized eigenvector associated with the ith eigenvalue of Σ, λ_i, and a_{ij} describes the projection of parameter θ_j onto the ith eigenparameter, θ_i.

The variance of the ith PC is given by λ_i and the total variance of all PCs equals $\sum_{i=1}^{p} \lambda_i = trace(\Sigma)$. Therefore, the eigenvalue λ_i associated with the ith PC explains a proportion

$$\frac{\lambda_i}{trace(\Sigma)}$$

of the variation in the population of points. The smaller is λ_i, the more sensitive is the system to the variation of the eigenparameter χ_i (because it explains a smaller proportion of the total variance). Therefore, in contrast to the interest in the first PC in most PCA applications, our main interest lies in the smallest PC. The last PC extends across the narrowest region of the posterior parameter distribution and, therefore, provides information on parameters to which the model is the most sensitive. In other words, the smallest PCs correspond to stiff parameter combinations, while the larger PCs may correspond to weak parameter combinations [29].

Figure 3.7 shows how much the variance is explained by each PC. Since the dimension of our parameter vector $\theta = (\mu, d_N, d_{c_1}, d_{c_2}, d_{c_3}, \beta, \gamma, \sigma, \varepsilon)$ is 9, we have 9 principal components. We can observe that the smallest PC corresponds to the ninth component.

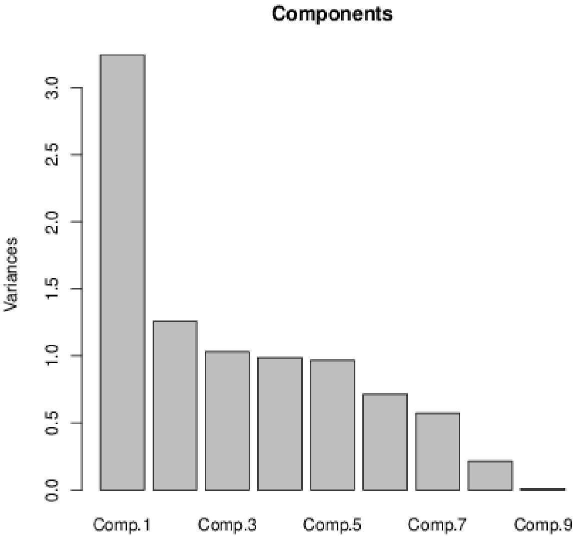


Figure 3.7: PCA of the set of accepted particles (with all the parameters). The first PC explains 36.03% of the total variance, the second 13.95%, the third 11.46%, the fourth 10.96%, the fifth 10.96%, the sixth 7.94%, the seventh 6.39%, the eighth 2.38% and the ninth 0.01% of the variance.

Table 3.6 summarizes which parameters contribute the most to these PCs, that is, it shows the contribution of each parameter on each component.

Table 3.6: Components matrix PCA (with all the parameters).

	Component								
Parameter	1	2	3	4	5	6	7	8	9
μ	-0.100	-0.015	0.492	0.842	0.187	0.027	0.053	-0.004	0.003
d_N	-0.056	0.137	-0.687	0.231	0.671	0.044	0.008	-0.017	-0.006
d_{c_1}	0.691	-0.597	-0.088	0.012	0.025	0.367	0.124	-0.057	0.064
d_{c_2}	0.835	-0.035	-0.007	0.053	0.093	-0.385	-0.220	0.304	0.025
d_{c_3}	0.613	0.451	0.043	-0.061	0.013	-0.169	0.622	-0.008	0.007
β	0.913	-0.370	0.014	0.031	0.012	0.136	0.042	0.009	-0.086
γ	0.819	0.281	0.072	0.018	0.034	-0.216	-0.302	-0.324	0.010
σ	0.401	0.672	0.094	-0.028	-0.029	0.578	-0.170	0.115	0.002
ϵ	-0.108	-0.091	0.541	-0.464	0.687	0.024	-0.002	0.000	0.000

As we commented above, our interest is in the ninth component. Looking at this, the analysis reveals that the last PC mainly extends in the direction of a linear combination of parameters β and d_{c_1}. Note that the values assigned to these parameters (-0.086 and 0.064, respectively) are higher than the others and that we have considered only the parameters whose weight in the component is higher than 0.05. Looking at the eighth component, the model is also somewhat less sensitive to variation in γ, d_{c_2} and σ (values of -0.324, 0.304 and 0.115, respectively). The model is therefore the least sensitive to changes in parameters d_N and ε, which is also supported by the composition of the other PCs. Thus, we can conclude that the model is most sensitive to changes parameters β, d_{c_1}, γ, d_{c_2} and σ.

With the aim to analyse the strategies of the Spanish Government against drug abuse, natality and mortality rates are not included in the PCA analysis, since we cannot change the value of these parameters to reduce cocaine consumption. This way, we carry out the PCA analysis, considering only four parameters: β, γ, σ and ε. Results can be seen in Figure 3.8 and Table 3.7.

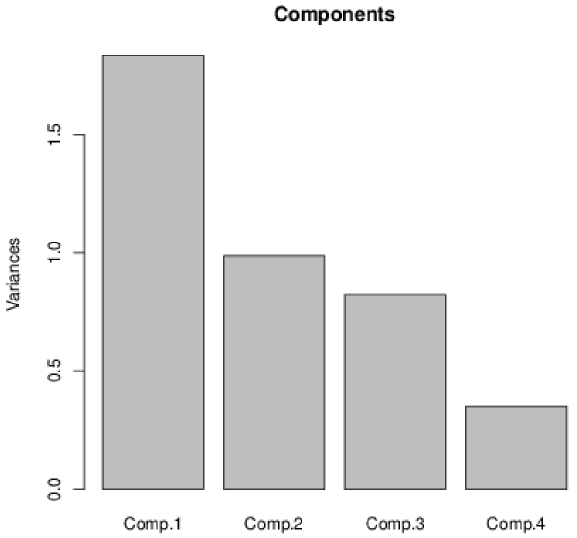


Figure 3.8: PCA of the set of accepted particles (without natality and mortality parameters). The first PC explains 20% of the total variance, the second 19.64%, the third 16.48% and the fourth 7.18% of the variance.

Table 3.7: Components matrix PCA (without natality and mortality parameters).

	Component			
Parameter	1	2	3	4
β	0.792	0.064	-0.492	0.356
γ	0.889	0.080	-0.100	-0.440
σ	0.627	0.060	0.757	0.175
ϵ	-0.161	0.987	-0.007	0.002

Looking at the last component, we can conclude that the model is most sensitive to changes in parameters γ, β and σ (values assigned of -0.440, 0.356 and 0.175), respectively, and therefore the least sensitive to changes in parameter ε. This outcome agrees with the results obtained considering the nine components, and with the information obtained from the previous chapter where we have shown that prevention policies (the ones related with parameters β, γ and σ) are the best effective strategy to modify the cocaine consumption, i.e., to reduce the population of regular and habitual consumers.

3.4 Conclusions

In this chapter we have presented an application of the Approximate Bayesian Computation scheme (ABC scheme) for model selection describing the evolution of cocaine consumption in Spain and we have been able to specify the dynamics of this process selecting the model that best explains the observed proportions of non-consumers, occasional consumers, regular consumers and habitual consumers. Taking into account the obtained results, we can conclude that if an individual starts with cocaine consumption he/she can increase his/her consumption to problematic levels.

The selection of model 1 instead of model 2 allows us to say that regular consumers (that is, individuals with a non-addictive level) do not reduce significantly the level of cocaine consumption, maybe because they do not consider that this consumption level can be a problem. On the other hand, as model 1 is considered better than model 3, we can reject the inclusion of a new subpopulation of habitual cocaine consumers in therapy, maybe because this subpopulation is very small compared to the others.

In addition, the ABC scheme provides us with an approximation to the posterior probability distributions for the parameters of the selected model. Thus, using an ABC approach, we obtain probabilistic predictions of the evolution of the cocaine consumption in Spain over the next few years. Particularly, the model predicts that 79.65% ([78.58%, 80.70%]), 13.0% ([12.11%, 13.86%]), 4.74% ([4.08%, 5.45%]) and 2.62% ([2.00%, 3.14%]) of 15-64-year-old individuals in Spain will be, by year 2015, non-consumer, occasional consumer, regular consumer and habitual consumer, respectively. If we compare these predictions with the obtained in the previous chapter (using the LHS method), we can observe than the credible intervals achieved with the ABC approach are smaller.

Furthermore, we carry out a sensitivity analysis of the selected model (model 1), that is the one that best explains the cocaine consumption in Spain. To do this, we have used PCA to identify which parameters influence more the model output, i.e., the prevalence of cocaine consumption in Spain. In this case, the conclusions obtained are similar to the ones in Chapter 2, where we also obtained that we have to focus the efforts on prevention policies to achieve a decrease in cocaine consumption.

In the next chapter, we propose another method to quantify the uncertainty of dynamical systems. Using the bootstrap method, we will obtain a new confidence interval estimation for the cocaine consumption model in Spain.

Chapter 4

The bootstrap method for confidence interval parameters estimation

To obtain accurate results from mathematical and computer models is often complicated due to the presence of uncertainties in experimental data used to estimate parameter values. In this chapter, we assess the uncertainty about parameters estimates and model predictions using the bootstrap method for confidence interval estimation. We will apply it to the dynamics of cocaine consumption in Spain model presented in Chapter 2 and improved in Chapter 3 (model 1), which is the one that best describes the cocaine consumption dynamics, as we demonstrated in the previous chapter. Thus, we obtain future predictions of cocaine consumption in Spain by credible intervals. Additionally, a sensitivity analysis is carried out in order to identify the most important parameters of the model. In this case, the method applied is different from the other ones employed in the previous chapters, although the objective is the same: to design strategies to control cocaine consumption.

4.1 Introduction

Confidence interval analysis is one of the most significant statistical tests to validate parameter reliability. However, while tools for estimating parameters have improved and are now easy to use in system dynamics software, less attention has been paid to the problem of finding confidence intervals around the estimated parameters. The goal of bootstrap confidence interval theory is to calculate confidence limits for the parameters from their distribution.

Bootstrapping, introduced in the late 1970s [21], is a widely used and robust method but its use in dynamic models is rare. Bootstrapping can handle violations of the maintained assumptions of traditional methods (e.g., the likelihood ratio method) such as error terms autocorrelated, heteroskedastic, censored and non-normally distributed. It is also valid for small samples whereas traditional methods are valid only asymptotically (which involves taking a large sample).

The main idea behind bootstrapping is resampling, that is, finding many "new" data sets by resampling the original data set, and estimate the parameter value(s) for each of these "new" data sets, generating a distribution of parameter estimates. Using the resulting empirical distribution of parameters, estimate the confidence intervals.

Considering the general procedure presented by G. Dogan [18], we will study error terms for the estimated parameters and resample these terms using a residual bootstrapping.

4.2 Methods

In this section we present the bootstrap method used. Following the general procedure presented by G. Dogan, we use the error terms for this purpose. An error term is actual data minus model output. First, we study the error terms to determine their probability distribution and then we will apply the parametric bootstrap method on the residuals to estimate the confidence intervals.

4.2.1 Error terms analysis

In order to know the probability distribution of the error terms (see step 3 in the process presented in section 4.2.2), the following points should be kept in mind to designing an appropriate resampling scheme for bootstrapping:

1. *Check whether the error terms are autocorrelated.*

2. *Check whether the error terms are normally distributed.*

Before checking these points, we compute the error terms. First, we fit the model to the actual data by optimizing the parameter values, using the Nelder-Mead algorithm described in Chapter 2. Note that the model is described by the equations (4.1)- (4.5).

$$
\begin{aligned}
\frac{dN(t)}{dt} &= \mu P(t) - d_N N(t) - \beta \frac{N(t)(C_o(t) + C_r(t) + C_b(t))}{P(t)} + \\
&\quad + \varepsilon C_b(t), \qquad (4.1) \\
\frac{dC_o(t)}{dt} &= \beta \frac{N(t)(C_o(t) + C_r(t) + C_b(t))}{P(t)} - d_{c_1} C_o(t) - \gamma C_o(t) \qquad (4.2) \\
\frac{dC_r(t)}{dt} &= \gamma C_o(t) - d_{c_2} C_r(t) - \sigma C_r(t) \qquad (4.3) \\
\frac{dC_b(t)}{dt} &= \sigma C_r(t) - d_{c_3} C_b(t) - \varepsilon C_b(t) \qquad (4.4) \\
P(t) &= N(t) + C_o(t) + C_r(t) + C_b(t) \qquad (4.5)
\end{aligned}
$$

Then, we compute the error terms for the optimum parameter values. The results can be seen in Table 4.1.

Table 4.1: Residual or error terms, denoted as $e_1(t)$, $e_2(t)$, $e_3(t)$ and $e_4(t)$. $N(t)$, $C_o(t)$, $C_r(t)$ and $C_b(t)$ are the real data (Table 3.1) and $\hat{N}(t)$, $\hat{C}_o(t)$, $\hat{C}_r(t)$ and $\hat{C}_b(t)$ are the model predictions.

Year	$e_1(t)$	$e_2(t)$	$e_3(t)$	$e_4(t)$
	$N(t)$ - $\hat{N}(t)$	$C_o(t)$ - $\hat{C}_o(t)$	$C_r(t)$ - $\hat{C}_r(t)$	$C_b(t)$ - $\hat{C}_b(t)$
1997 (t = 1)	-0.012389272	0.007266849	0.003880928	0.001241495
1999 (t = 2)	-0.021936760	0.014162516	0.005429750	0.002344494
2001 (t = 3)	0.004241248	0.002756792	-0.003428359	-0.003569682
2003 (t = 4)	0.000027072	0.000112998	-0.001734472	0.001594402
2005 (t = 5)	0.005305699	-0.002714428	-0.001506171	-0.001085100
2007 (t = 6)	-0.000030146	-0.003683237	0.002254647	0.001458736
2009 (t = 7)	-0.003073993	-0.015767403	0.010555656	0.008285740

Once the error terms are calculated, we are going to check the two points discussed above.

1. Check whether the error terms are autocorrelated.

Partial autocorrelation measures the degree of association between two time instants, when the effect of a set of controlling random variables is removed. It is the linear dependence of a variable with itself at two different points in time. Thus, if the value of a variable at time t depends on its value in the previous instant, plus a random term, the process is *autoregressive of first order (AR(1))*. If the dependency is established with the p previous values, the process will be *autoregressive of order p (AR(p))*. Our aim is to study if the error terms e_i, for i=1, 2, 3, 4, in t instant depends on the p previous values. Thus, the *AR(p)* process is defined as:

$$e_i(t) \quad = \quad \rho_{i,1}e_i(t-1) + \rho_{i,2}e_i(t-2) + ... + \rho_{i,p}e_i(t-p) + \varepsilon_i(t)$$

where:

- $\varepsilon_i(t)$ is a white noise process and, therefore, with zero mean, constant variance and zero covariance,
- $\rho_{i,j}$ are partial autocorrelation coefficients, that measure the additional effect of the variable $e_i(t-j)$ on $e_i(t)$, defined as [28]:

$$\rho_{i,j} \quad = \quad \frac{Cov(e_i(t), e_i(t-j))}{\sqrt{Var(e_i(t)Var(e_i(t-j))}}, j = 1, 2...t-1.$$

To detect the presence of autocorrelation we can use graphical methods and hypothesis tests. Partial autocorrelation plots are a commonly used tool to identify the order of an autoregressive model [6]. To determine when an estimated partial autocorrelation coefficient is considered zero, despite this empirical value, statistical contrasts have been performed to set confidence bands above which the coefficients are significant. If all the correlation coefficients are within these limits the process is considered white noise. When there are no coefficients within the bands, we have to find a pattern of behaviour as autoregressive scheme.

In Figure 4.1 the partial autocorrelation function (PACF) is plotted for each error term. As we can see, all the coefficients are inside these limits (between -1 and 1), for any possible lag. Then, we can affirm that the error terms are not autocorrelated.

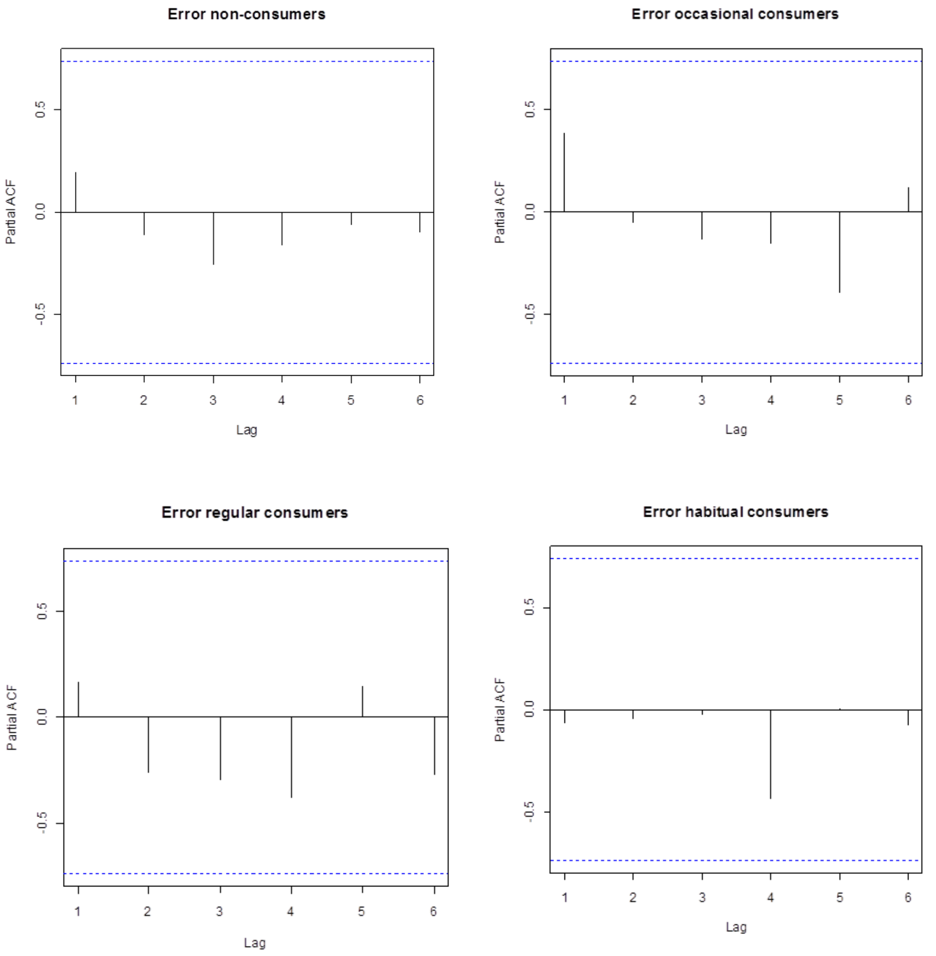


Figure 4.1: PACF for each error term.

This graphical method is complemented by other numerical methods such as the Ljung-Box test. It is one of the most used statistical tests to check the hypothesis of independence in a given time series. Instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a certain number of lags.

If the residuals are independent, their first p autocorrelations are zero for any value of p (where p is the lag). Then, the contrast of Ljung-Box tests the null hypothesis that the first p autocorrelations are zero. That is:

$$H_0\text{: } \rho_{i,1} = \rho_{i,2} = ... = \rho_{i,p} = 0 \text{ (No autocorrelation)}$$
$$H_1\text{: } \rho_{i,j} \neq 0 \text{ for j} \in \{1, 2, ..., p\} \text{ (Autocorrelation)}$$

The statistic test is [44]:

$$Q \quad = \quad n(n+2)\sum_{j=1}^{p} \frac{\hat{\rho}_{i,j}^2}{N-j} \sim \chi_p^2$$

where:

- n is the sample size,
- $\hat{\rho}_{i,j} = \dfrac{\sum\limits_{t=j+1}^{n} e_i(t)e_i(t-j)}{\sum\limits_{t=1}^{n} e_i(t)^2}$ is the sample autocorrelation at lag j,
- p is the number of lags being tested.

Results obtained after checking the Ljung-Box test are shown in Table 4.2.

Table 4.2: Ljung-Box test

error term	lag 1	lag 2	lag 3	lag 4	lag 5
$e_1(t)$	0.525	0.792	0.630	0.560	0.667
$e_2(t)$	0.206	0.419	0.598	0.578	0.078
$e_3(t)$	0.584	0.626	0.396	0.212	0.226
$e_4(t)$	0.818	0.963	0.994	0.420	0.538

None of the test statistic values are statistically significant (p-value$>$ 0.05) therefore the claim that there is autocorrelation should be rejected.

2. Check whether the error terms are normally distributed.

Normality of the distribution of errors is determined by using non-parametric tests. We carry out the Kolmogorov-Smirnov and Shapiro-Wilk tests to check the normality of each error term and, then, we complete the study using Mardia's multivariate test to check the normality of the random vector $(e_1(t),\, e_2(t),\, e_3(t))$. Note that $e_4(t) = -\sum_{i=1}^{3} e_i(t)$. That is why we will not consider $e_4(t)$ in the Mardia's test.

The Kolmogorov-Smirnov test tries to measure the fit between the empirical distribution function of a sample and the theoretical distribution function. It considers the hypothesis test:

$$H_0\text{: Normal distribution.}$$
$$H_1\text{: No normal distribution.}$$

This test is based on evaluating the statistic [41, 76]:

$$D_n = \sup_x |F_n(x) - F(x)|$$

where:

- $\sup_x$ is the supremum of the set of distances.
- $F_n = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x}$ is the empirical distribution function, where $I_{X_i \leq x}$ is the indicator function, equal to 1 if $X_i \leq x$ and equal to 0 otherwise.
- F is the theoretical distribution function. In our case, it will be the normal distribution, i.e., $F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2} du.$

The Shapiro-Wilk test is based on studying the fit of the data plotted on the probability graph in which each data is a point whose abscissa is the observed value of probability for a given value of the variable, and whose ordinate is the expected value of probability. This statistic measures the strength of a line adjustment. The higher this statistic is the greater disagreement with the normal line. Therefore we reject the null hypothesis. In this test the null and the alternative hypothesis are the same as the ones used for the previous test.

The Shapiro-Wilk test tests if a sample $x_1, ..., x_n$ comes from a normally distributed population using the following statistic test [74]:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_{(i)} - \overline{x})^2}$$

where:

- $x_{(i)}$ is the ith order statistic, i.e., the ith- smallest number in the sample,
- $x = (x_1 + ... + x_n)/n$ is the sample mean,
- the constants a_i are defined considered the following vector expression:

  $$(a_1, ..., a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$

  where $m = (m_1, ..., m_n)^T$ and $m_1, ..., m_n$ are the expected values of the order statistics $x_{(i)}$ of independent and identically-distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics.

Goodness-of-fit analysis suggests that each error term is normally distributed. The p-values for each error term and for both tests are presented in Table 4.3.

Table 4.3: Error terms normality

	$e_1(t)$	$e_2(t)$	$e_3(t)$	$e_4(t)$
Kolmogorov-Smirnov				
p-value	0.200	0.200	0.200	0.159
Shapiro-Wilk				
p-value	0.185	0.937	0.655	0.375

In order to complete the normality analysis of the error terms, Mardia's multivariate normality test is applied to the random vector $(e_1(t), e_2(t), e_3(t))$. Note that $e_4(t) = -\sum_{i=1}^{3} e_i(t)$.

Multivariate normality tests check if a given set of data are similar to the multivariate normal distribution. The null hypothesis is that the data set is similar to the normal distribution therefore a sufficiently small p-value indicates non-normal data. That is, it considers:

H_0: Multivariate normal distribution.

H_1: Non-multivariate normal distribution.

Mardia's test is based on multivariate extensions of skewness and kurtosis measures. For a random sample $X = (x_1, ..., x_n)$ from a p-variate distribution, Mardia's [47, 48] defined the p-variate skewness and kurtosis statistics by:

$$\begin{aligned} b_{1p} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{\left(x_{(i)} - \bar{x}\right)' S^{-1}\left(x_j - \bar{x}\right)\}^3 \\ b_{2p} &= \frac{1}{n}\sum_{i=1}^{n}\{\left(x_{(i)} - \bar{x}\right)' S^{-1}\left(x_i - \bar{x}\right)\}^2 \end{aligned}$$

where S is the sample covariance matrix.

Mardia's uses the skewness and kurtosis statistics to test for multinormality. If the data come from a multivariate normal distribution, i.e., H_0 is accepted, then:

$$\frac{n}{6}b_{1p} \sim \chi^2_{\left(\frac{p(p+1)(p+2)}{6}\right)}$$

$$\sqrt{n}\frac{(b_{2p} - p(p+2))}{\sqrt{8p(p+2)}} \sim N(0,1)$$

The p-values for Mardia's test are shown in Table 4.4:

Table 4.4: Mardia's test

	statistic	p-value
skewness	5.07483218	0.82188626
kurtosis	9.43779247	0.08957112

Since all p-values are higher than 0.05, we can accept that vector (e_1, e_2, e_3) presents a multivariate normal distribution. To be precise, we accept that:

$$(e_1, e_2, e_3) \sim N_3 \left[\begin{pmatrix} \mu_{e_1} \\ \mu_{e_2} \\ \mu_{e_3} \end{pmatrix}, \begin{pmatrix} \sigma^2_{e_1} & cov(e_1, e_2) & cov(e_1, e_3) \\ cov(e_2, e_1) & \sigma^2_{e_2} & cov(e_2, e_3) \\ cov(e_3, e_1) & cov(e_3, e_2) & \sigma^2_{e_3} \end{pmatrix} \right],$$

where μ_{e_i} and σ_{e_i}, $i = 1, 2, 3$, are the mean and the standard deviation of e_i, respectively, and cov(e_i, e_j) is the covariance between e_i and e_j, for $i, j = 1, 2, 3$, and $i \neq j$.

All these parameters can be easily estimated using the descriptive statistics of the error terms, calculated from Table 4.1. Finally, we obtained the following distributions for the error terms:

$$(e_1, e_2, e_3) \sim N_3 \left[\begin{pmatrix} -0.00398 \\ 0.00030 \\ 0.00221 \end{pmatrix}, \begin{pmatrix} 9.6576 \times 10^{-5} & -5.7724 \times 10^{-5} & -2.6800 \times 10^{-5} \\ -5.7724 \times 10^{-5} & 8.8315 \times 10^{-5} & -1.3318 \times 10^{-5} \\ -2.6700 \times 10^{-5} & -1.3318 \times 10^{-5} & 2.3995 \times 10^{-5} \end{pmatrix} \right]$$

Therefore, we can apply the bootstrap method to estimate cocaine consumption in Spain.

4.2.2 The bootstrap method

The residual method used is implemented as follows [18]:

1. Fit the model to the actual data (see Table 3.1) by optimizing the parameter values. We will use the Nelder-Mead method to search in the parameter space to find parameter values that minimize the sum of squared errors.

2. Compute the error terms for the optimum parameter values. An error term is actual data minus model output.

   $$e(t) = y(t) \text{ - } \hat{y}(t)$$

   where:

   - *e(t)* is the vector made up of the error terms,

   - *y(t)* is the vector defined by the real data,
   - $\hat{y}(t)$ is the vector generated by the model output.

3. Resample the error terms using a parametric technique and obtain new error values for each value of t. Denote the new error terms $e^{i*}(t)$. Each i value $(i = 1, 2, ..., n)$ represents a new error term set. Thus, $e^{i*}(t)$ is the resampled error term for the ith data set at time t. As we have seen in the previous section, $e_4(t) = -\sum_{i=1}^{3} e_i(t)$ and

$$(e_1, e_2, e_3) \sim N_3 \left[\begin{pmatrix} -0.00398 \\ 0.00030 \\ 0.00221 \end{pmatrix}, \begin{pmatrix} 9.6576 \times 10^{-5} & -5.7724 \times 10^{-5} & -2.6800 \times 10^{-5} \\ -5.7724 \times 10^{-5} & 8.8315 \times 10^{-5} & -1.3318 \times 10^{-5} \\ -2.6700 \times 10^{-5} & -1.3318 \times 10^{-5} & 2.3995 \times 10^{-5} \end{pmatrix} \right]$$

   Then, we use this distribution to resample the error terms.

4. Generate new perturbed data sets (y^{i*}(t)) by adding the resampled error terms to the model output:

$$y^{i*}(t) = \hat{y}^i(t) + e^{i*}(t)$$

5. For each new data perturbation calculated, compute the parameters that best fit the model with the perturbed dataset.

   Repeat steps 3, 4 and 5 many (n) times to obtain a sufficiently large bootstrap sample for the parameters of the model.

At the end of the process, we will have n estimates for the parameters. Then we will use these samples to run n times the model. The results of these evaluations will allow us to estimate confidence intervals for the model outputs, i.e., for the prevalence of cocaine consumption over the next few years. This approach has been used in other similar studies [65].

4.3 Results

4.3.1 Predictions

4.3.1.1 Generating new perturbed data

Once we have studied error terms, we are going to resample these terms using parametric bootstrapping. Bearing in mind data from Table 3.1, we generate random shortlists of three components (e_1, e_2, e_3) following the multivariate distribution given above. Thus, we have seven vectors $(e_1(t),\ e_2(t),\ e_3(t))$ for $t = 1997, 1999, 2001, 2003, 2005, 2007$ and 2009.

We add them to the model output, obtaining a new set of perturbed data. For each new data perturbation calculated, we compute the parameters that best fit the model with the new perturbed dataset, in the least square sense, using the same Nelder-Mead procedure we used to estimate the parameters of the model in Chapter 2, and store them.

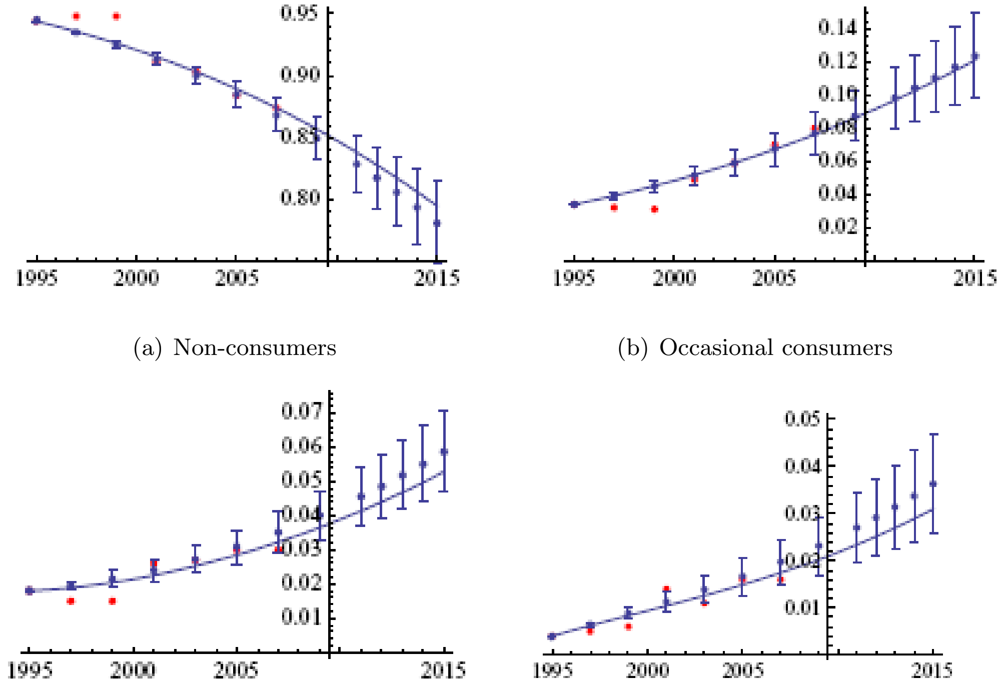
We repeat this procedure 5000 times in order to obtain 5000 set of parameters that fit each set of perturbed data.

4.3.1.2 Obtaining confidence intervals for model outputs

For each one of the 5000 set of parameters, we solve the system of differential equations (4.1)-(4.5) and compute the output of the solution, i.e., cocaine consumption for the four subpopulations $N(t)$, $C_o(t)$, $C_r(t)$ and $C_b(t)$, for $t = 1997, 1999, 2001, 2003, 2005, 2007$ and 2009. Thus, for each t and for each subpopulation, we have a set of 5000 model output values.

Then, we compute the mean and the 95% confidence interval by percentiles 2.5 and 97.5. Obtained results can be seen in Figure 4.2. In these graphs we can observe, for each subpopulation, the data from Table 3.1 (points), the deterministic model prediction (line) and the 95% confidence intervals (error bars). The points inside the confidence intervals are the mean of the 5000 outputs for every subpopulation in every time instant. We can observe that confidence intervals contain the real data from 2001 to 2009 (data obtained from Table 3.1).

In Table 4.5, some of the numerical values depicted in Figure 4.2 are presented, for the period 2011–2015. As we saw in the previous chapters, our model predicts a decreasing trend in non-consumer subpopulation $(N(t))$ and an increasing trend in all the populations of cocaine consumers $(C_o(t),\ C_r(t)$ and $C_b(t))$. In fact, if there are no changes in current cocaine consumption habits over the next few years, the model predicts that 78.15% ([0.7542,0.8082]), 12.38% ([0.1043,0.1442]), 5.87% ([0.0485,0.0686]) and 3.612% ([0.0254,0.0474]) of 15-64-year-old individuals in Spain will be, in the year 2015, non-consumers, occasional consumers, regular consumers and habitual consumers, respectively.



(a) Non-consumers

(b) Occasional consumers

(c) Regular consumers

(d) Habitual consumers

Figure 4.2: Model predictions over the next few years. The error bars corresponding to 95% confidence intervals in the same time instants as data.

4.3.2 Sensitivity analysis

Uncertainty and sensitivity analysis offer a way to assess the suitability of models and to establish what factors affect the model outputs. In our case, the knowledge of the most sensitive parameters can help us to design strategies to control cocaine consumption. In this section, we use PRCC (partial rank correlation coefficient) to identify the most important model parameters. PRCC analysis is a sensitivity analysis method that calculates the PRCC for the input variables and the outputs. This method is more robust than simple correlation coefficient analysis approach because it uses rank transformation statistic. However, PRCC is not useful for quantifying how much change occurs in the output variables by changing the value of the input parameters. Thus, PRCC can be informative on which parameters to target if we want to achieve specific goals. Calculation of PRCC enables the

Table 4.5: 95% credible intervals (CI) for the period 2011–2015 of percentage of non-consumers (N), occasional consumers (C_o), regular consumers (C_r) and habitual consumers (C_b). Credible intervals are calculated considering the 2.5% and 97.5% percentile for each year.

	N	C_o	C_r	C_b
Year 2011				
Median	0.8290	0.0986	0.0456	0.0270
95% CI	[0.8107,0.8468]	[0.0844,0.1134]	[0.03803,0.0528]	[0.0194,0.0347]
Year 2012				
Median	0.8179	0.1046	0.0486	0.0291
95% CI	[0.7975, 0.8377]	[0.0891,0.1206]	[0.0404,0.0564]	[0.02071,0.03755]
Year 2013				
Median	0.8062	0.1108	0.0518	0.0313
95% CI	[0.7837,0.8283]	[0.0940,0.1282]	[0.0430,0.0602]	[0.0222,0.0406]
Year 2014				
Median	0.7941	0.1172	0.0551	0.0337
95% CI	[0.7693,0.8184]	[0.0991,0.1361]	[0.0457,0.0643]	[0.0237,0.0439]
Year 2015				
Median	0.7815	0.1238	0.0587	0.03612
95% CI	[0.7542,0.8082]	[0.1043,0.1442]	[0.0485,0.0686]	[0.0254,0.0474]

determination of the statistical relationships between each input parameters and each outcome variable when controlling the effect of the rest of the parameters. This procedure allows us to determine the independent effects of each parameter, even when the parameters are correlated. The sign of the PRCC indicates the qualitative relationship between each input variable and each output variable. A positive value indicates that when the value of the input variable increases, the value of the output will also increase. A negative value indicates a negative correlation between the inputs and the output. The magnitude of PRCC measures the contribution of the input variables to the output variable. PRCC are determined for each input variable and the outcome variable as follows [4]:

- First, the outcome vector is added as an additional column in column number

$k+1$ to the matrix of input values (parameters of the model), where k is the number of input variables. The ordinal numbers representing the rank (1 to N) of each of these columns are defined as the set $(r_{1i}, r_{2i}, ..., r_{ki}, R_i)$, where i = run number. The average rank $\mu = (1+N)/2$. A $(k+1) \times (k+1)$ symmetric matrix (C) may be defined, with elements c_{ij},

$$c_{ij} = \frac{\sum_{t=1}^{N}(r_{it}-\mu)(r_{jt}-\mu)}{\sqrt{\sum_{t=1}^{N}(r_{it}-\mu)^2 \sum_{s=1}^{N}(r_{js}-\mu)^2}} \qquad i,j = 1,2,...,k. \tag{4.6}$$

- For the $c_{i,k+1}$ elements R_i replaces r_{jt} and r_{js}. The leading diagonal elements of C are all ones. The matrix B is defined as the inverse of C.

$$B = [b_{ij}] = C^{-1}.$$

- The PRCC (ρ_i) between the ith input parameter and the outcome variable is defined as

$$\rho_i = \frac{-b_{i,k+1}}{\sqrt{b_{i,i}b_{k+1,k+1}}}. \tag{4.7}$$

- The significance of a nonzero value of ρ_i is tested by computing t_i. The distribution of this variable approximates a Student's T of $N-2$ degrees of freedom:

$$t_i = \rho_i\sqrt{\frac{N-2}{\rho_i}} \tag{4.8}$$

The PRCC method assumes a monotonic relationship between the input parameters and the output. Thus, we assess this assumption of monotonicity by examining scatter plots, where each input variable is plotted against the outcome variable, *proportion of habitual consumers,* C_b. Figure 4.3 shows this monotonic relationship between the input variables and the *proportion of habitual consumers,* C_b in 2013. We can observe a strong correlation between the parameters γ and σ to C_b, some weak correlation between β and C_b and little or no correlation between C_b and the other parameters. For the remaining years of the study, we obtained similar results.

Then, the PRCC was calculated between each input parameter and C_b, *proportion of habitual consumers.* It can be found in Table 4.6 that uncertainties estimating the values of *γ, β, d_{C_2}, d_{C_3} and σ* are the most critical in the prediction accuracy of the future proportion of habitual consumers. These values are, respectively, in 2015, 0.74684, 0.56642, -0.510844, -0.49738 and 0.28924, which are higher than the others. Thus, we can focus our efforts in these variables. The sign of the PRCC identifies the qualitative relationship between the inputs and the output variable. The positive value of the PRCC for most of the variables implies that when the input variable increases, the future proportion of habitual consumers will also increase. However, the future number of habitual consumers decreases as the variables d_{C_2} and d_{C_3} increase.

Table 4.6: PRCC between each input variable and the output variable, for years 2011, 2012, 2013, 2014 and 2015 (with all the parameters).

Parameter	2011	2012	2013	2014	2015
μ	0.16995	0.17004	0.16986	0.16965	0.16962
p-value	1.03×10^{-33}	9.51×10^{-34}	1.11×10^{-33}	1.33×10^{-33}	1.37×10^{-33}
d_N	0.18188	0.18230	0.18293	0.18329	0.18377
p-value	1.90×10^{-38}	1.28×10^{-38}	7.05×10^{-39}	5.01×10^{-39}	3.16×10^{-39}
d_{c1}	-0.08829	-0.09670	-0.10493	-0.11257	-0.11934
p-value	4.00×10^{-10}	6.28×10^{-12}	1.02×10^{-13}	1.42×10^{-15}	2.52×10^{-17}
d_{c2}	-0.50433	-0.50551	-0.50659	-0.50778	-0.510844
p-value	0.00000	0.00000	0.00000	0.00000	0.00000
d_{c3}	-0.51024	-0.50695	-0.50352	-0.50023	-0.49738
p-value	0.00000	0.00000	0.00000	0.00000	0.00000
β	0.47940	0.50464	0.52721	0.547863	0.56642
p-value	9.11305×10^{-286}	0.00000	0.00000	0.00000	0.00000
γ	0.78726	0.77753	0.76742	0.75728	0.74684
p-value	0.00000	0.00000	0.00000	0.00000	0.00000
σ	0.36560	0.34430	0.32444	0.30585	0.28924
p-value	5.80×10^{-158}	3.78×10^{-139}	6.76×10^{-123}	9.41×10^{-109}	5.86×10^{-97}
ε	0.03104	0.03347	0.03587	0.03787	0.04024
p-value	2.82×10^{-2}	1.79×10^{-2}	1.12×10^{-2}	7.41×10^{-3}	4.43×10^{-3}

As we mentioned in the previous chapter, we cannot change the value of natality and mortality rates. Therefore, we repeat the PRCC analysis excluding these parameters and, in this second analysis, we consider only four parameters: β, γ, σ and ε. Results can be seen in Table 4.7.

Table 4.7: PRCC between each input variable and the output variable, for years 2011, 2012, 2013, 2014 and 2015 (without natality and mortality parameters).

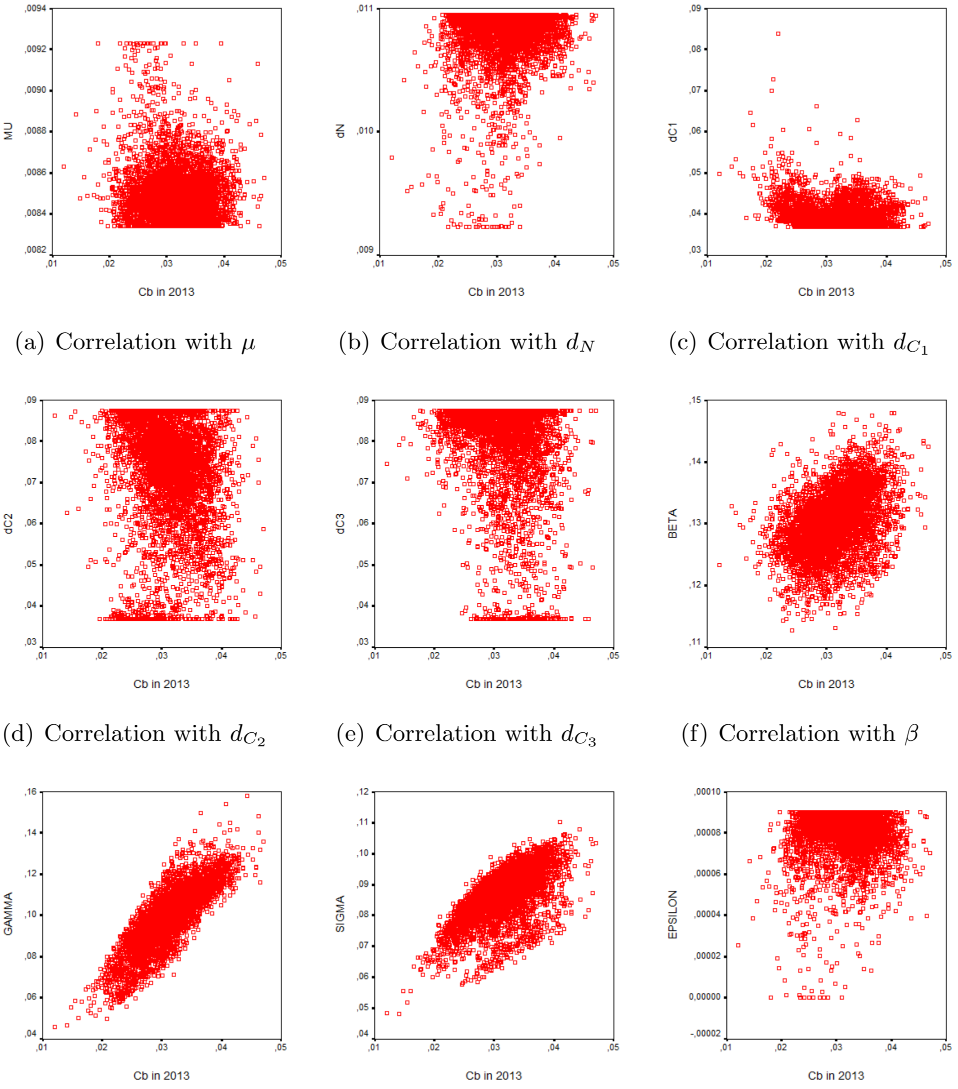
Parameter	2011	2012	2013	2014	2015
β	0.37345	0.39794	0.42038	0.44116	0.46020
p-value	2.94×10^{-165}	2.20×10^{-189}	2.44×10^{-213}	3.15×10^{-237}	1.27×10^{-260}
γ	0.74870	0.73800	0.72689	0.71571	0.70435
p-value	0.00000	0.00000	0.00000	0.00000	0.00000
σ	0.37846	0.36080	0.34476	0.33000	0.31648
p-value	5.05×10^{-170}	1.35×10^{-153}	1.52×10^{-139}	2.52×10^{-127}	1.03×10^{-116}
ε	-0.01291	-0.00860	-0.00445	-0.00072	0.00314
p-value	3.61×10^{-1}	5.43×10^{-1}	7.53×10^{-1}	9.59×10^{-1}	8.24×10^{-1}

It shows that the parameters γ, β and σ (with values of 0.70435, 0.46020 and 0.31648, respectively, in 2015) are the most critical in the prediction precision of the future proportion of habitual consumers. Thus, we can focus our efforts on these variables. Note that these parameters are related to an increasing consumption. Therefore we should decrease the rate of consumption in order to decrease cocaine consumption over the next few years. In other words, we should focus on prevention programmes. Note that this conclusion agrees with those obtained in the previous chapters, although the techniques used are different.

4.4 Conclusions

In this chapter, we have used the bootstrap method for confidence interval estimation and prediction to assess uncertainty about the model estimations. Using this method, the model predicts that 78.15% ([75.42%, 80.82%]), 12.38% ([10.43%, 14.42%]), 5.87% ([4.85%, 6.86%]) and 3.61% ([2.54%, 4.74%]) of 15-64-year-old individuals in Spain will be, by year 2015, non-consumer, occasional consumer, regular consumer and habitual consumer, respectively. If we compare these results with the ones obtained in the previous chapters, we can conclude that the intervals obtained with the bootstrap method are smaller than the intervals obtained with LHS methods, but higher than the ABC confidence intervals.

In addition, we have identified, using the PRCC, the parameters which most influence the prevalence of cocaine consumption in Spain. Note that these parameters (β, γ and σ) are related to cocaine consumption contagion or transitions which involve an increase in consumption. The conclusions obtained are similar to the ones reached in the previous chapters, i.e., in order to reduce cocaine consumption, prevention is the best strategy.



(a) Correlation with μ

(b) Correlation with d_N

(c) Correlation with d_{C_1}

(d) Correlation with d_{C_2}

(e) Correlation with d_{C_3}

(f) Correlation with β

(g) Correlation with γ

(h) Correlation with σ

(i) Correlation with ε

Figure 4.3: Scatter plots comparing the total number of habitual consumers in 2013 (C_b) against each parameter. (g) and (h) show a strong correlation, (f) shows some weak correlation and the other graphs show little or no correlation (see Table 4.6 for correlation coefficients).

Chapter 5

Conclusions

Due to the importance of the use of cocaine in developed countries and, in particular, in Spain, and considering this consumption as an epidemic that is transmitted socially, in this Ph.D. Thesis we have defined a mathematical model to study the evolution of cocaine consumption in Spain and predict that consumption in the coming years.

In order to ensure that the defined model is the one that best fits the data, we have considered other possible models (built taking into account the suggestion of a clinical psychologist), and we found that, indeed, the first defined model is the one that best reflects the data.

To the best of our knowledge, this is the only model defined in Spain considering the populations non consumers, occasional consumers, regular consumers and habitual consumers, and using actual data provided by the surveys of the Drug National Observatory Reports, PNSD.

Furthermore, given the uncertainty due to human behaviour, to the errors in the data, rounding errors, etc., it is necessary to consider uncertainty in the definition of the model. Thus, we are not satisfied only in the results the deterministic model provides, and we wanted to go further, considering uncertainty in the model parameters. In this way, we will be able to predict the use of cocaine in Spain over the next years with 95% confidence intervals. To get this and in order to improve and control the uncertainty in the estimations, we used three different techniques to estimate 95% confidence intervals and, consequently, quantify the uncertainty in the predictions: LHS (Latin Hipercube Sampling), ABC (Approximate Bayesian Computation) and Bootstrap. The predictions obtained with each one of these techniques for the year 2015 can be found in the Table 5.1:

	LHS	ABC	Bootstrap
Non-consumers	[0.394, 0.957]	[0.786, 0.807]	[0.754, 0.808]
Occasional consumers	[0.008, 0.471]	[0.121, 0.139]	[0.104, 0.144]
Regular consumers	[0.006, 0.200]	[0.041, 0.055]	[0.049, 0.069]
Habitual consumers	[0.005, 0.103]	[0.020, 0.031]	[0.025, 0.047]

Table 5.1: Model predictions with uncertainty for year 2015 of percentage of non-consumers, occasional consumers, regular consumers and habitual consumers using each one of the techniques: LHS, ABC and Bootstrap.

Studying the predictions with uncertainty provided for the model, we note that the model predicts well in the short term with independence of the used technique, although the intervals obtained with the LHS method are higher than the other ones obtained with the ABC and bootstrap method. However, as time goes on, the predictions are not as good as expected. This is because there are two issues we did not include during the modeling: the Plan of Action Against Cocaine Consumption proposed in 2007 by the Spanish Health Ministry, and the economic crisis that emerged in 2008. Both facts have meant a considerable and unexpected decline in the trend of consumption.

In addition, in this Ph.D. Thesis, we conducted a model sensitivity analysis to determine which parameters are those that most influence cocaine consumption in Spain. The result of this sensitivity analysis allows us to design public health strategies and to analyse their effects on reducing cocaine consumption in the future. We have used several techniques to carry out this sensitivity analysis and the conclusion obtained in all cases is the same: prevention policies are the most effective strategy in reducing cocaine consumption.

Finally, we want to summarize the main objectives we achieved in the present Ph.D. dissertation:

- To define the first mathematical model to study the evolution of cocaine consumption in Spain considering the populations non-consumers, occasional consumers, regular consumers and habitual consumers, and with real data.
- To use techniques that allowed us to choose the model that best explains reality, from among three proposed models taking into account the expert opinion of a clinical psychologist.

- To quantify the uncertainty in the model using different techniques (LHS, ABC and Bootstrap) and be able to predict with accuracy cocaine consumption in Spain with 95% confidence intervals.

- To carry out a model sensitivity analysis to determine which model parameters are those that most influence the cocaine consumption in Spain and then, to conclude that the most efficient way of reducing this consumption is through prevention policies. Consequently, it is worth investing efforts in such policies.

Bibliography

[1] H. Bae, R. Grandhi, and R. A. Canfield. An approximation approach for uncertainty quantification using evidence theory. *Reliability Engineering and System safety*, 86 (3):215–225, 2004.

[2] N.A. Battista. *Heroin Epidemic Models.* Technical report, Department of Applied Mathematics & Statistics. Stony brook university, New York, United States, 2009.

[3] D.A. Behrens, J.P. Caulkins, G. Tragler, J. L. Haunschmied, and G. Feichtinger. A dynamic model of drug initiation: implications for treatment and drug control. *Mathematical Biosciences*, 159:1–20, 1999.

[4] S.M. Blower and H. Dowlatabadi. Sensitivity and uncertainty analysis of complex models of a disease transmission: and HIV model, as an example. *International Statistical Review*, 62:229–243, 1994.

[5] D. Bose and M. Wright. *Uncertainty analysis of laminar aeroheating predictions for mars entries.* In 38th AIAA Thermophysics Conference (ed. V. Alexandrov et al.), pp.652-662. AIAA 2005-4682, 2006.

[6] G.E.P. Box and D.A. Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65:1509–1526, 1970.

[7] F. Brauer, P. Driessche, and J. Wui. *Mathematical Epidemiology.* Springer, Canada, 2008.

[8] A. Budney, S. Higgins, D. Mercer, and G. Carpenter. *Therapy Manuals for Drug Abuse. Manual 2. A Community Reinforcement Plus Vouchers Approach.* Treating Cocaine Addiction. National Institute on Drug Abuse. EE.UU, 1998.

[9] Spanish Official Bulletin. Retrieved from http://www.boe.es, 2009.

[10] L. Caballero. *Adicción a cocaína: neurobiología, clínica, diagnóstico y tratamiento.* Delegación del Gobierno para el Plan Nacional sobre Drogas. Madrid: Ministerio del Interior [Addiction to cocaine: neurobiology, clinic, diagnosis and treatment. National Plan on Drugs. Spain], 2005.

[11] G.C. Castillo, S.G. Jordan, and A.H. Rodriguez. *Mathematical models for the dynamics of tobacco use, recovery and relapse*, volume 3. Technical Report Series, BU-1505-M. Department of Biometrics, Cornell University, 1997.

[12] J.P. Caulkins, D.A. Behrens, C. Knoll, G. Tragler, and D. Zuba. Markov chain modeling of initiation and demand: The case of the us epidemic. *Health Care Management Science*, 7:319–329, 2004.

[13] J.P. Caulkins, G. Tragler, and D. Waqllner. Optimal timing of use reduction vs. harm reduction in a drug epidemic model. *International Journal of Drug Policy*, 20:480–487, 2009.

[14] N.A. Christakis and J.H. Fowler. *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives.* Hachette Book Group, New York, 2009.

[15] A. Cintrón-Arias, F. Sánchez, and C. Castillo-Chávez. *Sensitivity Analysis of Drinking Dynamics: From Deterministic to Stochastic Formulations.* Technical report, Center for research in scientific computation, North Carolina, State University, United States, 2008.

[16] A.C. Cullen and H.C. Frey. *Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs.* Plenum Press, New York, 1994.

[17] L. Degenhardta, J. Singletona, B. Calabriaa, J. McLarena, T. Kerrb, S. Mehtac, G. Kirkc, and W.D. Halld. Mortality among cocaine users: A systematic review of cohort studies. *Drug and Alcohol Dependence*, 113:88–95, 2011.

[18] G. Dogan. Bootstrapping for confidence interval estimation and hypothesis testing for parameters of system dynamics models. *System Dynamics Review*, 23:415–436, 2007.

[19] D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society*, Series B 57 (1):45–97, 1995.

[20] L. Dutra, G. Stathopoulou, S.L. Basdem, T.M. Leyro, M.B. Powers, and M.W. Otto. A meta-analytic review of psychosocial interventions for substance use disorders. *American Journal of Psychiatry*, 165:179–187, 2008.

[21] B. Efron. 1977 rietz lecture: bootstrap methods - another look at the jackknife. *Annals os Statistics*, 7(1):1–26, 1979.

[22] S. L. R. Ellison, M. Rosslein, and A. Williams. *Quantifying uncertainty inanalytical measurement.* CEPIS, 2000.

[23] EMCDDA. *European Monitoring Centre for drugs and drug addiction: Drug dependent problem in Europe.* Lisbon: EMCDDA, 2007.

[24] S.S. Everingham and C.P. Rydell. *Modelling the demand for cocaine.* RAND, Santa Monica, United States, 1994.

[25] D.M. Gorman, J. Mezic, and P.J. Gruenewald. Agent- based modeling of drinking behavior: A preliminary model and potential applications to theory and practice. *Research and practice*, 96:2055–2060, 2006.

[26] L. Green, H. Lin, and M. Khalessi. *Probabilistic methods for uncertainty propagation applied to aircraft design.* In 20th AIAA Applied Aerodynamics Conference, 2002.

[27] F. Guerrero, F.J. Santonja, and R.J. Villanueva. Analysing the spanish smoke-free legislation of 2006: A new method to quantify its impact using a dynamic model. *International Journal of Drug Policy*, 22:247–251, 2011.

[28] D.N. Gujarati. *Basic Econometrics (4th edition).* Mc Graw-Hill, 2003.

[29] R. Gutenkunst, J. Waterfall, F. Casey, K. Brow, C. Myers, and J. Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput.Biol. In press, doi:10.1371/journal.pcbi.0030189*, 3:e189, 2007.

[30] J. C. Helton and W. L. Oberkampf. Alternative representations of epistemic uncertainty. *Reliability Engineering and System safety*, 85 (1):1–10, 2004.

[31] H. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42:599–653, 2000.

[32] A. Hoare, D.G. Regan, and D.P. Wilson. Sampling and sensitivity analyses tools (saSAT) for computational modelling. *Theoretical Biology and Medical Modelling*, 5:1–18, 2008.

[33] IMM. http://scaling.imm.upv.es.

[34] MathWorks; Inc. Retrieved from http://www.mathworks.es/products/matlab/index.html.

[35] INE. *Spanish Statistic Institute*. Retrieved from http://www.ine.es, 2008.

[36] H.W. Hethcote J. Mena-Lorca. Dynamic models of infectious diseases as regulators of population sizes. *Journal of Mathematical Biology*, 30:693–716, 1992.

[37] L. Jódar, F.J. Santonja, and G. González. Modeling dynamics of infant obesity in the region of Valencia, Spain. *Computers and Mathematics with Aplications*, 56:679–689, 2008.

[38] B. Johnson, J. Roach, N. Ait-Daoud, M. Javors, J. Harrison, and A. Elkashef. Preliminary, randomized, double-blind, placebo-controlled study of the safety and efficacy of ondansetron in the treatment of cocaine dependence. *Drug and Alcohol Dependence*, 84:256–263, 2006.

[39] C.Y. Kaya. Time- optimal switching control for the us cocaine epidemic. *Socio-economic planning sciences*, 38:57–72, 2004.

[40] W.O. Kermack and A.G. McKendrick. Contributions to the mathematical theory of epidemics, part 1. *Proceeding Royal Society London*, A115:700–721, 1927.

[41] A. Kolmogorov. Sulla determinazione empirica de una legge in distribuzione. *Giornate Inst. Ital. Attuari*, 4:83–91, 1933.

[42] F. Levin, S. Evans, D. Brooks, and F. Garawi. Treatment of cocaine dependent treatment seekers with adult adhd: double-blind comparison of methylphenidate and placebo. *Drug and Alcohol Dependence*, 87:20–29, 2007.

[43] J. Liu and T. Zhang. Global behaviour of a heroin epidemic model with distributed delays. *Applied Mathematics Letters*, 24:1685–1692, 2011.

[44] G.M. Ljung and G.E.P. Box. On a measure of a lack of fit in time series models. *Biometrika*, 66:265–270, 1979.

[45] R. J. Beckman M. D. Mckay and W.J.i W.J. Conover. Comparison of 3 methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21 (2):239–245, 1994.

[46] Z. Ma and J. Li. *Dynamical Modeling and Analysis of Epidemics.* World Scientific, Singapore, 2009.

[47] K.V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57:519–530, 1970.

[48] K.V. Mardia. *Tests of univariate and multivariate normality.* In: P.R. Krishnaiah (ed.), Handbook of Statistics, vol 1, pp. 279-320, Amsterdam: North Holland, 1980.

[49] S. Marino, I.B. Hogue, C.J. Ray, and D.E. Kirschner. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of Theoretical Biology*, 254:178–196, 2008.

[50] M. Martcheva and C. Castillo-Chavez. Diseases with chronic stage in a population with varying size. *Mathematical Biosciences*, 182(1):1–25, 2003.

[51] Mathematica. Technical and scientific software. Available at htpp://www.wolfram.com/products/mathematica.

[52] D. Mercer and G.Woody. *Therapy Manuals for Drug Abuse. Manual 3. An Individual Drug Counselling Approach to treat Cocaine Addiction.* The Collaborative Cocaine Treatment Study. National Institute on Drug Abuse. EE.UU, 1998.

[53] Spanish Health Ministry. *Spanish drugs survey 2007-2008.* Retrieved from http://www.msc.es/gabinetePrensa/notaPrensa/pdf/EncuestaDomiciliariaDrogas Alcohol(EDADES)MINISTRO.pdf, 2008.

[54] Spanish Health Ministry. *Spanish drugs survey 2009.* Retrieved from http://www.pnsd.msc.es/Categoria2/observa/estudios/home.htm, 2009.

[55] M.G. Morgan and M. Henrion. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis.* 1st Ed., Cambridge University Press, Cambridge, UK, 1990.

[56] Z.P. Mourelatos and J. Zhou. Uncertainty quantification using evidence theory in multidisciplinary design optimization. *Reliability Engineering and System safety*, 85 (1):281–294, 2004.

[57] A. Mubayi, P. E. Greenwood, C. Castillo-Chávez, P.J. Gruenewald, and D.M. Gorman. The impact of relative residence times on the distribution of heavy drinkers in highly distinct environments. *Socio-Economic Planning Sciences*, 44:45–56, 2010.

[58] G. Mulone and B. Straughan. A note on heroin epidemics. *Mathematical Biosciences*, 218:138–141, 2009.

[59] J.D. Murray. *Mathematical Biology.* Springer, New York, 2002.

[60] J.A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1964.

[61] NPD. *National drug strategy 2000-2008.* National Plan on Drugs. Spain: Health Ministry, 2000.

[62] NPD. *Cocaine: Clinic Comission Report.* National Plan on Drugs. Spain: Health Ministry, 2007.

[63] F. Nyabadza and S.D. Hove-Musekwa. From heroin epidemics to methamphetamine epidemics: Modelling substance abuse in a south african province. *Mathematical Biosciences*, 225:132–140, 2010.

[64] A. Olsson, G. Sandberg, and O. Dahlblom. On latin hypercube sampling for structural reliability analysis. *Structural Safety*, 25:47–68, 2003.

[65] M. Peco, F.J. Santonja, A.C. Tarazona, R.J. Villanueva, and J. Villanueva-Oller. The effect of the spanish law of political parties (lpp) on the attitude of the basque country population towards eta: A dynamic modelling approach. *Mathematical and Computer Modelling. In press, doi:10.1016/j.mcm.2011.11.007*, 2011.

[66] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W. Vetterling. *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, Cambridge, 1986.

[67] C. J. Roy and W. L. Oberkampf. *A Complete Framework for Verification, Validation, and Uncertainty Quantification in Scientific Computing (Invited).* In 48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition, Orlando, Florida, AIAA 2010-124, 2010.

[68] G.P. Samanta. Dynamic behaviour for a nonautonomous heroin epidemic model with time delay. *Journal of Applied Mathematics and Computing*, 35:161–178, 2011.

[69] E. Sánchez, R.J. Villanueva, F.J. Santonja, and M. Rubio. Predicting cocaine consumption in spain: a mathematical modelling approach. *Drugs: Education, Prevention and Policy*, 18(2):108–115, 2011.

[70] F. Sánchez, X. Wang, C. Castillo-Chávez, D. M. Gorman, and P. J. Gruenewald. *Drinking as an Epidemic- a simple mathematical model with recovery and relapse, in: K. Witkiewitz, G. Alan Marlett (Eds.).* Therapists Guide to Evidence Based Relapse Prevention, Princeton and Oxford, 2007.

[71] F.J. Santonja, A.C. Tarazona, and R.J. Villanueva. A mathematical model of the pressure of an extreme ideology on a society. *Computers and Mathematics with Aplications*, 56:836–846, 2008.

[72] F.J. Santonja, R.J. Villanueva, L. Jódar, and G. González. Mathematical modeling of social obesity epidemic in the region of Valencia, Spain. *Mathematical and Computer Modelling of Dynamical Systems*, 16:23–34, 2010.

[73] J. Schmitz, A. Stotts, H. Rhoades, and J. Grabowski. Naltrexone and relapse prevention treatment for cocaine-dependent patients. *Addictive Behaviors*, 26:167–180, 2001.

[74] S.S Shapiro and M.B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1965.

[75] O. Sharomi and A.B. Gumel. Curtailing smoking dynamics: A mathematical modeling approach. *Applied Mathematics and Computation*, 195:475–499, 2008.

[76] N. V. Smirnov. Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin of Moscou University*, 2:3–16, 1939.

[77] B. Song, M. Castillo-Garsow, K.R. Ríos-Soto, M. Mejran, L. Henso, and C. Castillo-Chávez. Raves, clubs and ecstasy: The impact of peer pressure. *Mathematical Biosciences and Engineering*, 3:249–266, 2006.

[78] A. Stotts, M. Mooney, S. Sayre, M. Novy, M. Schmitz, and J. Grabowsk. Illusory predictors: generalizability of findings in cocaine treatment retention research. *Addictive Behaviors*, 32:2819–2836, 2007.

[79] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M.P.H. Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6:187–202, 2009.

[80] UNODC. *World drug report.* Vienna: United Nations Office on Drugs and Crime, 2008.

[81] D. Vose. *Risk Analysis: A Quantitative Guide.* 3rd Ed., Wiley, New York, 2008.

[82] E. White and C. Comiskey. Heroin epidemics, treatment and ode modelling. *Mathematical Biosciences*, 208:312–324, 2007.

[83] S. F. Wojtkiewicz, M. S. Eldred, R. V. Field, A. Urbina, and J. R. Red-Horse. *Uncertainty quantification in large computational engineering models.* In 42nd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials (SDM) Conference, 2001.

[84] H. Zimmermann. *Fuzzy set theory -and its applications.* 3rd edn. Kluwer Academic Publishers, 1996.