



Universitat Politècnica de València  
Departament de Matemàtica Aplicada

PhD. THESIS

**Modelling the evolution dynamics  
of the academic performance  
in high school in Spain.  
Probabilistic predictions of future trends  
and their economical consequences.**

Ph.D. CANDIDATE

Almudena Sánchez Sánchez

ADVISORS

Dr. Juan Carlos Cortés López  
Dr. Francisco José Santonja Gómez  
Dr. Rafael Jacinto Villanueva Micó

Valencia - June 2013



# Declaration of Authorship

Dr. Juan Carlos Cortés López and Rafael Jacinto Villanueva Micó, professors at the Universitat Politècnica de València and Francisco José Santonja Gómez professor at the Universitat de València,

CERTIFY that the present thesis entitled: *Modelling the evolution dynamics of the academic performance in high school. Probabilistic predictions of future trends and their economical consequences* has been performed under our supervision in the Department of Applied Mathematics at the Universitat Politècnica de València by Almudena Sánchez Sánchez. It constitutes her thesis dissertation to obtain the PhD degree in Mathematics.

In compliance with the current legislation, we authorize the presentation of this dissertation signing the present certificate.

**Valencia - July 2013**

**Dr. Juan Carlos  
Cortés López**

**Dr. Francisco José  
Santonja Gómez**

**Dr. Rafael Jacinto  
Villanueva Micó**



*“With good education is human a docile and divine creature but without it, they are the fiercest of animals.*

*Education and teaching improve the good ones and make good to the bad ones.”*

Plato. Dialogues. Volume IV.

(República, trad. de Eggers Lan). Gredos.



# Abstract (English)

In this dissertation, we use epidemiologic-mathematical techniques to model the academic performance in Spain (paying special attention on the academic underachievement) to understand better the mechanisms behind this important issue as well as to predict how academic results will evolve in the Spanish *Bachillerato* over the next few years. The Spanish *Bachillerato* educational level is made up of the last courses before accessing to the university or to the work market and corresponds to students of 16 – 18 years old. This educational level is a milestone in the career training of students because it represents a period to make important decisions about academic and professional future.

In a first step, in the Chapter 2 we will present a deterministic model where academic performance is analyzed assuming the negative attitude of *Bachillerato* students may be due to their autonomous behavior and the influence of classmates with bad academic results. Then, in the Chapter 3, the model is improved based on the idea that not only the bad academic habits are socially transmitted but also the good study habits. Besides, we decompose the transmission academic habits into good and bad academic habits, in order to analyze with more detail which group of students are more susceptible to be influenced by good or bad academic students. The consideration of quantifying the abandon rates is also a new issue dealt with in it. The adopted approach allow to provide both punctual and confidence intervals predictions to the evolution of academic performance (including the abandon rates) in *Bachillerato* in Spain over the next few years. The adopted approach allows us to model academic performance in academic levels other than *Bachillerato* and/or beyond the Spanish academic system. This issue is assessed in Chapter 4, where the model is satisfactorily applied to the current academic system of the German region of North Rhine-Westphalia.

To conclude this dissertation, we provide an estimation of the cost related to the Spanish academic underachievement based on our predictions. This estimation represents the investment in the Spanish *Bachillerato* from the Spanish Government and families over the next few years, paying special attention on the groups of students who do not promote and abandon during their corresponding academic year.





# Abstract (Spanish)

En esta tesis, se utilizan técnicas matemático-epidemiológicas para modelar el rendimiento académico en España (prestando especial atención en el fracaso escolar) para comprender mejor los mecanismos detrás de esta importante cuestión, así como para predecir cómo evolucionarán los resultados académicos en el *Bachillerato* español en los próximos años. El nivel educativo de *Bachillerato* en España está formado por los dos últimos cursos antes de acceder a la universidad o al mercado de trabajo y corresponde a los estudiantes de 16 – 18 años. Este nivel educativo es muy importante para la formación de los estudiantes ya que representa un periodo en el que deberán tomar importantes decisiones sobre el futuro académico y profesional.

En primer lugar, en el Capítulo 2, se presenta un modelo determinista donde se analiza el rendimiento académico asumiendo que la actitud negativa de los alumnos de *Bachillerato* puede ser debida a su comportamiento autónomo y la influencia de compañeros con malos resultados académicos. Luego, en el Capítulo 3, se mejora el modelo basado en la idea de que no sólo los malos hábitos académicos se transmiten socialmente sino también los buenos hábitos de estudio. Además, descomponemos los parámetros de transmisión de hábitos académicos con el fin de analizar con más detalle qué grupos de estudiantes son más susceptibles a ser influenciados por compañeros con buenos o malos hábitos académicos. El abandono escolar también han sido incluido en este modelo. El enfoque adoptado permite proporcionar predicciones deterministas y con intervalos de confianza de la evolución del rendimiento escolar (incluyendo las tasas de abandono) en *Bachillerato* en España en los próximos años. Este enfoque, además, nos permite modelar el rendimiento académico en otros niveles educativos del sistema académico español o de fuera de España tal y como se muestra en el Capítulo 4, donde el modelo se aplica satisfactoriamente al sistema académico actual de la región alemana de Renania del Norte-Westfalia.

Para concluir esta tesis, proporcionamos una estimación de los costes relacionados con el rendimiento académico español basado en nuestras predicciones. Esta estimación representa la inversión en *Bachillerato* por parte del Gobierno español y las familias en los próximos años, con especial atención en los grupos de estudiantes que no promocionan y abandonan en los diferentes cursos académicos.



# Abstract (Valencià)

En aquesta tesi, s'utilitzen tècniques matemàtic-epidemiològiques per a modelitzar el rendiment acadèmic a Espanya (parant especial atenció en el fracàs escolar) per a comprendre millor els mecanismes darrere d'aquesta qüestió tan important, així com per a predir com evolucionaran els resultats acadèmics en el Batxillerat espanyol en els pròxims anys. El nivell educatiu de Batxillerat a Espanya està format pels dos últims cursos abans d'accedir a la universitat o al mercat de treball i correspon als estudiants de 16 a 18 anys. Aquest nivell educatiu és molt important per a la formació dels estudiants ja que representa un període en què hauran de prendre decisions importants sobre el futur acadèmic i professional.

En primer lloc, en el Capítol 2, es presenta un model determinista on s'analitza el rendiment acadèmic assumint que l'actitud negativa dels alumnes de Batxillerat pot ser deguda al seu comportament autònom i la influència dels companys amb resultats acadèmics dolents. Després, en el Capítol 3, es millora el model basat en la idea que no només els mals hàbits acadèmics es transmeten socialment sinó també els bons hàbits d'estudi. A més, descomposem els paràmetres de transmissió d'hàbits acadèmics a fi d'analitzar amb més detall quins grups d'estudiants són més susceptibles de ser influenciats pels companys amb hàbits acadèmics bons o dolents. L'abandonament escolar també han sigut inclòs en aquest model. L'estudi des de aquest punt de vista, a més a més, ens permet modelitzar el rendiment acadèmic en altres nivells educatius del sistema acadèmic espanyol o de fora d'Espanya tal com es mostra en el Capítol 4, on el model s'aplica satisfactòriament al sistema acadèmic actual de la regió alemanya de Renània del Norte-Westfalia.

Aquesta tesi conclou proporcionant una estimació dels costos relacionats amb el rendiment acadèmic espanyol en base a les nostres prediccions. Aquesta estimació representa la inversió en Batxillerat per part del Govern espanyol i les famílies en els pròxims anys, parant especial atenció en els grups d'estudiants que no promocionen i abandonen en els diferents cursos acadèmics.



# Contributions

Some of the chapters presented in this dissertation have been previously published or are in process for publication.

- Chapter 2 appears in its entirety as:

J. Camacho and J.C. Cortés and R.M. Micle and A. Sánchez-Sánchez. Predicting the academic underachievement in a high school in Spain over the next few years: A dynamic modeling approach. *Mathematical and Computer Modelling*, 7-8(57):1703–1708, 2013. DOI:10.1016/j.mcm.2011.11.011.

- Chapter 3 appears in its entirety as:

J.C. Cortés and A. Sánchez-Sánchez and F.J. Santonja and R.J. Villanueva. Non-parametric probabilistic forecasting of academic performance in Spanish high school using an epidemiological modelling approach. *Applied Mathematics and Computation*, Jun 2013. DOI:10.1016/j.amc.2013.06.070.

- Chapter 4 appears in its entirety as:

J.C. Cortés and M. Ehrhardt and A. Sánchez-Sánchez and F.J. Santonja and R.J. Villanueva. Modelling the dynamics of the students academic performance in the German region of North Rhine-Westphalia: an epidemiological approach with uncertainty. *International Journal of Computer Mathematics*. DOI:10.1080/00207160.2013.813937.

- Appendix A appears in its entirety as:

J.C. Cortés and A. Sánchez-Sánchez and F.J. Santonja and R.J. Villanueva. epiModel : A system to build automatically systems of differential equations of compartmental type-epidemiological models. *Computers in Biology and Medicine*, 41(11):999-1005, 2011. DOI:10.1016/j.combiomed.2011.06.018.



# Acknowledgements

Thanks be to God for allowing me to write this dissertation.

The completion of this thesis has been possible thanks to the support of many people to whom I will always be grateful.

To Juan Carlos Cortés, Francisco Santonja and Rafael Villanueva, my dissertation advisors, because without their time, dedication and motivation this thesis would never have existed, but especially for their friendship and because of them I really enjoyed doing this work. And I will be forever grateful for their efforts to obtain economic financing that would allow me to continue keep learning from them.

To the colleagues of the *Multidisciplinary Institute of Mathematics* at the *Universitat Politècnica de València* for their hospitality and allowing me to develop my work there day to day during the last two years. This has been a valuable source of inspiration both human and academic.

To whole *Departament de Matemàtica Aplicada*, for their kindness and all the received help whenever I needed.

To Prof. Matthias Ehrhardt, for facilitating my stay in Wuppertal (Germany) for 3 months, for their kindness and valuable suggestions.

My entire family and friends, and especially my parents Enrique and Pilar, my brother Enrique and my boyfriend Raúl for their help and encouragement needy at times, to be always at my side, and because without their love and trust would never have succeeded.

Thank you!





*To my parents, Enrique and Pilar.*

*To my brother, Enrique.*



# Contents

|  |           |
|--|-----------|
| Declaration of Authorship  | ii        |
| Abstract (English)   | vi        |
| Abstract (Spanish)   | viii      |
| Abstract (Valencià)  | x         |
| Citations Published Work   | xii       |
| List of Figures  | xxii      |
| List of Tables   | xxiv      |
| <b>1 Introduction</b>  | <b>1</b>  |
| <b>2 Predicting the academic underachievement in high school in Spain over the next few years: A dynamic modelling approach</b>              | <b>9</b>  |
| 2.1 Introduction . . . . .   | 9         |
| 2.2 Building the mathematical model . . . . .  | 10        |
| 2.2.1 Available data . . . . .   | 10        |
| 2.3 Parameter estimation . . . . .   | 14        |
| 2.4 Prediction over next few years . . . . .   | 15        |
| 2.5 Conclusions . . . . .  | 16        |
| <b>3 Non-parametric probabilistic forecasting of academic performance in Spanish high school using an epidemiological modelling approach</b> | <b>19</b> |
| 3.1 Introduction . . . . .   | 19        |
| 3.2 The epidemiological-mathematical model . . . . .   | 20        |
| 3.2.1 Available data . . . . .   | 20        |
| 3.2.2 Model building . . . . .   | 21        |
| 3.2.3 Scaling the model . . . . .  | 26        |
| 3.3 Deterministic parameter estimation and prediction over the next few years . . . . .  | 27        |

|  |  |           |
|--|--|-----------|
| 3.4  | Introducing uncertainty in the model parameters and predicting the next few years . . . . .  | 30        |
| 3.4.1  | Error term analysis . . . . .  | 31        |
| 3.4.2  | Generating new output perturbed data . . . . .   | 33        |
| 3.4.3  | Obtaining confidence intervals for model outputs . . . . .   | 33        |
| 3.5  | Abandon analysis . . . . .   | 37        |
| 3.6  | Conclusions . . . . .  | 39        |
| <b>4</b>   | <b>Modelling the dynamics of the students academic performance in the German region of North Rhine-Westphalia: an epidemiological approach with uncertainty</b>          | <b>41</b> |
| 4.1  | Introduction . . . . .   | 41        |
| 4.2  | Model building . . . . .   | 42        |
| 4.2.1  | Available data . . . . .   | 42        |
| 4.2.2  | The type-epidemiological model . . . . .   | 43        |
| 4.3  | Scaling, fitting and predictions . . . . .   | 48        |
| 4.4  | Introducing uncertainty in the model parameters and predicting the next few years . . . . .  | 50        |
| 4.4.1  | Error term analysis . . . . .  | 51        |
| 4.4.2  | Generating new output perturbed data . . . . .   | 53        |
| 4.4.3  | Obtaining confidence intervals for model outputs . . . . .   | 54        |
| 4.5  | Conclusions . . . . .  | 57        |
| <b>5</b>   | <b>Estimation of the cost of the academic underachievement in high school in Spain over the next few years</b>   | <b>59</b> |
| 5.1  | Introduction . . . . .   | 59        |
| 5.2  | Estimation with 95% confidence intervals of the cost of the academic underachievement in <i>Bachillerato</i> for the next few years for the Spanish Government . . . . . | 60        |
| 5.3  | Estimation with 95% confidence intervals of the investment in education by Spanish families of <i>Bachillerato</i> students in the next few years . . . . .              | 66        |
| 5.4  | Conclusions . . . . .  | 70        |
| <b>6</b>   | <b>Conclusion and discussion</b>   | <b>73</b> |
| <br>   |  |           |
| <b>Appendix A epiModel: A system to build automatically systems of differential equations of compartmental type-epidemiological models</b> |  | <b>77</b> |
| A.1  | Introduction . . . . .   | 77        |
| A.2  | How to build the file "ModelDefinition" . . . . .  | 80        |
| A.2.1  | General variable . . . . .   | 80        |
| A.2.2  | Definition of the subpopulations . . . . .   | 80        |
| A.2.3  | Defining Parameters . . . . .  | 81        |

---

|  |  |            |
|--|--|------------|
| A.2.3.1  | Parameters of independent term and linear term . . . . .         | 82         |
| A.2.3.2  | Parameters of non-linear terms . . . . .                         | 84         |
| A.3  | Steps to building the system of differential equations . . . . . | 85         |
| A.3.1  | The file "Model.data" . . . . .                                  | 86         |
| A.3.2  | The file "parameters.data" . . . . .                             | 87         |
| A.4  | Examples . . . . .   | 87         |
| A.4.1  | SIRS model . . . . .   | 88         |
| A.4.2  | SIR model with two age groups . . . . .                          | 90         |
| A.4.3  | SIR model with two age groups and two sexes . . . . .            | 93         |
| A.5  | Conclusions . . . . .  | 96         |
| <b>Appendix B Validation of our Spanish mathematical model results</b>   |  | <b>99</b>  |
| <b>Appendix C Validation of our German mathematical model results</b>    |  | <b>103</b> |
| <b>Appendix D Time series analysis: Forecasting models in Statgraph-</b> |  |            |
| <b>ics Plus 5.1.</b>   |  | <b>105</b> |
| D.1  | Introduction . . . . .   | 105        |
| D.2  | Forecasting models . . . . .                                     | 106        |
| D.3  | Validation of the model . . . . .                                | 108        |
| D.4  | Obtaining 95% confidence intervals . . . . .                     | 109        |
| <b>Bibliography</b>  |  | <b>111</b> |



# List of Figures

|     |   |    |
|-----|---|----|
| 1.1 | Structure of the Spanish educational system for students aged between 12 – 18 years old. . . . .  | 4  |
| 2.1 | Flow diagram of the mathematical model for <i>Bachillerato</i> academic underachievement in Spain. The boxes represent subpopulations under study classified according to genre (women (W) and men (M)) and academic level (First and Second Stage of <i>Bachillerato</i> ). Students that belong to the promote/non-promote group are denoted by W, M, $\bar{W}$ and $\bar{M}$ , respectively. The arrows represent the transitions between the subpopulations, and they are labeled by their corresponding terms and parameters according to the model. . . . .   | 13 |
| 2.2 | Fitting and prediction of the academic performance of <i>Bachillerato</i> Spanish students over the academic years 1999 – 2000 to 2014 – 2015. . . . .  | 16 |
| 3.1 | Flow diagram of the epidemiological-mathematical model for dynamics of <i>Bachillerato</i> academic performance in Spain. The boxes represent the students depending on their sex, stage and academic results. The arrows denote the transits of students labeled by the expressions and parameters governing these transits. . . . .   | 26 |
| 3.2 | Real data (red points on the left side of vertical axis) and prediction (line) with confidence intervals (on the right side of vertical axis) of the academic performance of <i>Bachillerato</i> Spanish students over the academic years 1999 – 2000 to 2014 – 2015. Smaller confidence intervals, represent less uncertainty in the predictions, the points in the middle of the confidence intervals are their means. The square black point represents the last academic results published recently corresponding to the academic year 2009 – 2010. Notice that each graph has its own scale. . . . . | 36 |
| 4.1 | Flow diagram of the model (4.3)-(4.5). The boxes represent the students depending on their gender, level and academic results. The arrows denote the transit of students labelled by the cause of the flow. . . . .   | 48 |

|     |  |    |
|-----|--|----|
| 4.2 | Real data (black points) and prediction with 95% confidence intervals (red line) of the academic performance of German students in the North Rhine-Westphalia over the academic years 2006 – 2007 to 2014 – 2015. Smaller confidence intervals, represent less uncertainty in the predictions, the dashed lines in the middle of the confidence intervals are their means. Note that there are high differences in the scale of the graphs between the promotable and non-promotable students, specially with very low rates in the non-promotable groups. | 57 |
| 5.1 | Graph of the prediction of euros invested by the Spanish Government in each Spanish student in the First and Second Stage of <i>Bachillerato</i> , in both, state and private high schools during the academic years from 2009 – 2010 to 2014 – 2015. . . . .  | 63 |
| 5.2 | Graph of the prediction (in euros) the Spanish families will invest in each <i>Bachillerato</i> student during the academic years from 2009 – 2010 to 2014 – 2015. . . . .   | 69 |
| A.1 | Process of how <i>epiModel</i> works. "ModelBuilder.nb" loads data from "ModelDefinition" and "epiModel" creates "Model.data" and "parameters.data". . . . .   | 79 |
| A.2 | Parameter types dependent on where the arrows enter and exit in compartmental models. . . . .  | 82 |
| A.3 | Screenshot of "ModelBuilder.nb" in <i>Mathematica</i> . . . . .  | 86 |
| A.4 | Diagram of a Susceptible-Infectious-Recovered-Susceptible model. . . . .   | 88 |
| A.5 | Diagram of a Susceptible-Infectious-Recovered model with two age groups. . . . .   | 90 |



# List of Tables

|     |  |    |
|-----|--|----|
| 2.1 | The available data corresponding to the First and Second Stage of <i>Bachillerato</i> , in both, state and private high schools all over Spain during academic years 1999 – 2000 to 2007 – 2008. Each row shows the percentage of women/men that promote ( $W_i   M_i$ ) and do not promote ( $\overline{W}_i   \overline{M}_i$ ) for each level $i = 1, 2$ (that corresponds to the First and Second Stage of <i>Bachillerato</i> , respectively) over the total Spanish <i>Bachillerato</i> students. These data are referred to the month of September, when each academic year ends officially. . . . .        | 10 |
| 2.2 | Estimation of the model parameters. . . . .  | 14 |
| 2.3 | The model output corresponding to the First and Second Stage of <i>Bachillerato</i> , in both, state and private high schools all over Spain during academic years 2008 – 2009 to 2014 – 2015. Each row shows the rate of women and boys who promote ( $W_i   M_i$ ) and do not promote ( $\overline{W}_i   \overline{M}_i$ ) for each level $i = 1, 2$ . Graphically, it can be seen in Figure 2.2. . . . .   | 15 |
| 3.1 | The available data corresponding to the First and Second Stage of <i>Bachillerato</i> , in both, state and private high schools all over Spain from academic year 1999 – 2000 to 2008 – 2009. Each row shows the percentage of Girls and Boys who promote ( $G_i   B_i$ ) and do not promote ( $\overline{G}_i   \overline{B}_i$ ) for each level $i = 1, 2$ over the total Spanish <i>Bachillerato</i> students. . . . .  | 20 |
| 3.2 | Estimation of positive and negative autonomous decision and abandonment rates. . . . .   | 28 |
| 3.3 | Estimation of positive and negative transmission parameters. . . . .   | 28 |
| 3.4 | The model output corresponding to the First and Second Stage of <i>Bachillerato</i> , in both, state and private high schools all over Spain during academic years 1999 – 2000 to 2014 – 2015. Each row shows the percentage of girls and boys who promote ( $G_i   B_i$ ) and do not promote ( $\overline{G}_i   \overline{B}_i$ ) for each level $i = 1, 2$ . It can be compared the model output values for $t = 1999 - 2000, \dots, 2008 - 2009$ to the data values in Table 3.1 to verify the goodness of the fitting. Graphically, it can be seen in the left-hand side of the graphs in Figure 3.2. . . . . | 29 |

|      |   |    |
|------|---|----|
| 3.5  | Differences between the real data in Table 3.1 and the output model in Table 3.4 corresponding to the First and Second Stage of <i>Bachillerato</i> , both state and private high schools all over Spain during academic years 1999 – 2000 to 2008 – 2009, for Girls Boys who promote ( $G_i B_i$ ) and do not promote ( $\overline{G}_i   \overline{B}_i$ ) for each level $i = 1, 2$ . . . . .  | 31 |
| 3.6  | Matrix of Pearson correlation coefficients for the error terms. . . . .   | 32 |
| 3.7  | The 95% confidence interval prediction corresponding to the First and Second Stage of <i>Bachillerato</i> , in both, state and private high schools all over Spain during academic years 2009 – 2010 to 2014 – 2015. Each row shows the rate of girls/boys who promote ( $G_i B_i$ ) and do not promote ( $\overline{G}_i \overline{B}_i$ ) for each level $i = 1, 2$ . Graphically, it can be seen in Figure 3.7. . . . .  | 35 |
| 3.8  | The 95% confidence interval prediction, the real academic results and the distance between these real data from its corresponding confidence interval. The dash indicates that the point lies inside its confidence interval. Data corresponding to the First and Second Stage of <i>Bachillerato</i> , in both, state and private high schools all over Spain during the academic year 2009 – 2010. Each row shows the rate of girls/boys who promote ( $G_i B_i$ ) and do not promote ( $\overline{G}_i \overline{B}_i$ ) for each level $i = 1, 2$ . . . . . | 37 |
| 3.9  | Estimation of the percentage of abandon in Spanish <i>Bachillerato</i> during the academic years from 1999 – 2000 to 2014 – 2015. . . . .   | 38 |
| 3.10 | Descriptive analysis of the percentage of abandon in Spanish <i>Bachillerato</i> during the academic years from 2009 – 2010 to 2014 – 2015. . . . .   | 38 |
| 4.1  | The available data corresponding to Levels 11, 12 and 13, in both, state and private high schools all over North Rhine-Westphalia from academic year 2006 – 2007 to 2010 – 2011 divided by gender, level and promote/non-promote over the total number of students in the three levels. . . . .   | 43 |
| 4.2  | Estimation of positive and negative autonomous decision and abandon rates. . . . .  | 50 |
| 4.3  | Estimation of positive and negative transmission parameters. . . . .  | 50 |
| 4.4  | The 95% confidence interval predictions corresponding to the Levels 11, 12 and 13, in both, state and private high schools all over the German region of North Rhine-Westphalia during academic years 2011 – 2012 to 2014 – 2015. Each row shows the percentage of girls/boys who promote and do not promote for each academic level. . . . .   | 56 |
| 5.1  | Investment per Spanish student in the First and Second Stage of <i>Bachillerato</i> , in both, state and private high schools all over Spain from academic year 1999 – 2000 to 2008 – 2009 by the Government [1]. . . . .   | 61 |
| 5.2  | The indicators (RMSE and MAPE) considered for the validation of the different models in order to determine the model that best fit the data in Table 5.1. The best is the <i>Linear Trend Model</i> . . . . .   | 62 |

|      |   |     |
|------|---|-----|
| 5.3  | The prediction of euros invested by the Spanish Government in each Spanish student in the First and Second Stage of <i>Bachillerato</i> , in both, state and private high schools during the academic years from 2009 – 2010 to 2014 – 2015. . . . .  | 63  |
| 5.4  | Number of Spanish student in the First and Second Stage of <i>Bachillerato</i> in both, state and private high schools, all over Spain from academic year 1999 – 2000 to 2008 – 2009 [2]. . . . .   | 64  |
| 5.5  | The indicators (RMSE and MAPE) considered for the validation of the different models in order to determine the model that best fit the data in Table 5.4. The best is the <i>Random Walk with Trend Model</i> . . . . .   | 64  |
| 5.6  | Estimations with 95% confidence intervals of the number of Spanish students in the First and Second Stage of <i>Bachillerato</i> in both, state and private high schools, all over Spain from academic year 2009 – 2010 to 2014 – 2015. . . . .   | 65  |
| 5.7  | Estimation with 95% confidence intervals of the number of <i>Bachillerato</i> students who do not promote and abandon in the First and Second Stage of <i>Bachillerato</i> , in both, state and private high schools all over Spain from academic year 2009 – 2010 to 2014 – 2015 and their corresponding cost for the Spanish Government also given with 95% confidence intervals. . . . . | 66  |
| 5.8  | Spanish families investment, on average, per Spanish student in the First and Second Stage of <i>Bachillerato</i> in both, state and private high schools, all over Spain from academic year from 1999 – 2000 to 2008 – 2009 [1]. . . . .   | 68  |
| 5.9  | The indicators (RMSE and MAPE) considered for the validation of the different models in order to determine the model that best fit the data in Table 5.8. The best is the <i>Linear Trend Model</i> . . . . .   | 69  |
| 5.10 | The prediction of euros Spanish families will invest in each Spanish student in the First and Second Stage of <i>Bachillerato</i> , in both, state and private high schools during the academic years from 2009 – 2010 to 2014 – 2015. . . . .  | 69  |
| 5.11 | 95% confidence intervals of the Spanish families cost in the group of <i>Bachillerato</i> students with academic underachievement over the next few years. . . . .  | 70  |
| B.1  | The model output corresponding to the mathematical model shown in both, Chapter 2 and Chapter 3 and the predictions with corresponding 95% confidence intervals obtained in Chapter 3 of the First and Second Stage of <i>Bachillerato</i> , in both, state and private high schools all over Spain during academic years 2008 – 2009 to 2014 – 2015. . . . .                               | 100 |

- 
- B.2 Absolute errors corresponding to the distance between the deterministic predictions given in Chapter 2 and 3 (also shown in Table B.1) and the low or high 95% confidence interval extremes stated in Chapter 3 of the First and Second Stage of *Bachillerato*, in both, state and private high schools all over Spain during academic years 2008 – 2009 to 2014 – 2015. The dashes indicate that the deterministic prediction lies inside its corresponding 95% confidence interval. 101
- C.1 The model output obtained with the estimated parameters (Tables 4.2 and 4.3) in our German model, the real data and their associated absolute errors corresponding the Levels 11, 12 and 13, in both, state and private high schools all over the German region of North Rhine-Westphalia during academic years 2006 – 2007 to 2010 – 2011. Each row shows the percentage of girls/boys who promote and do not promote for each academic level. . . . . 104

# Chapter 1

## Introduction

In recent years, there has been increased awareness of the importance of education in society by both governments and society in general. Education largely determines the professional life of an individual. It has an impact on the ease of getting and keeping a job and also influences on the conditions and characteristics of the job. This important issue has led to the educational experts and policy makers to focus their attention on the evolution of the academic results of students. There are many contributions which show us the increasing concern about young students' academic performance in worldwide [1, 3, 4], mainly, focusing on the bad academic results which could have serious negative influences on the country's economic development. Obviously, the better education of the population, the greater benefits could be brought by the population to the country.

In general terms, the definition of obtaining bad academic results is commonly called 'academic underachievement'. However, this expression presents several definitions due to the different perspectives given by recognized educational experts. In [5] academic underachievement is defined as: 'Students who, during their stay in school, have obtained a minimum preparation to enable them to live independently in society: find a job, organize independently and behave civic, responsible and tolerant'. In [6], the author says that 'academic underachievement has, at least, three points: low academic performance, difficulties in adapting to the rules of coexistence and destruction of self-esteem'. In [7], the author defines it as 'academic underachievement is any insufficiency detected in the results achieved by pupils in schools regarding the proposed objectives for their level, age and development, and is usually expressed through negative grades'. In addition,

the Spanish educational law in force stated that 'academic underachievement is considered when a student has not realized any improvement in their academic performance from 18 to 25 years old'.

It is very difficult to quantify the academic underachievement if we use any of the above-mentioned definitions. It leads us to propose our own definition of academic underachievement, more strict than the ones in the previous paragraph, but it will allow the quantification. Therefore, in this dissertation, we identify academic underachievement when a student who, during her/his stay in school in a specific academic year, has not been able to get the proposed objectives for her/his academic level such as it is established by the educational law in force and, as a consequence, she/he is not allowed to pass into a higher level or she/he has decided to abandon her/his studies.

During the last few years, Spanish academic authorities have taken numerous educational measures in order to combat academic underachievement in high schools including an increase in funding and resources destined to education improvement [8]. These efforts have been carried out by several changes in the educational laws looking for an improvement in academic results [9, 10]. This concern has even become a compelling reason for the enactment of the last education law [10], which states in its abroad outlines: 'It pretends to reduce the rate of academic underachievement and to improve the level of performance of students'. These legislative measures have been focused, mainly, on the educational levels in which have appeared an increasing of academic underachievement: Compulsory Secondary Education (in the Spanish terminology, Educación Secundaria Obligatoria (ESO)) and the last two high school courses (in the Spanish terminology, *Bachillerato*) that theoretically correspond to students between 12–16 and 16–18 years old, respectively. As a result of these educational measures, although the rate of the academic underachievement has slightly reduced over the last years, these rates in these educational stages are still at very worrying levels about 30% of the pupils [11]. This situation is more alarming if we compare these figures with the corresponding ones to the rest of the countries of the European Union in which the levels of academic underachievement are around the 15% [3, 4, 12].

Nowadays, the job opportunities of people depend on their qualification, their ability to acquire, use and interpret the information, including their skills to adapt the new knowledge to a very demanding and competitive society in constant change. In order to acquire them, students go to basic schools first and high

schools later, learning the contents determined in the corresponding legislations. As a consequence, high rates of academic underachievement could lead that the number of qualified workers is less than the number of qualified jobs expected over the following years [13, 14]. This is also a primer concern of the European Union [15–17].

The interest about academic underachievement in Spain is completely justified, not only by the high rates but also it is becoming a major social and political concern [18–20], especially in the unemployment and its serious consequences. This issue is of primer importance in the current context of economic crisis affecting particularly Spain. In fact, when the economic crisis started around the year 2008 affecting negatively on the international labor market, in Spain, the unemployment rates were twice higher than the rest of the European Countries [21]. Moreover, in 2012, the 80% of the Spanish people that had finished their higher studies, accessed to the work market while the Spanish population who had only got ESO or lower educational levels was around 27% [22].

To deal with this problem, in this work we focus our attention on the Spanish *Bachillerato* educational stage for several reasons. First, from a mathematical standpoint, *Bachillerato* has a simpler academical structure that seems to be an adequate start-point to introduce these type of modeling approach. Second, from a social viewpoint, as we pointed out previously, *Bachillerato* is a milestone in the career training of students because it represents a period to make important decisions about academic and professional future [23]. When they finish *Bachillerato* they can decide whether to continue their higher studies (university or professional training) or to access the work market. This is of paramount importance for society because, although the percentage of high school academic underachievement has slightly reduced over the last years, nowadays it seems to be at a worrying steady-level. This constitutes a serious problem not only for these individuals and their families but also for the society that has invested an important amount of money in their previous training.

The access to this academic educational level takes place after finishing the Compulsory Secondary Education (ESO) where students should stay for four years (from 12 to 16 years old). Then, a great part of them decide to access to *Bachillerato*, which is not mandatory and, as it was pointed out, corresponds to students who are 16 – 18 years old. *Bachillerato* consists of two years and the subjects are more specialized in science, literature or art than in ESO. In Spain, students can

especialise in three different *Bachillerato* branches: Arts; Sciences and Technology; Humanities and Social Sciences (see Figure 1.1). In accordance with educational regulations in force in Spain [10], in September, a student of First Stage of *Bachillerato* will transit into the Second Stage of *Bachillerato* if she/he has passed successfully all the subjects except, maybe, at most two out of a total of ten. In order to get the *Bachillerato* degree, it is necessary to pass all the subjects of both *Bachillerato* Stages, in particular, a student will pass any subject of *Bachillerato* subject if she/he gets, at least, 5 points out of 10 in the final academic results.

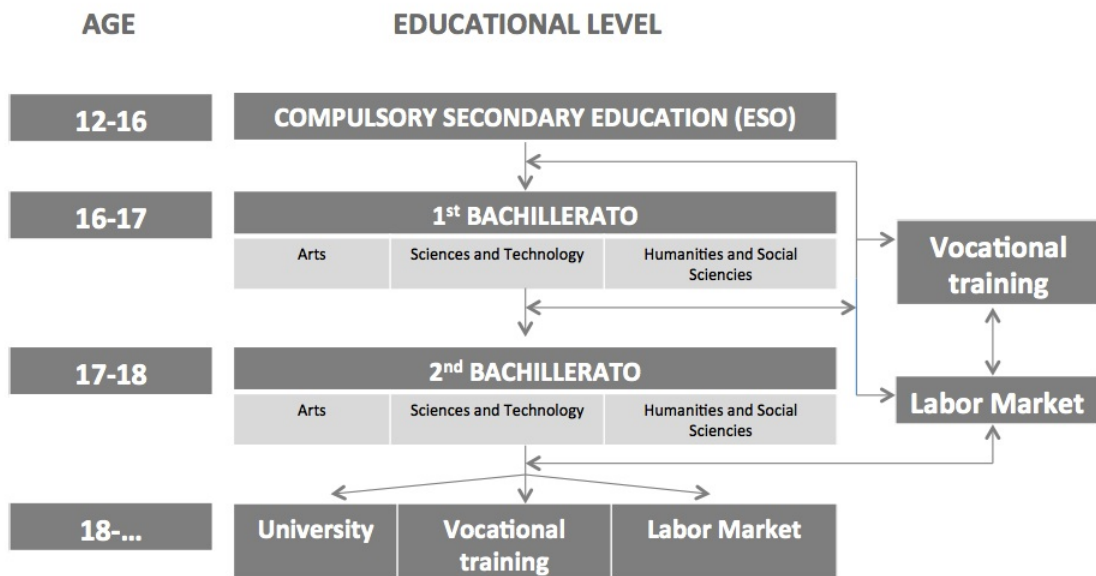


FIGURE 1.1: Structure of the Spanish educational system for students aged between 12 – 18 years old.

The different learning theories [24, 25], in particular, the Vygotskian perspective [26–28] and the recent studies published [29, 30], confirm that habits and behavior may be socially transmitted. In our context, we assume that social contacts have a great influence on transmission of study habits [23, 31]. The main idea behind our study of the academic performance is that these appropriate or inappropriate habits may spread from one student to another, more probably between students of the same academic level [23].

Some examples of analogous situations (related to social behavior modelling) using type-epidemiological mathematical models are encountered in public health, obesity [30, 32], alcoholism [33], drug abuse [34], shopaholism [35], spread of ideas [36], evaluation of law effects on societies [37], and so on. To point out that the use of epidemiological models is widely known by the scientific community, especially,



for the spread of diseases. However, although there is a high interest in analysing the academic performance [38–40], to the best of our knowledge, there is no evidence that these techniques have been applied to analyze academic performance previously.

To address this approach, we propose a non-linear system of differential equations to model the evolution of the academic performance in the educational level of *Bachillerato* in Spain using modeling techniques in mathematical epidemiology. To analyze the academic performance, we have focused on the available academic results belonging to the students of the First and Second Stage of *Bachillerato* during the academic years from 1999 – 2000 to 2007 – 2008, in both, state and private high schools all over Spain [2, 41]. *Bachillerato* students will be classified in promote and non-promote group according to the obtained academic results at the end of each academic year. Moreover, in order to reflect as truthfully as possible the attitude of students towards their studies, we have taken into account pedagogical studies [42–45] which confirm that exist a significative difference of academic performance depending on genre considering, in general, women obtain better academic results than men. This motivates the fact that our model considers genre into its formulation. We also consider that the transmission of academic habits is carried out between students of the same academic level [23, 46, 47].

Once the model is stated, we will be able to monitor the promoted and graduated students. The estimation of the parameters of our model will allow us to predict the evolution of the academic performance in specific confidence intervals. Furthermore, we include the estimation of the abandon rates. Abandon is an important and sensitive aspect still under debate in the pedagogical studies and is not commonly available. We have made a decision in order to include this issue in the model and this is to consider *abandon* when, during the academic year, the student leaves the academic system. As a consequence, we could predict the percentage of Spanish population may be less qualified at time to enter in the work market with their negative consequences. An important characteristic of the proposed model is its ability to be adapted, although in this research we have focused on *Bachillerato* educational level, to other Spanish educational levels and also foregoing any educational systems. This is illustrated in Chapter 4 where the proposed model for Spanish *Bachillerato* has been adapted successfully to study the academic performance in the German region of North Rhine-Westphalia.

Moreover, high rates of academic underachievement have strong negative effects on the economic situation of families and the Spanish Government, mainly, in the current economic crisis in Spain. On the one hand, families must spend a lot of money on each of their children's education and it has been increasing as time goes reaching values that, on average, are around 1 300 euros per year for each Spanish student [48, 49]. There are many needs that parents have to deal with such as the school fees, books, uniform and, in some cases, accommodation each academic year, money that would be invested again if the student is not able to promote during the academic year. In the same way, each year the Spanish State Management spends a high percentage of its budget on education [1], a waste of large amounts of money if the rates of academic underachievement are increasing [20]. The predictions given in this dissertation of the Spanish *Bachillerato* academic results in the coming years will allow to provide an estimation of the investment could be made in this educational level by both, the Spanish Government and families. We will pay special attention on the groups of students who abandon and do not promote during their corresponding academic year whose academic attitude could lead a high economic costs for the Spanish Government and families.

The proposed mathematical approach will allow us to understand better the mechanisms behind the academic performance as well as to predict how things will evolve in the Spanish *Bachillerato* over the next few years and this way, to provide relevant information to make appropriate decisions to policymakers. In addition, this predictions will allow us to quantify the large costs that would entail for the Spanish society the high rates of academic underachievement in the coming years.

We conclude this introduction by showing more details of the contents of this dissertation.

In Chapter 2, we propose a dynamic model based on a non-linear system of differential equations to understand the evolution of the academic underachievement in students First and Second Stage of *Bachillerato* in Spain taking into account the different attitude of students according their gender. To build the model we suppose that a student obtains negative academic results when she/he decides autonomously to adopt a negative academic habits and also when she/he gets into study habits socially transmitted by students with bad academic habits. From the available *Bachillerato* academic results in both, state and private Spanish high schools during the period 1999 – 2000 to 2007 – 2008, we fit the model to the data and obtain the parameters of the model. Then, we predict the academic

underachievement evolution over the next few years. In Chapter 3, our approach is to improve the dynamical model statement. In this case, the improved model is based on the idea that not only the bad academic habits are transmitted but also the good study habits. We also decompose the transmission parameters into good and bad academic habits, in order to analyze with more detail which group of students are more susceptible to be influenced by good or bad academic students. Besides, we introduce uncertainty in the model, a bootstrapping approach is employed. The model presented in this chapter is validated verifying that the 95% confidence intervals predicted collect the estimations given in Chapter 2 and also with new available data which were published during the development of this dissertation. Other important improvement in this model is the quantification of the abandon rates. In Chapter 4, we show an application of our improved mathematical model to the German educational system. This has been possible thanks to the grant received from the Spanish Ministry of Education which allowed to stay during three months (from 7th April to 29th June, 2012) in Bergische Universität Wuppertal (Germany) working under the supervision of Prof. M. Ehrhardt. During this stay, we had the opportunity to access to the academic results belonging to students in high school of the German region of North Rhine-Westphalia which enabled us to apply our model and to analyze the evolution of the academic performance of students in this German region. Additionally, it also allows us to test the ability of the proposed method to be adapted to other educational systems as well as to compare the educational training of Spanish and German students. This dissertation ends by studying in Chapter 5 the negative economical effects which could have for both, the Spanish Government and families, the high rates of academic underachievement in the next few years according to the predictions obtained from our dynamic model. Finally, the conclusions are given in Chapter 6.



# Chapter 2

## Predicting the academic underachievement in high school in Spain over the next few years: A dynamic modelling approach

### 2.1 Introduction

In this chapter we propose a dynamic model to understand the evolution of the academic underachievement in high school in Spain. This model is based on ideas of Christakis and Fowler which confirm that individual habits may be transmitted by social contact [29, 30]. Moreover, in order to reflect as truthfully as possible the attitude of students towards their studies, we have taken into account pedagogical studies [42–45] which state that exists a significative difference of academic performance depending on genre. This motivates the fact that our model considers genre into its formulation. We also consider that the transmission of academic habits is carried out between students of the same academic level [23, 46, 47]. Thus, to build the model we suppose that a student has academic underachievement when she/he gets into study habits transmitted by students with bad academic habits. From the available academic results of the Spanish high school educational system during the period 1999–2008 [2, 41], we fit the model to the data in order to obtain the parameters of the model. Then, we predict the academic underachievement evolution over the next few years.

## 2.2 Building the mathematical model

### 2.2.1 Available data

We have considered the available data corresponding to the First and Second Stage of *Bachillerato*, in both, state and private high schools all over Spain during academic years 1999 – 2000 to 2007 – 2008. We point out that we study this period because after the academic year 2007 – 2008 only partial data corresponding to specific regions of Spain are available. For each of these academic years, Table 2.1 collects the percentage of women/men that pass ( $W_i | M_i$ ) and do not pass ( $\bar{W}_i | \bar{M}_i$ ) for each level  $i = 1, 2$  (that corresponds to the First and Second Stage of *Bachillerato*, respectively) over the total Spanish *Bachillerato* students. These data have been obtained from the official database [2, 41] and they are referred to the month of September, when each academic year ends officially.

| Academic year | First Stage of <i>Bachillerato</i><br>(Women   Men) |  | Second Stage of <i>Bachillerato</i><br>(Women   Men) |  |
|---------------|---|--|--|--|
|               | % Pass<br>( $W_1   M_1$ )                           | % Do not Pass<br>( $\bar{W}_1   \bar{M}_1$ ) | % Pass<br>( $W_2   M_2$ )                            | % Do not Pass<br>( $\bar{W}_2   \bar{M}_2$ ) |
| 1999 – 00     | 19.68   15.24                                       | 9.75   9.33                                  | 16.21   11.64  | 9.52   8.63                                  |
| 2000 – 01     | 22.65   17.54                                       | 9.91   10.12                                 | 14.07   10.04  | 8.24   7.43                                  |
| 2001 – 02     | 19.23   14.23                                       | 8.61   9.10                                  | 17.86   13.06  | 9.32   8.59                                  |
| 2002 – 03     | 18.87   14.19                                       | 8.36   8.51                                  | 19.14   13.97  | 8.76   8.20                                  |
| 2003 – 04     | 19.93   15.06                                       | 7.74   7.88                                  | 19.19   13.80  | 8.44   7.96                                  |
| 2004 – 05     | 20.11   15.14                                       | 7.65   7.94                                  | 18.90   13.92  | 8.39   7.95                                  |
| 2005 – 06     | 20.07   15.39                                       | 7.64   7.93                                  | 19.14   13.97  | 8.08   7.78                                  |
| 2006 – 07     | 20.06   15.34                                       | 7.67   7.87                                  | 19.14   14.29  | 7.98   7.65                                  |
| 2007 – 08     | 20.25   15.82                                       | 7.57   7.66                                  | 19.37   14.61  | 7.60   7.12                                  |

TABLE 2.1: The available data corresponding to the First and Second Stage of *Bachillerato*, in both, state and private high schools all over Spain during academic years 1999 – 2000 to 2007 – 2008. Each row shows the percentage of women/men that promote ( $W_i | M_i$ ) and do no promote ( $\bar{W}_i | \bar{M}_i$ ) for each level  $i = 1, 2$  (that corresponds to the First and Second Stage of *Bachillerato*, respectively) over the total Spanish *Bachillerato* students. These data are referred to the month of September, when each academic year ends officially.

We build our mathematical model, based on an epidemiological-type model, by considering that academic underachievement is a process that takes place when a female (W) or a male (M) student that initially belongs to the promotable group of a specific level,  $W_i$  or  $M_i$ ,  $i = 1, 2$ , leaves her/his good academic habits due to the negative influence (*contagion*) from other students of the same educational level

who belong to the group of students with academic underachievement,  $\bar{W}_i$  and  $\bar{M}_i$ . We emphasize that, in spite of the fact that Spanish *Bachillerato* students share the same educational center with students belonging to other lower educational levels, it is realistic to assume that from an academic performance point of view, in general, the *contagion* between them is not significative. This statement is justified by the differences in age and maturity between students that belong to different educational levels, their physical location in the high school, etc. [23, 46, 47]. In the model, these considerations will be taken into account even for *Bachillerato* students of different academic level. We shall consider that, for each specific academic level under study, the bad academic habits just spread between students of the same course, independently of their genre. Thus, the transitions described can be modeled as follows:

- For a specific *Bachillerato* academic level  $i = 1, 2$ , a student in  $W_i$  (respec.  $M_i$ ) transits to  $\bar{W}_i$  (respec.  $\bar{M}_i$ ) because students in  $\bar{W}_i$  and  $\bar{M}_i$  transmit their negative academic habits at rates  $\beta_i^W$  (respec.  $\beta_i^M$ ). Therefore, this is a non-linear term modeled by  $\beta_i^W W_i(\bar{W}_i + \bar{M}_i)$  (respec.  $\beta_i^M M_i(\bar{W}_i + \bar{M}_i)$ ). Note that modeling above assumes implicitly homogeneous population mixing for each academic level under consideration. This is a usual assumption in type-continuous epidemiological models [30, 32–36, 50]. In this context, it is realistic to consider that a particular attitude of any student towards their studies not only can be influenced by her/his autonomous behavior but also by their mates, the study center and, in general, by the social environment surrounded.
- For a specific *Bachillerato* academic level  $i = 1, 2$ , students can also acquire bad academic habits because they autonomously decide to strive less and to give up appropriate study habits due to lack of self-motivation, personal problems, etc.
- Data collected in Table 2.1 refer to the end of each academic course, i.e., in September when, according to educational regulation in force in Spain, every student in  $W_1$  and  $M_1$  will pass to  $W_2$  and  $M_2$ , respectively. Taking into account that time  $t$  is measured in years and we identify each academic year in the period 1999 – 2000 to 2007 – 2008, with  $0, 1, \dots, 8$ , respectively, this is modeled by the  $\delta$  parameter indicated in Figure 2.1 and is defined as

follows:

$$\delta = \begin{cases} 1 & \text{if } \frac{9}{12} + j \leq t \leq \frac{10}{12} + j, \quad j = 0, 1, 2, \dots, 8, \\ 0 & \text{otherwise,} \end{cases}$$

where  $\delta = 1$  if a *Bachillerato* student (woman/men) passes into a higher academic level in September and 0 otherwise. This parameter allows to model the transitions of students who pass successfully in September (ninth month of the year) from First to Second Stage of *Bachillerato*.

- For a specific *Bachillerato* academic level  $i = 1, 2$ , a student in  $\overline{W}_i$  (respec.  $\overline{M}_i$ ) transits to  $W_i$  (respec.  $M_i$ ), when she/he gives up her/his bad academic habit. An individual in  $\overline{W}_i$  (respec.  $\overline{M}_i$ ) transits to  $W_i$  (respec.  $M_i$ ) at rate  $\gamma_i^W$  (respec.  $\gamma_i^M$ ) proportionally to the size of  $\overline{W}_i$  (respec.  $\overline{M}_i$ ). Analogously to  $\alpha_i^W$  and  $\alpha_i^M$ , parameters  $\gamma_i^W$  and  $\gamma_i^M$  also contain those autonomous decisions adopted by students belonging to  $\overline{W}_i$  and  $\overline{M}_i$ , respectively.

Then, the transitions between these different subpopulations are described by the following coupled non-linear system of differential equations where the unknowns are  $W_i = W_i(t)$ ,  $M_i = M_i(t)$ ,  $\overline{W}_i = \overline{W}_i(t)$  and  $\overline{M}_i = \overline{M}_i(t)$  ( $t$  denotes time in years),

$$\begin{aligned} W_1'(t) &= -\delta W_1(t) - \alpha_1^W W_1(t) - \beta_1^W W_1(t)[\overline{W}_1(t) + \overline{M}_1(t)] + \gamma_1^W \overline{W}_1(t), \\ \overline{W}_1'(t) &= \alpha_1^W W_1(t) + \beta_1^W W_1(t)[\overline{W}_1(t) + \overline{M}_1(t)] - \gamma_1^W \overline{W}_1(t), \\ W_2'(t) &= \delta W_1(t) - \alpha_2^W W_2(t) - \beta_2^W W_2(t)[\overline{W}_2(t) + \overline{M}_2(t)] + \gamma_2^W \overline{W}_2(t), \\ \overline{W}_2'(t) &= \alpha_2^W W_2(t) + \beta_2^W W_2(t)[\overline{W}_2(t) + \overline{M}_2(t)] - \gamma_2^W \overline{W}_2(t), \\ M_1'(t) &= -\delta M_1(t) - \alpha_1^M M_1(t) - \beta_1^M M_1(t)[\overline{W}_1(t) + \overline{M}_1(t)] + \gamma_1^M \overline{M}_1(t), \\ \overline{M}_1'(t) &= \alpha_1^M M_1(t) + \beta_1^M M_1(t)[\overline{W}_1(t) + \overline{M}_1(t)] - \gamma_1^M \overline{M}_1(t), \\ M_2'(t) &= \delta M_1(t) - \alpha_2^M M_2(t) - \beta_2^M M_2(t)[\overline{W}_2(t) + \overline{M}_2(t)] + \gamma_2^M \overline{M}_2(t), \\ \overline{M}_2'(t) &= \alpha_2^M M_2(t) + \beta_2^M M_2(t)[\overline{W}_2(t) + \overline{M}_2(t)] - \gamma_2^M \overline{M}_2(t). \end{aligned} \tag{2.1}$$

The parameters of the model are:

- $\alpha_i^g$ , denotes the rate at which a student of Spanish *Bachillerato* academic level  $i$  and genre  $g$ , who belongs to the promoting group, passes to have bad academic habits by an autonomous decision. In accordance with personal trait patterns and academic performance of adolescents [51], it is considered



that for the same educational stage, women are more responsible with respect to studies matters than men. Whereas comparing the same genre, students that belong to the Second Stage of *Bachillerato* get more involved than students of the First Stage due not only to maturity but also they feel more self-motivated because they are about to get the *Bachillerato* degree [42–45]. In consequence, we consider these academic behavioral conclusions by assuming, respectively, that

$$\alpha_1^W < \alpha_1^M, \alpha_2^W < \alpha_2^M; \quad \alpha_1^W > \alpha_2^W, \alpha_1^M > \alpha_2^M. \quad (2.2)$$

- $\beta_i^g$ , denotes the transmission rate at which a student of Spanish *Bachillerato* academic level  $i$  and genre  $g$  adopts bad academic habits due to the negative influence from students that do not pass and belong to the same academic level  $i$  including both genre.
- $\gamma_i^g$ , denotes the rate at which a student of Spanish *Bachillerato* academic level  $i$  and genre  $g$  who has bad academic habits, by an autonomous decision, decides to change her/his bad academic habits and she/he ends up getting into the passing group.

The flow diagram, associated to the model, is depicted in Figure 2.1.

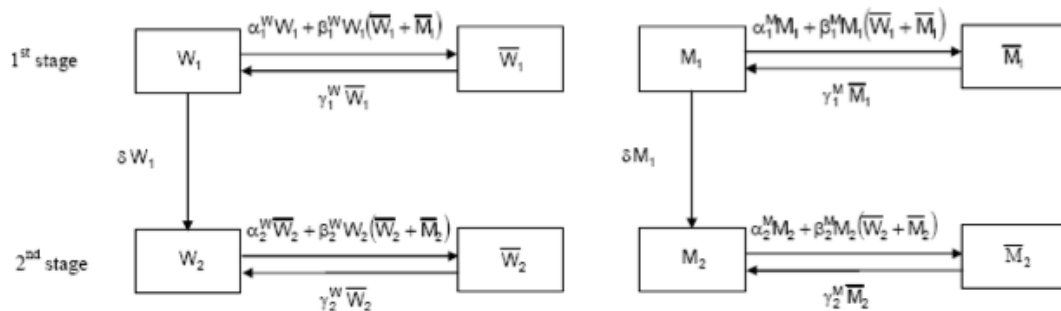


FIGURE 2.1: Flow diagram of the mathematical model for *Bachillerato* academic underachievement in Spain. The boxes represent subpopulations under study classified according to genre (women (W) and men (M)) and academic level (First and Second Stage of *Bachillerato*). Students that belong to the promote/non-promote group are denoted by  $W$ ,  $M$ ,  $\bar{W}$  and  $\bar{M}$ , respectively. The arrows represent the transitions between the subpopulations, and they are labeled by their corresponding terms and parameters according to the model.

## 2.3 Parameter estimation

In this section we are going to present the estimation of the parameters of (2.1) by fitting the model in the mean square sense to the data collected in Table 2.1.

In order to do that, we have implemented a function of the parameters in *Mathematica 8.0* [52] taking as initial conditions of the system of differential equations (2.1), we take data of the academic year 1999 – 2000 (corresponding to  $t = 0$ ), so  $W_1(0) = 19.68$ ,  $M_1(0) = 15.24$ ,  $\overline{W}_1(0) = 9.75$ ,  $\overline{M}_1(0) = 9.33$ ,  $W_2(0) = 16.21$ ,  $M_2(0) = 11.64$ ,  $\overline{W}_2(0) = 9.52$  and  $\overline{M}_2(0) = 8.63$ . The process for obtaining the parameters that best fit (in the mean square sense) is as follows:

- Use *Mathematica* command `NDSolve` to solve the system of differential equations obtained (2.1).
- Substitute the solution in the time instants 1999 – 2000, ..., 2007 – 2008 corresponding to the academic courses.
- Compute the root mean square error between the model outputs (obtained in the previous step) and the data in Table 2.1.

The parameters that best fit the model to data will be those that minimize the above function. The minimization process has been done using the Nelder-Mead algorithm [53] included in the *Mathematica* command `NMinimize`.

The estimated parameters are collected in Table 2.2. The least square error that we have obtained is 0.0095.

| Gender | Negative autonomous decision |             | Negative transmission |          | Positive autonomous decision |          |
|--------|------------------------------|-------------|-----------------------|----------|------------------------------|----------|
|        | Parameter                    | Value       | Parameter             | Value    | Parameter                    | Value    |
| Women  | $\alpha_1^W$                 | 0.0180156   | $\beta_1^W$           | 0.538867 | $\gamma_1^W$                 | 0.285557 |
|        | $\alpha_2^W$                 | 0.000119175 | $\beta_2^W$           | 0.668998 | $\gamma_2^W$                 | 0.270119 |
| Men    | $\alpha_1^M$                 | 0.0180175   | $\beta_1^M$           | 2.59115  | $\gamma_1^M$                 | 0.853703 |
|        | $\alpha_2^M$                 | 0.00307025  | $\beta_2^M$           | 1.38138  | $\gamma_2^M$                 | 0.405319 |

TABLE 2.2: Estimation of the model parameters.

Regarding Table 2.2, for both educational stages, the transmission rate at which male students of *Bachillerato* adopt bad academic habits due to influence of students that do not pass is greater than the corresponding one for female students, i.e.,  $\beta_1^M > \beta_1^W$  and  $\beta_2^M > \beta_2^W$ . Whereas the rate at which a male student of *Bachillerato* who has bad academic habits decides autonomously to change his bad academic habits and he ends up getting into the passing group is greater than the corresponding one for a female student, i.e.,  $\gamma_1^M > \gamma_1^W$  and  $\gamma_2^M > \gamma_2^W$ . We point out that both results agree with conclusions derived from other pedagogical studies [51].

## 2.4 Prediction over next few years

Now, once the model is stated and the parameters estimated, we are able to give predictions of each subpopulation over the next few years computing the solutions of the model,  $W_1(t)$ ,  $\bar{W}_1(t)$ ,  $M_1(t)$ ,  $\bar{M}_1(t)$ ,  $W_2(t)$ ,  $\bar{W}_2(t)$ ,  $M_2(t)$  and  $\bar{M}_2(t)$  for values of time  $t$  in the future. The solution to model (2.1) is given in percentage in Table 2.3 and plotted in Figure 2.2.

| Academic year | First Stage of <i>Bachillerato</i><br>(Women   Men) |  | Second Stage of <i>Bachillerato</i><br>(Women   Men) |  |
|---------------|---|--|--|--|
|               | % Pass<br>( $W_1$   $M_1$ )                         | % Do not Pass<br>( $\bar{W}_1$   $\bar{M}_1$ ) | % Pass<br>( $W_2$   $M_2$ )                          | % Do not Pass<br>( $\bar{W}_2$   $\bar{M}_2$ ) |
| 2008 – 2009   | 0.21138   0.15997                                   | 0.08293   0.08573                              | 0.18798   0.13589                                    | 0.06932   0.06681                              |
| 2009 – 2010   | 0.21197   0.16033                                   | 0.08233   0.08537                              | 0.18945   0.13692                                    | 0.06785   0.06578                              |
| 2010 – 2011   | 0.21246   0.16063                                   | 0.08184   0.08508                              | 0.19073   0.13783                                    | 0.06657   0.06487                              |
| 2011 – 2012   | 0.21287   0.16087                                   | 0.08143   0.08483                              | 0.19187   0.13865                                    | 0.06543   0.06406                              |
| 2012 – 2013   | 0.21324   0.16109                                   | 0.08106   0.08461                              | 0.19289   0.13939                                    | 0.06441   0.06331                              |
| 2013 – 2014   | 0.21360   0.16129                                   | 0.08070   0.08441                              | 0.19386   0.14010                                    | 0.06344   0.06260                              |
| 2014 – 2015   | 0.21397   0.16150                                   | 0.08033   0.08420                              | 0.19480   0.14080                                    | 0.06250   0.06190                              |

TABLE 2.3: The model output corresponding to the First and Second Stage of *Bachillerato*, in both, state and private high schools all over Spain during academic years 2008 – 2009 to 2014 – 2015. Each row shows the rate of women and boys who promote ( $W_i$  |  $M_i$ ) and do not promote ( $\bar{W}_i$  |  $\bar{M}_i$ ) for each level  $i = 1, 2$ . Graphically, it can be seen in Figure 2.2.

According to the global prediction of the model, note that the total percentage of *Bachillerato* students that will not pass is worrying because it will lie

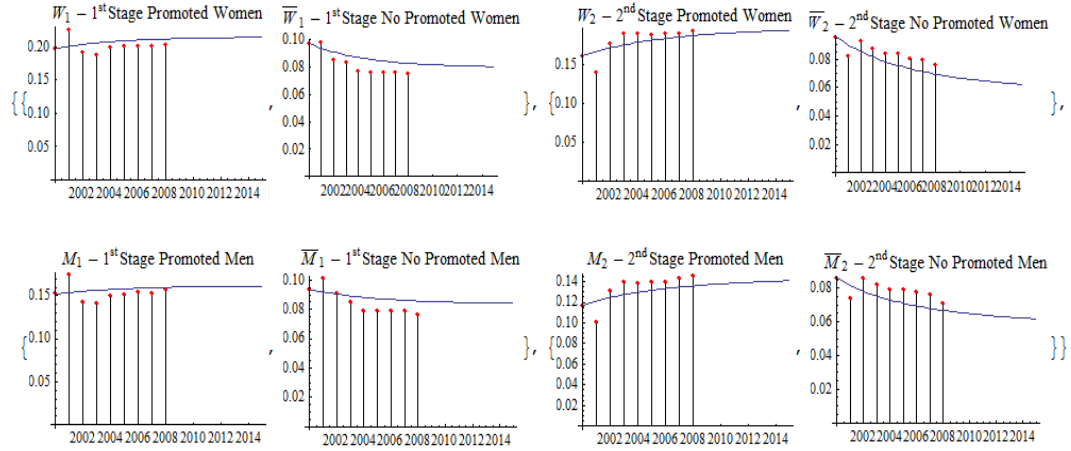


FIGURE 2.2: Fitting and prediction of the academic performance of *Bachillerato* Spanish students over the academic years 1999 – 2000 to 2014 – 2015.

around 30% which still constitutes a high rate. In accordance with the model, we consider that the First Stage of *Bachillerato* is the key level to begin to combat academic underachievement in *Bachillerato* because it has greater associated academic underachievement rates than Second Stage (around 17% over the total Spanish *Bachillerato* students).

## 2.5 Conclusions

In this chapter we have proposed a continuous model to study academic underachievement in the last educational stage of the Spanish high school, called *Bachillerato*. The major novelty of this contribution is the treatment of academic underachievement as a problem that is transmitted through social contact including its mathematical type-epidemiological modeling. We point out that we used hypotheses appearing in other studies [23, 29, 31, 42–47], as homogeneous mixing, habits transmission dynamics and genre groups. In addition, the model allows us to see how things will evolve over the next future. Hence, it enables us to make an estimation of, both the number of pupils who will apply for higher education and the students who do not get the minimum knowledge to pass into the next course. According to the obtained results in Table 2.3 and Figure 2.2, the academic underachievement in the Spanish *Bachillerato* could be around of 30% over the total number of Spanish *Bachillerato* students.

In this chapter, the first proposed model has been presented. In the next chapter, we will show an improved mathematical model based on the presented one. In the next model, in addition to the assumptions included in the current one, we will consider the transmission of good academic habits and we also decompose the transmission parameters into good and bad academic habits, in order to analyze with more detail which group of students are more susceptible to be influenced by good or bad academic students. In addition, we will obtain the estimation of the abandon rates. Other important improvement will be seen in the model to be presented in the next chapter, and it is the inclusion of uncertainty in it which will allow us to present our predictions using confidence intervals.



# Chapter 3

## Non-parametric probabilistic forecasting of academic performance in Spanish high school using an epidemiological modelling approach

### 3.1 Introduction

In this chapter, we propose an improved mathematical model respect to the model developed in Chapter 2. Our approach is also based on the idea that academic habits of any student is a mixture of personal decisions and influence of classmates. In this case, we consider that not only the bad academic habits are transmitted but also the good ones. We want to analyze in more detail which groups of students are more susceptible to be influenced by good or bad academic students. Moreover, in order to consider the uncertainty in the model, a bootstrapping approach is employed. This technique permits to forecast model trends in the next few years using confidence intervals. In addition, the model presented is validated verifying that the 95% confidence intervals predicted collect the deterministic estimations given in Chapter 2 in the academic years 2008 – 2009 to 2014 – 2015, and also with new available data which have been published during the development of this

dissertation. This improved model also allows us to forecast the abandon rates, information which is not commonly available.

## 3.2 The epidemiological-mathematical model

### 3.2.1 Available data

The available data that we have considered in this chapter correspond to the academic results belonging to the students of the First and Second Stage of *Bachillerato* during the academic years from 1999 – 2000 to 2008 – 2009, in both, state and private high schools all over Spain. Notice that these data are the same as it has been used in Chapter 2 although, in this case, it was possible to include the academic results of Spanish *Bachillerato* belonging to the academic year 2008 – 2009 [41] since as time goes new *Bachillerato* academic results of this academic results have been published. The available data can be seen in Table 3.1.

| Academic year | First Stage (Girls   Boys)  |   | Second Stage (Girls   Boys) |   |
|---------------|-----------------------------|---|-----------------------------|---|
|               | % Promote ( $G_1$   $B_1$ ) | % Non-Promote ( $\bar{G}_1$   $\bar{B}_1$ ) | % Promote ( $G_2$   $B_2$ ) | % Non-Promote ( $\bar{G}_2$   $\bar{B}_2$ ) |
| 1999 – 2000   | 19.68   15.24               | 9.75   9.33                                 | 16.21   11.64               | 9.52   8.63                                 |
| 2000 – 2001   | 22.65   17.54               | 9.91   10.12                                | 14.07   10.04               | 8.24   7.43                                 |
| 2001 – 2002   | 19.23   14.23               | 8.61   9.10                                 | 17.86   13.06               | 9.32   8.59                                 |
| 2002 – 2003   | 18.87   14.19               | 8.36   8.51                                 | 19.14   13.97               | 8.76   8.20                                 |
| 2003 – 2004   | 19.93   15.06               | 7.74   7.88                                 | 19.19   13.80               | 8.44   7.96                                 |
| 2004 – 2005   | 20.11   15.14               | 7.65   7.94                                 | 18.90   13.92               | 8.39   7.95                                 |
| 2005 – 2006   | 20.07   15.39               | 7.64   7.93                                 | 19.14   13.97               | 8.08   7.78                                 |
| 2006 – 2007   | 20.06   15.34               | 7.67   7.87                                 | 19.14   14.29               | 7.98   7.65                                 |
| 2007 – 2008   | 20.25   15.82               | 7.57   7.66                                 | 19.37   14.61               | 7.60   7.12                                 |
| 2008 – 2009   | 20.72   16.57               | 7.28   7.43                                 | 19.43   14.86               | 7.05   6.66                                 |

TABLE 3.1: The available data corresponding to the First and Second Stage of *Bachillerato*, in both, state and private high schools all over Spain from academic year 1999 – 2000 to 2008 – 2009. Each row shows the percentage of Girls and Boys who promote ( $G_i$  |  $B_i$ ) and do not promote ( $\bar{G}_i$  |  $\bar{B}_i$ ) for each level  $i = 1, 2$  over the total Spanish *Bachillerato* students.



### 3.2.2 Model building

We build our improved mathematical model based on the same ideas as we presented in the mathematical model in Chapter 2, that is, based on idea that individual habits may be transmitted by social contact [29, 30] considering as main idea that the academic performance of a student, Girl (G) or Boy (B), is a mixture of her/his own study habits and the study habits of their classmates. We have also taken into account pedagogical studies [42–45] which confirm that exist a significative difference of academic performance depending on genre and also consider that the transmission of academic habits is carried out between students of the same academic level [23, 46, 47]. However, in this improved model we consider not only the transmission of bad academic habits but also the transmissions of good academic habits among *Bachillerato* students in the same academic level.

The subpopulations we have considered in this model are the same as the ones we considered in the previous chapter (also time  $t$  in years), that is,

- $G_i = G_i(t)$  is the number of girls who promote at time  $t$ , for  $i = 1, 2$ .
- $B_i = B_i(t)$  is the number of boys who promote at time  $t$ , for  $i = 1, 2$ .
- $\bar{G}_i = \bar{G}_i(t)$  is the number of girls who do not promote at time  $t$ , for  $i = 1, 2$ .
- $\bar{B}_i = \bar{B}_i(t)$  is the number of boys who do not promote at time  $t$ , for  $i = 1, 2$ .

Furthermore, we consider, besides the assumptions we presented in the previous model, additionally assumptions to build this improved model. The assumptions considered to define this model are the following:

- Let us assume homogeneous population mixing following the lead of other recently published works [30, 32–36, 50]. As it was stated in the previous model (see Chapter 2), in this context, it is also considered that a particular attitude of any student towards their studies not only can be influenced by her/his autonomous behavior but also by their mates, the study center and, in general, the social environment surrounded which may condition a student to adopt different academic attitude.

- *Negative autonomous decision:* For each *Bachillerato* academic level  $i = 1, 2$ , students belonging to the promotable groups  $G_i$  or  $B_i$ , may change their personal habit towards the study and this change may lead the students to obtain bad academic results, moving to  $\overline{G}_i$  or  $\overline{B}_i$ . We assume that this transition is proportional to the number of pupils in  $G_i$  and  $B_i$ , and it is modeled by the linear terms  $\alpha_i^G G_i$  and  $\alpha_i^B B_i$ . According to educational experts, it is assumed that the academic attitude is different in the same educational level depending on gender: girls are usually more responsible for their academic performance than boys in both the First and Second Stage of *Bachillerato* [51]. This leads us to suppose the following restrictions:

$$\alpha_1^G < \alpha_1^B, \alpha_2^G < \alpha_2^B. \quad (3.1)$$

In addition we will assume that:

$$\alpha_1^G > \alpha_2^G, \alpha_1^B > \alpha_2^B, \quad (3.2)$$

because students in the Second Stage are more mature than their mates in the First Stage [51].

- *Negative habits transmission:* For each *Bachillerato* academic level  $i = 1, 2$ , students in  $G_i$  or  $B_i$  may move to the non-promotable group,  $\overline{G}_i$  or  $\overline{B}_i$  respectively, due to the negative influence transmitted in the encounters between students (girls and boys) in the non-promotable group in the same academic level. Hence, these transitions are modeled by the non-linear terms  $\beta_i^{G\overline{G}} G_i \overline{G}_i + \beta_i^{G\overline{B}} G_i \overline{B}_i$  and  $\beta_i^{B\overline{G}} B_i \overline{G}_i + \beta_i^{B\overline{B}} B_i \overline{B}_i$ , where  $\beta_i^{G\overline{G}}$ ,  $\beta_i^{G\overline{B}}$ ,  $\beta_i^{B\overline{G}}$  and  $\beta_i^{B\overline{B}}$  are the corresponding transmission rates where the first letter in the superscript denotes the group susceptible to acquire the bad study habit and the second one denotes the group that transmit the bad study habit,  $i = 1, 2$ , at time  $t$ . All specific factors and social encounters involved in the transmission of the bad academic habits are embedded in  $\beta$  parameters.
- *Positive autonomous decision:* For each *Bachillerato* academic level  $i = 1, 2$ , students belonging to the non-promotable group  $\overline{G}_i$  or  $\overline{B}_i$ , may change their personal behavior towards their study habits and this change may lead the students to improve their academic results, moving to  $G_i$  or  $B_i$ . We assume that this transition is proportional to the number of pupils in  $\overline{G}_i$  and  $\overline{B}_i$ , and it is modeled by the linear terms  $\gamma_i^G \overline{G}_i$  and  $\gamma_i^B \overline{B}_i$ .

- *Positive habits transmission:* For each *Bachillerato* academic level  $i = 1, 2$ , students in  $\bar{G}_i$  or  $\bar{B}_i$  may move to the promotable group,  $G_i$  or  $B_i$  respectively, due to the positive influence transmitted in the encounters between students (girls and boys) in the promotable group in the same academic level. Hence, these transitions are modeled by the non-linear terms  $\delta_i^{\bar{G}G}\bar{G}_iG_i + \delta_i^{\bar{G}B}\bar{G}_iB_i$  and  $\delta_i^{\bar{B}G}\bar{B}_iG_i + \delta_i^{\bar{B}B}\bar{B}_iB_i$ , where  $\delta_i^{\bar{G}G}$ ,  $\delta_i^{\bar{G}B}$ ,  $\delta_i^{\bar{B}G}$  and  $\delta_i^{\bar{B}B}$  are the corresponding transmission rates where the first letter in the superscript denotes the group susceptible to acquire the good study habit and the second one denotes the group that transmit the good study habit,  $i = 1, 2$ , at time  $t$ . All specific factors and social encounters involved in the transmission of the good academic habits are embedded in  $\delta$  parameters.
- *Passing courses and graduation:* According to the Spanish educational law, all the students in  $G_1$  and  $B_1$ , in September, transit automatically to  $G_2$  and  $B_2$ , respectively. Analogously, all the students in  $G_2$  and  $B_2$  will graduate in September. These transitions are modeled by  $\varepsilon G_1, \varepsilon G_2, \varepsilon B_1, \varepsilon B_2$ , where

$$\varepsilon = \begin{cases} 1 & \text{if } \frac{9}{12} + j \leq t \leq \frac{10}{12} + j, \\ 0 & \text{otherwise,} \end{cases}$$

where  $j = 0, 1, \dots, 9$  correspond to academic years 1999 – 2000, 2000 – 2001, ..., 2008 – 2009, respectively.  $\varepsilon$  will take value 1 if a *Bachillerato* student (girl/boy) passes into a higher academic level in September and 0 otherwise. As in the previous model shown in Chapter 2, this parameter allows to model the transitions of students who pass successfully in September (ninth month of the year) from First to Second Stage of *Bachillerato*. This parameter also models *Bachillerato* graduation, that is, when students finish the Second Stage of *Bachillerato* (see Figure 3.1).

- *Abandon:* As we said previously, this new formulation of the model considers non-completion of *Bachillerato*. For each *Bachillerato* academic level  $i = 1, 2$ , a proportion of the students in  $\bar{G}_i$  or  $\bar{B}_i$  with bad academic results may leave the studies by autonomous decision. This situation is modeled by the linear terms  $\eta_i^G\bar{G}_i$  and  $\eta_i^B\bar{B}_i$ . We also assume that these transitions are proportional to the number of pupils in  $\bar{G}_i$  and  $\bar{B}_i$ .
- *Access:* New students enter into the *Bachillerato* in the month of September in the promotable group, both girls and boys. It is modeled by the functions

$$\sigma^G = \begin{cases} \tau^G & \text{if } \frac{9}{12} + j \leq t \leq \frac{10}{12} + j, \\ 0 & \text{otherwise,} \end{cases}$$

$$\sigma^B = \begin{cases} \tau^B & \text{if } \frac{9}{12} + j \leq t \leq \frac{10}{12} + j, \\ 0 & \text{otherwise,} \end{cases}$$

where  $j = 0, 1, \dots, 9$  correspond to academic years 1999 – 2000, 2000 – 2001,  $\dots$ , 2008 – 2009, respectively, and  $\tau^G$  and  $\tau^B$  to be determined. This parameters allow us to model student input in the system.

Notice that, in contrast to the model developed in Chapter 2, in this improved model it has considered to include these new assumptions: the access and abandon to the system, the transmission of positive habits and the graduation of *Bachillerato* students. Moreover, we also decompose the transmission parameters into good and bad academic habits in order to analyze with more detail which group of students are more susceptible to be influenced by good or bad academic students.

Thus, under the above assumptions, we build the non-linear system of ordinary differential equations (3.3) using *epiModel* software (see Appendix A). *epiModel* facilitates the implementation all the equations in *Mathematica 8.0* [52] saving developing time. This non-linear system of ordinary differential equations is built in order to describe the dynamics of *Bachillerato* students academic performance, where the unknown functions are  $G_i = G_i(t)$ ,  $B_i = B_i(t)$ ,  $\bar{G}_i = \bar{G}_i(t)$  and  $\bar{B}_i = \bar{B}_i(t)$  ( $t$  denotes time in years). For the sake of clarity, each equation is written in three lines, in the first one the linear terms, in the second one the non-linear terms related to the transmission of bad study habits and in the third line the non-linear terms related to the transmission of good study habits. We point out that *epiModel* software was not applied in the previous mathematical model developed in Chapter 2 since it contains equations which are easier to implement by hand.

$$\begin{aligned}
G_1'(t) &= \sigma^G - \varepsilon G_1(t) - \alpha_1^G G_1(t) + \gamma_1^G \overline{G}_1(t) \\
&\quad - \beta_1^{G\overline{G}} G_1(t) \frac{\overline{G}_1(t)}{T(t)} - \beta_1^{G\overline{B}} G_1(t) \frac{\overline{B}_1(t)}{T(t)} \\
&\quad + \delta_1^{\overline{G}G} \overline{G}_1(t) \frac{G_1(t)}{T(t)} + \delta_1^{\overline{G}B} \overline{G}_1(t) \frac{B_1(t)}{T(t)}, \\
\overline{G}_1'(t) &= \alpha_1^G G_1(t) - \gamma_1^G \overline{G}_1(t) - \eta_1^G \overline{G}_1(t) \\
&\quad + \beta_1^{G\overline{G}} G_1(t) \frac{\overline{G}_1(t)}{T(t)} + \beta_1^{G\overline{B}} G_1(t) \frac{\overline{B}_1(t)}{T(t)} \\
&\quad - \delta_1^{\overline{G}G} \overline{G}_1(t) \frac{G_1(t)}{T(t)} - \delta_1^{\overline{G}B} \overline{G}_1(t) \frac{B_1(t)}{T(t)}, \\
G_2'(t) &= \varepsilon G_1(t) - \varepsilon G_2(t) - \alpha_2^G G_2(t) + \gamma_2^G \overline{G}_2(t) \\
&\quad - \beta_2^{G\overline{G}} G_2(t) \frac{\overline{G}_2(t)}{T(t)} - \beta_2^{G\overline{B}} G_2(t) \frac{\overline{B}_2(t)}{T(t)} \\
&\quad + \delta_2^{\overline{G}G} \overline{G}_2(t) \frac{G_2(t)}{T(t)} + \delta_2^{\overline{G}B} \overline{G}_2(t) \frac{B_2(t)}{T(t)}, \\
\overline{G}_2'(t) &= \alpha_2^G G_2(t) - \gamma_2^G \overline{G}_2(t) - \eta_2^G \overline{G}_2(t) \\
&\quad + \beta_2^{G\overline{G}} G_2(t) \frac{\overline{G}_2(t)}{T(t)} + \beta_2^{G\overline{B}} G_2(t) \frac{\overline{B}_2(t)}{T(t)} \\
&\quad - \delta_2^{\overline{G}G} \overline{G}_2(t) \frac{G_2(t)}{T(t)} - \delta_2^{\overline{G}B} \overline{G}_2(t) \frac{B_2(t)}{T(t)}, \\
B_1'(t) &= \sigma^B - \varepsilon B_1(t) - \alpha_1^B B_1(t) + \gamma_1^B \overline{B}_1(t) \\
&\quad - \beta_1^{B\overline{G}} B_1(t) \frac{\overline{G}_1(t)}{T(t)} - \beta_1^{B\overline{B}} B_1(t) \frac{\overline{B}_1(t)}{T(t)} \\
&\quad + \delta_1^{\overline{B}G} \overline{B}_1(t) \frac{G_1(t)}{T(t)} + \delta_1^{\overline{B}B} \overline{B}_1(t) \frac{B_1(t)}{T(t)}, \\
\overline{B}_1'(t) &= \alpha_1^B B_1(t) - \gamma_1^B \overline{B}_1(t) - \eta_1^B \overline{B}_1(t) \\
&\quad + \beta_1^{B\overline{G}} B_1(t) \frac{\overline{G}_1(t)}{T(t)} + \beta_1^{B\overline{B}} B_1(t) \frac{\overline{B}_1(t)}{T(t)} \\
&\quad - \delta_1^{\overline{B}G} \overline{B}_1(t) \frac{G_1(t)}{T(t)} - \delta_1^{\overline{B}B} \overline{B}_1(t) \frac{B_1(t)}{T(t)}, \\
B_2'(t) &= \varepsilon B_1(t) - \varepsilon B_2(t) - \alpha_2^B B_2(t) + \gamma_2^B \overline{B}_2(t) \\
&\quad - \beta_2^{B\overline{G}} B_2(t) \frac{\overline{G}_2(t)}{T(t)} - \beta_2^{B\overline{B}} B_2(t) \frac{\overline{B}_2(t)}{T(t)} \\
&\quad + \delta_2^{\overline{B}G} \overline{B}_2(t) \frac{G_2(t)}{T(t)} + \delta_2^{\overline{B}B} \overline{B}_2(t) \frac{B_2(t)}{T(t)}, \\
\overline{B}_2'(t) &= \alpha_2^B B_2(t) - \gamma_2^B \overline{B}_2(t) - \eta_2^B \overline{B}_2(t) \\
&\quad + \beta_2^{B\overline{G}} B_2(t) \frac{\overline{G}_2(t)}{T(t)} + \beta_2^{B\overline{B}} B_2(t) \frac{\overline{B}_2(t)}{T(t)} \\
&\quad - \delta_2^{\overline{B}G} \overline{B}_2(t) \frac{G_2(t)}{T(t)} - \delta_2^{\overline{B}B} \overline{B}_2(t) \frac{B_2(t)}{T(t)}. \\
T(t) &= G_1(t) + \overline{G}_1(t) + B_1(t) + \overline{B}_1(t) + G_2(t) + \overline{G}_2(t) + B_2(t) + \overline{B}_2(t)
\end{aligned} \tag{3.3}$$

The flow diagram, associated to the model, is plotted in Figure 3.1.

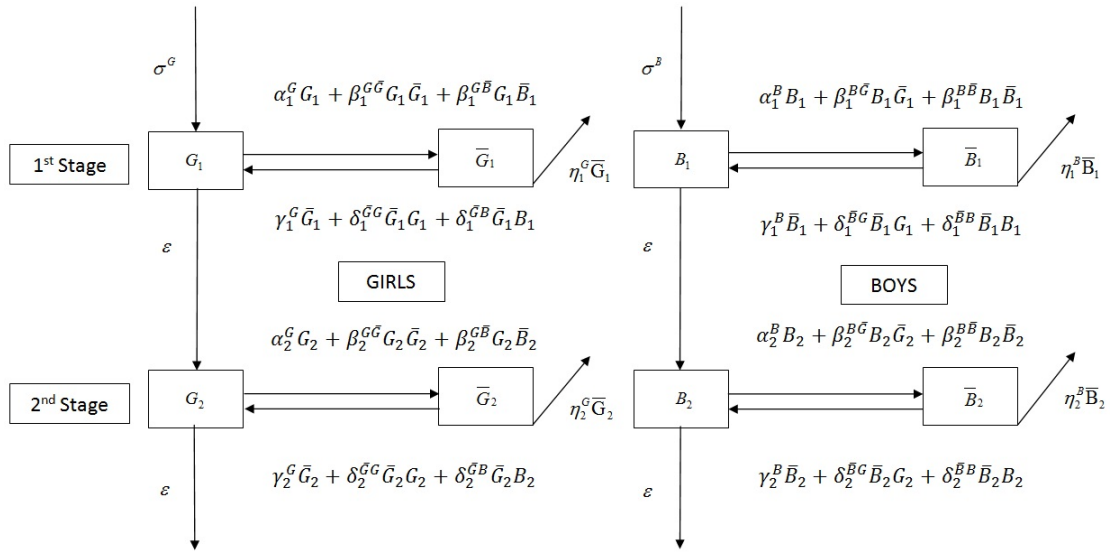


FIGURE 3.1: Flow diagram of the epidemiological-mathematical model for dynamics of Bachillerato academic performance in Spain. The boxes represent the students depending on their sex, stage and academic results. The arrows denote the transits of students labeled by the expressions and parameters governing these transits.

### 3.2.3 Scaling the model

Data in Table 3.1 are related to the percentages of population meanwhile model (3.3) is referred to the number of students where the total population is varying in size over the time. Notice that the *Access* and *Abandon* of students in the system is changing in each academic year. It leads us to transform (by scaling) the model into the same units as data in order to fit the data with the model. Hence, following the ideas developed in [54–56] about how to scale models where the population is varying in size, our model is scaled. The resulting equations are more complex and longer although they contain the same parameters as the non-scaled model (3.3) keeping their meaning. For sake of clarity and, due to the large size of the equations, we can see in detail an example of the development of this process in [57].

Now, in order to avoid introducing new notation corresponding to the obtained scaled model, we are going to keep the same notation to the subpopulation considered previously in our model (Section 3.2.2) corresponding to the percentage of Girls and Boys in the promotable and non-promotable groups, in the First and Second course of *Bachillerato* ( $G_1(t), \bar{G}_1(t), B_1(t), \bar{B}_1(t), G_2(t), \bar{G}_2(t), B_2(t), \bar{B}_2(t)$ ).

Notice that in the model developed in Chapter 2 this process was not needed because, in that case, we considered a mathematical model where the population are constant and there were not inputs and outputs as in this improved mathematical model.

### 3.3 Deterministic parameter estimation and prediction over the next few years

This section is addressed to estimate the parameters of model (3.3). This task has been performed by fitting the scaled model in the mean square sense to the available data collected in Table 3.1.

Once the model built by *epiModel* is scaled as it is indicated in Section 3.2.3, the system of differential equations (3.3) (in its scaled version) is numerically solved by taking as initial conditions the data of the academic year 1999 – 2000 (corresponding to  $t = 0$ ) in Table 3.1. Note that we take the same initial conditions as in the model shown in Chapter 2. Then, we obtain the model parameters that best fit (in the mean square sense) following the same procedure shown in Section 2.3, that is, the parameters that best fit the model using the Nelder-Mead algorithm [53] included in the *Mathematica* command `NMinimize`. Tables 3.2 and 3.3 collect the estimation of the model parameters.

In Table 3.2, we show these values for both, negative and positive autonomous decision and abandon rates. In Table 3.3, we show the corresponding values for negative and positive transmission rates. According to the obtained parameters, we could read into that:

- *Negative and positive autonomous decision ( $\alpha$  and  $\gamma$  parameters)* is higher for girls and boys in the First Stage of *Bachillerato* (see Table 3.2).
- *Abandon rates ( $\eta$  parameter)* are higher for girls belonging in both stages, First and Second Stage of *Bachillerato* (see Table 3.2).
- *Negative transmission of academic habits ( $\beta$  parameter)* is higher from non-promotable boys in First and non-promotable girls and boys in Second Stage

of *Bachillerato* to the promotable girls in the First and Second Stage, respectively. However, the boys are more negatively influence by the girls in case of the First Stage and the boys in case of the Second one (see Table 3.3).

- *Positive transmission of academic habits ( $\delta$  parameter)* is low from girls who promote to non-promotable girls in the Second Stage while it is higher to the non-promotable boys in the same academic level (see Table 3.3).

| Gender | Negative autonomous decision |         | Positive autonomous decision |         | Abandon rates |         |
|--------|------------------------------|---------|------------------------------|---------|---------------|---------|
|        | Parameter                    | Value   | Parameter                    | Value   | Parameter     | Value   |
| Girls  | $\alpha_1^G$                 | 0.04501 | $\gamma_1^G$                 | 0.08685 | $\eta_1^G$    | 0.07480 |
|        | $\alpha_2^G$                 | 0.00366 | $\gamma_2^G$                 | 0.00385 | $\eta_2^G$    | 0.06431 |
| Boys   | $\alpha_1^B$                 | 0.04610 | $\gamma_1^B$                 | 0.11643 | $\eta_1^B$    | 0.02676 |
|        | $\alpha_2^B$                 | 0.01208 | $\gamma_2^B$                 | 0.04163 | $\eta_2^B$    | 0.00232 |

TABLE 3.2: Estimation of positive and negative autonomous decision and abandon rates.

| Gender | Negative transmission |         | Positive transmission      |         |
|--------|-----------------------|---------|----------------------------|---------|
|        | Parameter             | Value   | Parameter                  | Value   |
| Girls  | $\beta_1^{GG}$        | 0.00002 | $\delta_1^{\overline{GG}}$ | 0.03699 |
|        | $\beta_1^{GB}$        | 0.11093 | $\delta_1^{\overline{GB}}$ | 0.09793 |
|        | $\beta_2^{GG}$        | 0.08939 | $\delta_2^{\overline{GG}}$ | 0.00607 |
|        | $\beta_2^{GB}$        | 0.09837 | $\delta_2^{\overline{GB}}$ | 0.06962 |
| Boys   | $\beta_1^{BG}$        | 0.08700 | $\delta_1^{\overline{BG}}$ | 0.01881 |
|        | $\beta_1^{BB}$        | 0.02852 | $\delta_1^{\overline{BB}}$ | 0.04922 |
|        | $\beta_2^{BG}$        | 0.01837 | $\delta_2^{\overline{BG}}$ | 0.11703 |
|        | $\beta_2^{BB}$        | 0.11679 | $\delta_2^{\overline{BB}}$ | 0.07805 |

TABLE 3.3: Estimation of positive and negative transmission parameters.

Notice that, in contrast the model parameters shown in Chapter 2 (see Table 2.2), this improved model includes new parameters which are estimated (Tables 3.2 and 3.3). This makes difficult the comparison between both sets of estimated parameters. However, according to the common parameters in both models related to the autonomous decision of *Bachillerato* students ( $\alpha$  and  $\gamma$  parameters), we can see that the negative and positive autonomous decision parameters, in both



models, seems to follow the same pattern, that is, they are higher in students in the First Stage of *Bachillerato* than in the Second one.

Then, once the model is stated and the parameters estimated, we are able to give predictions of each subpopulation over the next few years by computing the solutions of the model for values of time  $t$  in the forthcoming future. In Table 3.4 we can see the model output for  $t = 1999 - 2000, \dots, 2014 - 2015$ . Note that from 2009 - 2010 the obtained values are the model predictions. It can be compared the model output values for  $t = 1999 - 2000, \dots, 2008 - 2009$  to the data values in Table 3.1 to verify the goodness of the fitting. Graphically, it can be seen in the left-hand side of the graphs in Figure 3.2 (Page 35).

| Academic year | First Stage of <i>Bachillerato</i><br>(Girls   Boys) |  | Second Stage of <i>Bachillerato</i><br>(Girls   Boys) |  |
|---------------|--|--|---|--|
|               | % Promote<br>( $G_1$   $B_1$ )                       | % Non-Promote<br>( $\bar{G}_1$   $\bar{B}_1$ ) | % Promote<br>( $G_2$   $B_2$ )                        | % Non-Promote<br>( $\bar{G}_2$   $\bar{B}_2$ ) |
| 1999 - 2000   | 19.68   15.24  | 9.75   9.33                                    | 16.21   11.64   | 9.52   8.63                                    |
| 2000 - 2001   | 19.94   15.29  | 9.24   9.07                                    | 16.73   12.13   | 9.16   8.44                                    |
| 2001 - 2002   | 20.14   15.34  | 8.80   8.82                                    | 17.18   12.61   | 8.85   8.26                                    |
| 2002 - 2003   | 20.29   15.38  | 8.43   8.56                                    | 17.58   13.07   | 8.60   8.09                                    |
| 2003 - 2004   | 20.39   15.42  | 8.11   8.31                                    | 17.94   13.51   | 8.39   7.93                                    |
| 2004 - 2005   | 20.46   15.46  | 7.84   8.07                                    | 18.25   13.93   | 8.21   7.78                                    |
| 2005 - 2006   | 20.49   15.50  | 7.60   7.83                                    | 18.54   14.35   | 8.06   7.63                                    |
| 2006 - 2007   | 20.49   15.53  | 7.40   7.60                                    | 18.79   14.75   | 7.94   7.50                                    |
| 2007 - 2008   | 20.47   15.57  | 7.22   7.37                                    | 19.02   15.14   | 7.84   7.37                                    |
| 2008 - 2009   | 20.44   15.60  | 7.06   7.15                                    | 19.22   15.51   | 7.76   7.26                                    |
| 2009 - 2010   | 20.38   15.63  | 6.92   6.94                                    | 19.40   15.88   | 7.69   7.15                                    |
| 2010 - 2011   | 20.31   15.67  | 6.80   6.73                                    | 19.57   16.43   | 7.64   7.05                                    |
| 2011 - 2012   | 20.22   15.70  | 6.69   6.53                                    | 19.72   16.58   | 7.60   6.96                                    |
| 2012 - 2013   | 20.13   15.72  | 6.59   6.34                                    | 19.86   16.92   | 7.56   6.87                                    |
| 2013 - 2014   | 20.03   15.75  | 6.50   6.16                                    | 19.99   17.25   | 7.54   6.79                                    |
| 2014 - 2015   | 19.91   15.78  | 6.42   5.98                                    | 20.10   17.57   | 7.52   6.72                                    |

TABLE 3.4: The model output corresponding to the First and Second Stage of *Bachillerato*, in both, state and private high schools all over Spain during academic years 1999 - 2000 to 2014 - 2015. Each row shows the percentage of girls and boys who promote ( $G_i$  |  $B_i$ ) and do not promote ( $\bar{G}_i$  |  $\bar{B}_i$ ) for each level  $i = 1, 2$ . It can be compared the model output values for  $t = 1999 - 2000, \dots, 2008 - 2009$  to the data values in Table 3.1 to verify the goodness of the fitting. Graphically, it can be seen in the left-hand side of the graphs in Figure 3.2.

### 3.4 Introducing uncertainty in the model parameters and predicting the next few years

Uncertainty is a key part of the real world and it should be considered in modeling. Therefore, the assumption that parameters always are constant or the parameter estimation does not contain errors is not appropriate. Thus, it is natural to consider that the model parameters contain uncertainties. Hence, the deterministic prediction can give us an idea about the future trends but the obtained values may not be as accurate as expected.

Thus, we propose forecasting future evolutions using confidence intervals. In order to calculate these confidence intervals, let us use the technique called bootstrapping. Bootstrapping is an efficient method for determining a non-parametric probabilistic estimation of model parameters [58–61]. Specifically, the non-parametric probabilistic estimation of the parameters is performed using a *residual* bootstrapping approach. In order to do it, we are going to follow the next steps:

- Step 1** Compute the error terms for the estimated parameters (deterministic parameters) by the difference between our predictions model in Table 3.4 and their corresponding real data in Table 3.1. We analyze these error terms to find out their probabilistic distribution to resample them using bootstrapping.
- Step 2** Obtain new perturbed data by adding the resampled error (obtained in Step 1) to output of the model collected in Table 3.4 for  $t = 1999 - 2000, \dots, 2008 - 2009$ .
- Step 3** For each new data perturbation calculated (in Step 2), we compute the parameters that best fit the model (in the mean square sense).
- Step 4** For each set of parameter values obtained by fitting the model with the perturbed data, we solve the model with these parameters and compute the outputs in the required time instants.
- Step 5** Taking 95% confidence interval (of each output) from each subpopulation by percentile 2.5 and percentile 97.5 we will be able to conclude the percentage of students who promote/do not promote.

Now, we will show the details of the procedure followed in this section.

### 3.4.1 Error term analysis

In order to analyse the error terms and to obtain their probability distribution, we have followed the next steps:

- We compute the output of the model with the parameters in Tables 3.2 and 3.3 at the time instants  $t = 1999 - 2000, \dots, 2008 - 2009$  and compute their differences (errors) with the corresponding data from Table 3.1. Let us denote these errors by  $e_{G_1}(t)$ ,  $e_{G_2}(t)$ ,  $e_{B_1}(t)$ ,  $e_{B_2}(t)$ ,  $e_{\overline{G}_1}(t)$ ,  $e_{\overline{G}_2}(t)$ ,  $e_{\overline{B}_1}(t)$ ,  $e_{\overline{B}_2}(t)$  for each subpopulation. The results are shown in Table 3.5.

| Academic<br>year  | First Stage of <i>Bachillerato</i><br>(Girls   Boys) |  | Second Stage of <i>Bachillerato</i><br>(Girls   Boys) |  |
|-------------------|--|--|---|--|
|                   | % Promote<br>$e_{G_1}(t)   e_{B_1}(t)$               | % Non-Promote<br>$e_{\overline{G}_1}(t)   e_{\overline{B}_1}(t)$ | % Promote<br>$e_{G_2}(t)   e_{B_2}(t)$                | % Non-Promote<br>$e_{\overline{G}_2}(t)   e_{\overline{B}_2}(t)$ |
| 2000 – 2001 (t=1) | -0.02699   -0.02250                                  | -0.00660   -0.01048  | 0.02657   0.02091                                     | 0.00919   0.01009  |
| 2001 – 2002 (t=2) | 0.00901   0.01107                                    | 0.00194   -0.00293   | -0.00678   -0.00454                                   | -0.00465   -0.00331  |
| 2002 – 2003 (t=3) | 0.01418   0.01191                                    | 0.00072   0.00054  | -0.01556   -0.00907                                   | -0.00160   -0.00112  |
| 2003 – 2004 (t=4) | 0.00462   0.00372                                    | 0.00373   0.00435  | -0.01251   -0.00294                                   | -0.00053   -0.00033  |
| 2004 – 2005 (t=5) | 0.00347   0.00321                                    | 0.00189   0.00131  | -0.00645   0.00013                                    | -0.00181   -0.00175  |
| 2005 – 2006 (t=6) | 0.00419   0.00099                                    | -0.00038   -0.00097  | -0.00604   0.00377                                    | -0.00019   -0.00147  |
| 2006 – 2007 (t=7) | 0.00434   0.00195                                    | -0.00273   -0.00270  | -0.00353   0.00458                                    | -0.00041   -0.00151  |
| 2007 – 2008 (t=8) | 0.00225   -0.00241                                   | -0.00351   -0.00286  | -0.00357   0.00526                                    | 0.00239   0.00255  |
| 2008 – 2009 (t=9) | -0.00284   -0.00968                                  | -0.00217   -0.00276  | -0.00204   0.00653                                    | 0.00707   0.00598  |

TABLE 3.5: Differences between the real data in Table 3.1 and the output model in Table 3.4 corresponding to the First and Second Stage of *Bachillerato*, both state and private high schools all over Spain during academic years 1999 – 2000 to 2008 – 2009, for Girls|Boys who promote ( $G_i|B_i$ ) and do not promote ( $\overline{G}_i | \overline{B}_i$ ) for each level  $i = 1, 2$ .

- We analyse if the error terms in Table 3.5 are correlated. Pearson correlation coefficient was used. The results obtained indicate that the set of all pairs of errors were correlated (see Table 3.6) as all the p-values associated to the coefficients are statistically significant ( $p - values < 0.05$ ).
- Taking into account the Box-Ljung test [62], we also analyze if each error term is autocorrelated. Note that this non-parametric test can be used to check the hypothesis that the elements of a sequence are mutually independent. In our case, none of the test statistic values associated to each error corresponding to each academic year is statistically significant

|   | $e_{G_1}(t)$ | $e_{B_1}(t)$ | $e_{\overline{G}_1}(t)$ | $e_{\overline{B}_1}(t)$ | $e_{G_2}(t)$ | $e_{B_2}(t)$ | $e_{\overline{G}_2}(t)$ | $e_{\overline{B}_2}(t)$ |
|---|--------------|--------------|-------------------------|-------------------------|--------------|--------------|-------------------------|-------------------------|
| $e_{G_1}(t)$ Pearson correlation coef.            | 1            | 0.744        | -0.968                  | -0.851                  | 0.957        | 0.783        | -0.945                  | -0.887                  |
| P-value   |              | 0.022*       | 0.000*                  | 0.004*                  | 0.000*       | 0.012*       | 0.000*                  | 0.001*                  |
| $e_{B_1}(t)$ Pearson correlation coef.            | 0.744        | 1            | -0.811                  | -0.791                  | 0.803        | 0.880        | -0.860                  | -0.775                  |
| P-value   | 0.022*       |              | 0.008*                  | 0.011*                  | 0.009*       | 0.002*       | 0.003*                  | 0.014*                  |
| $e_{\overline{G}_1}(t)$ Pearson correlation coef. | -0.968       | -0.811       | 1                       | 0.771                   | -0.897       | -0.900       | 0.939                   | 0.819                   |
| P-value   | 0.000*       | 0.008*       |                         | 0.015*                  | 0.001*       | 0.001*       | 0.000*                  | 0.007*                  |
| $e_{\overline{B}_1}(t)$ Pearson correlation coef. | -0.851       | -0.791       | 0.771                   | 1                       | -0.951       | -0.661       | 0.851                   | 0.978                   |
| P-value   | 0.004*       | 0.011*       | 0.015*                  |                         | 0.000*       | 0.053*       | 0.004*                  | 0.000*                  |
| $e_{G_2}(t)$ Pearson correlation coef.            | 0.957        | 0.803        | -0.897                  | -0.951                  | 1            | 0.735        | -0.961                  | -0.940                  |
| P-value   | 0.000*       | 0.009*       | 0.001*                  | 0.000*                  |              | 0.024*       | 0.000*                  | 0.000*                  |
| $e_{B_2}(t)$ Pearson correlation coef.            | 0.783        | 0.880        | -0.900                  | -0.661                  | 0.735        | 1            | -0.813                  | -0.710                  |
| P-value   | 0.012*       | 0.002*       | 0.001*                  | 0.053*                  | 0.024*       |              | 0.008*                  | 0.032*                  |
| $e_{\overline{G}_2}(t)$ Pearson correlation coef. | -0.945       | -0.860       | 0.939                   | 0.851                   | -0.961       | -0.813       | 1                       | 0.836                   |
| P-value   | 0.000*       | 0.003*       | 0.000*                  | 0.004*                  | 0.000*       | 0.008*       |                         | 0.005*                  |
| $e_{\overline{B}_2}(t)$ Pearson correlation coef. | -0.887       | -0.775       | 0.819                   | 0.978                   | -0.940       | -0.710       | 0.836                   | 1                       |
| P-value   | 0.001*       | 0.014*       | 0.007*                  | 0.000*                  | 0.000*       | 0.032*       | 0.005*                  |                         |

\* P-value statistically significant to 95%

TABLE 3.6: Matrix of Pearson correlation coefficients for the error terms.

( $p$  - value > 0.05), therefore the claim that there is autocorrelation should be rejected.

- The normality of the distribution of errors is determined by using a non-parametric test. Taking into account that our error terms are correlated, *E-statistic (Energy) Test of Multivariate Normality* is applied [63]. In our case, this test has a p-value equal to 0.9963. Therefore, we can accept that errors  $e_{G_1}(t)$ ,  $e_{B_1}(t)$ ,  $e_{\overline{G}_1}(t)$ ,  $e_{\overline{B}_1}(t)$ ,  $e_{G_2}(t)$ ,  $e_{B_2}(t)$ ,  $e_{\overline{G}_2}(t)$  and  $e_{\overline{B}_2}(t)$  present a multivariate normal distribution. To be precise, we accept that

$$\mathbf{e}(t) = (e_{G_1}(t), e_{\overline{G}_1}(t), e_{B_1}(t), e_{\overline{B}_1}(t), e_{G_2}(t), e_{\overline{G}_2}(t), e_{B_2}(t), e_{\overline{B}_2}(t)),$$

where

$$\mathbf{e}(t) \sim N_8(\mu_{\mathbf{e}}, \Sigma_{\mathbf{e}}), \quad (3.4)$$

being the components of vector  $\mu_{\mathbf{e}}$  the expectations of each component of vector  $\mathbf{e}(t)$  and  $\Sigma_{\mathbf{e}}$  its variance-covariance matrix. These parameters have been estimated using the errors in Table 3.5:

$$\mu_{\mathbf{e}} = (0.0014, -0.0008, -0.0033, 0.0011, -0.0002, -0.0018, 0.0027, 0.0010),$$

$$\Sigma_{\mathbf{e}} = \begin{pmatrix} 1.3464 \cdot 10^{-4} & \cdots & -4.5276 \cdot 10^{-5} \\ 2.8025 \cdot 10^{-5} & \cdots & -1.1065 \cdot 10^{-5} \\ \vdots & \ddots & \vdots \\ -4.52763 \cdot 10^{-5} & \cdots & 1.9350 \cdot 10^{-5} \end{pmatrix}.$$

### 3.4.2 Generating new output perturbed data

The process carried out to generate new output perturbed data has been the following. To our approach, from a computational standpoint, we consider enough to generate 1 000 random error terms following the multivariate normal distribution given by the expression (3.4) assumed by *E-statistic (Energy) Test of Multivariate Normality*. For each one of these 1 000 random error terms:

- We add these error terms (1 000 times) to data in Table 3.4 for  $t = 1999 - 2000, \dots, 2008 - 2009$ , obtaining a new set of perturbed data. Note that we obtained 1 000 sets of perturbed data.
- And we compute the parameters which best fit the model with the set of perturbed data (in the least square sense) and store them, using the same procedure we used to estimate the parameters in Tables 3.2 and 3.3. Note that this procedure allows us to have 1 000 sets of values for the parameters of the model.

### 3.4.3 Obtaining confidence intervals for model outputs

Finally, the confidence intervals are obtained as follows:

- For each one of the 1 000 set of parameters, we solve the system of differential equations (3.3) in order to compute the model output for each subpopulation of students and  $t = 2009 - 2010, \dots, 2014 - 2015$ .
- For each  $t$  and each subpopulation, we have a set of 1 000 model output values. Then, we compute the mean, median and the 95% confidence interval

by percentiles 2.5 and 97.5. These confidence intervals give us the non-parametric probabilistic prediction of the evolution in the next few years. The obtained results can be seen in Table 3.7.

Thus, in Figure 3.2 (Page 35) we can see graphically, for each subpopulation, the data from Table 3.4 (red points), the deterministic model prediction (line), the 95% confidence intervals (error bars). The points in the middle of the confidence intervals are the mean of the 1 000 outputs for each subpopulation at each time instant where we have data about the academic results of Spanish *Bachillerato* students. These mean values are the ones appearing in Table 3.7. Also, we can observe that there is a slight decreasing in the non-promotable groups. However, the sum of the students in non-promotable groups predicted is around 27% similar to the corresponding percentage obtained in the mathematical model developed in Chapter 2 (around 30%). Furthermore, we can see how uncertainty increases in the predictions for the two stages of promoted girls and the First Stage of promoted boys due to the own uncertainty in the process to generate the 1 000 set of parameters to obtain the 1 000 model outputs.

As time goes, new academic results have been published, in particular, the academic results of students of *Bachillerato* in the academic year 2009 – 2010, data that could not be used initially to fit the model because they were not available at that time. These new data allow us to compare the obtained predictions from our model with the new real data. In Figure 3.2 (Page 35), we also include, for each subpopulation, the obtained academic results during the academic year 2009 – 2010 (square black points). These values are the ones appearing in Table 3.8. We can see that probabilistic predictions by 95% confidence intervals either the black points lie inside the confidence intervals or they are close of them making and error in order of, at most,  $10^{-2}$  in absolute terms.

Furthermore, to remark that the predictions in Chapter 2 are also either in the the obtained confidence intervals or near of them. The order of the absolute error is, at most, of  $10^{-2}$  (see Tables B.1 and B.2 in Appendix B).

| Group       | Time (t)    | Mean    | Median  | Confidence interval   |
|-------------|-------------|---------|---------|-----------------------|
| $G_1$       | 2009 – 2010 | 0.20205 | 0.20414 | [ 0.17993 , 0.21227 ] |
| $\bar{G}_1$ | 2009 – 2010 | 0.06851 | 0.06866 | [ 0.06512 , 0.07041 ] |
| $G_2$       | 2009 – 2010 | 0.18859 | 0.19092 | [ 0.16987 , 0.19554 ] |
| $\bar{G}_2$ | 2009 – 2010 | 0.07847 | 0.07875 | [ 0.07564 , 0.08020 ] |
| $B_1$       | 2009 – 2010 | 0.16101 | 0.15807 | [ 0.15286 , 0.18575 ] |
| $\bar{B}_1$ | 2009 – 2010 | 0.06853 | 0.06834 | [ 0.06574 , 0.07340 ] |
| $B_2$       | 2009 – 2010 | 0.16176 | 0.16099 | [ 0.15632 , 0.17254 ] |
| $\bar{B}_2$ | 2009 – 2010 | 0.07100 | 0.07097 | [ 0.06852 , 0.07391 ] |
| $G_1$       | 2010 – 2011 | 0.20126 | 0.20346 | [ 0.17870 , 0.21222 ] |
| $\bar{G}_1$ | 2010 – 2011 | 0.06719 | 0.06734 | [ 0.06333 , 0.06920 ] |
| $G_2$       | 2010 – 2011 | 0.18999 | 0.19241 | [ 0.16938 , 0.19734 ] |
| $\bar{G}_2$ | 2010 – 2011 | 0.07787 | 0.07816 | [ 0.07471 , 0.07969 ] |
| $B_1$       | 2010 – 2011 | 0.16165 | 0.15852 | [ 0.15283 , 0.18840 ] |
| $\bar{B}_1$ | 2010 – 2011 | 0.06646 | 0.06624 | [ 0.06346 , 0.07173 ] |
| $B_2$       | 2010 – 2011 | 0.16557 | 0.16470 | [ 0.15965 , 0.17752 ] |
| $\bar{B}_2$ | 2010 – 2011 | 0.06994 | 0.06991 | [ 0.06728 , 0.07287 ] |
| $G_1$       | 2011 – 2012 | 0.19969 | 0.20220 | [ 0.17673 , 0.21202 ] |
| $\bar{G}_1$ | 2011 – 2012 | 0.06607 | 0.06630 | [ 0.06179 , 0.06830 ] |
| $G_2$       | 2011 – 2012 | 0.19053 | 0.19355 | [ 0.16803 , 0.19898 ] |
| $\bar{G}_2$ | 2011 – 2012 | 0.07744 | 0.07770 | [ 0.07392 , 0.07949 ] |
| $B_1$       | 2011 – 2012 | 0.16306 | 0.15924 | [ 0.15291 , 0.19165 ] |
| $\bar{B}_1$ | 2011 – 2012 | 0.06453 | 0.06430 | [ 0.06137 , 0.07001 ] |
| $B_2$       | 2011 – 2012 | 0.16957 | 0.16870 | [ 0.16287 , 0.18219 ] |
| $\bar{B}_2$ | 2011 – 2012 | 0.06903 | 0.06899 | [ 0.06613 , 0.07221 ] |
| $G_1$       | 2012 – 2013 | 0.19730 | 0.20043 | [ 0.17223 , 0.21172 ] |
| $\bar{G}_1$ | 2012 – 2013 | 0.06510 | 0.06536 | [ 0.06044 , 0.06749 ] |
| $G_2$       | 2012 – 2013 | 0.19021 | 0.19409 | [ 0.16952 , 0.20041 ] |
| $\bar{G}_2$ | 2012 – 2013 | 0.07715 | 0.07760 | [ 0.07322 , 0.07948 ] |
| $B_1$       | 2012 – 2013 | 0.16530 | 0.16126 | [ 0.15361 , 0.19406 ] |
| $\bar{B}_1$ | 2012 – 2013 | 0.06277 | 0.06247 | [ 0.05941 , 0.06848 ] |
| $B_2$       | 2012 – 2013 | 0.17378 | 0.17278 | [ 0.16616 , 0.18630 ] |
| $\bar{B}_2$ | 2012 – 2013 | 0.06826 | 0.06826 | [ 0.06513 , 0.07170 ] |
| $G_1$       | 2013 – 2014 | 0.19497 | 0.19850 | [ 0.17036 , 0.21135 ] |
| $\bar{G}_1$ | 2013 – 2014 | 0.06416 | 0.06444 | [ 0.05934 , 0.06691 ] |
| $G_2$       | 2013 – 2014 | 0.18999 | 0.19112 | [ 0.16989 , 0.20172 ] |
| $\bar{G}_2$ | 2013 – 2014 | 0.07686 | 0.07736 | [ 0.07242 , 0.07916 ] |
| $B_1$       | 2013 – 2014 | 0.16736 | 0.16410 | [ 0.15372 , 0.19705 ] |
| $\bar{B}_1$ | 2013 – 2014 | 0.06109 | 0.06072 | [ 0.05772 , 0.06713 ] |
| $B_2$       | 2013 – 2014 | 0.17783 | 0.17656 | [ 0.16943 , 0.19088 ] |
| $\bar{B}_2$ | 2013 – 2014 | 0.06756 | 0.06760 | [ 0.06426 , 0.07116 ] |
| $G_1$       | 2014 – 2015 | 0.19361 | 0.19730 | [ 0.16837 , 0.21089 ] |
| $\bar{G}_1$ | 2014 – 2015 | 0.06315 | 0.06351 | [ 0.05798 , 0.06610 ] |
| $G_2$       | 2014 – 2015 | 0.19077 | 0.19262 | [ 0.16786 , 0.20304 ] |
| $\bar{G}_2$ | 2014 – 2015 | 0.07645 | 0.07694 | [ 0.07180 , 0.07888 ] |
| $B_1$       | 2014 – 2015 | 0.16821 | 0.16442 | [ 0.15377 , 0.19860 ] |
| $\bar{B}_1$ | 2014 – 2015 | 0.05940 | 0.05898 | [ 0.05586 , 0.06554 ] |
| $B_2$       | 2014 – 2015 | 0.18140 | 0.17994 | [ 0.17236 , 0.19516 ] |
| $\bar{B}_2$ | 2014 – 2015 | 0.06680 | 0.06685 | [ 0.06333 , 0.07060 ] |

TABLE 3.7: The 95% confidence interval prediction corresponding to the First and Second Stage of *Bachillerato*, in both, state and private high schools all over Spain during academic years 2009 – 2010 to 2014 – 2015. Each row shows the rate of girls/boys who promote ( $G_i|B_i$ ) and do not promote ( $\bar{G}_i|\bar{B}_i$ ) for each level  $i = 1, 2$ . Graphically, it can be seen in Figure 3.7.

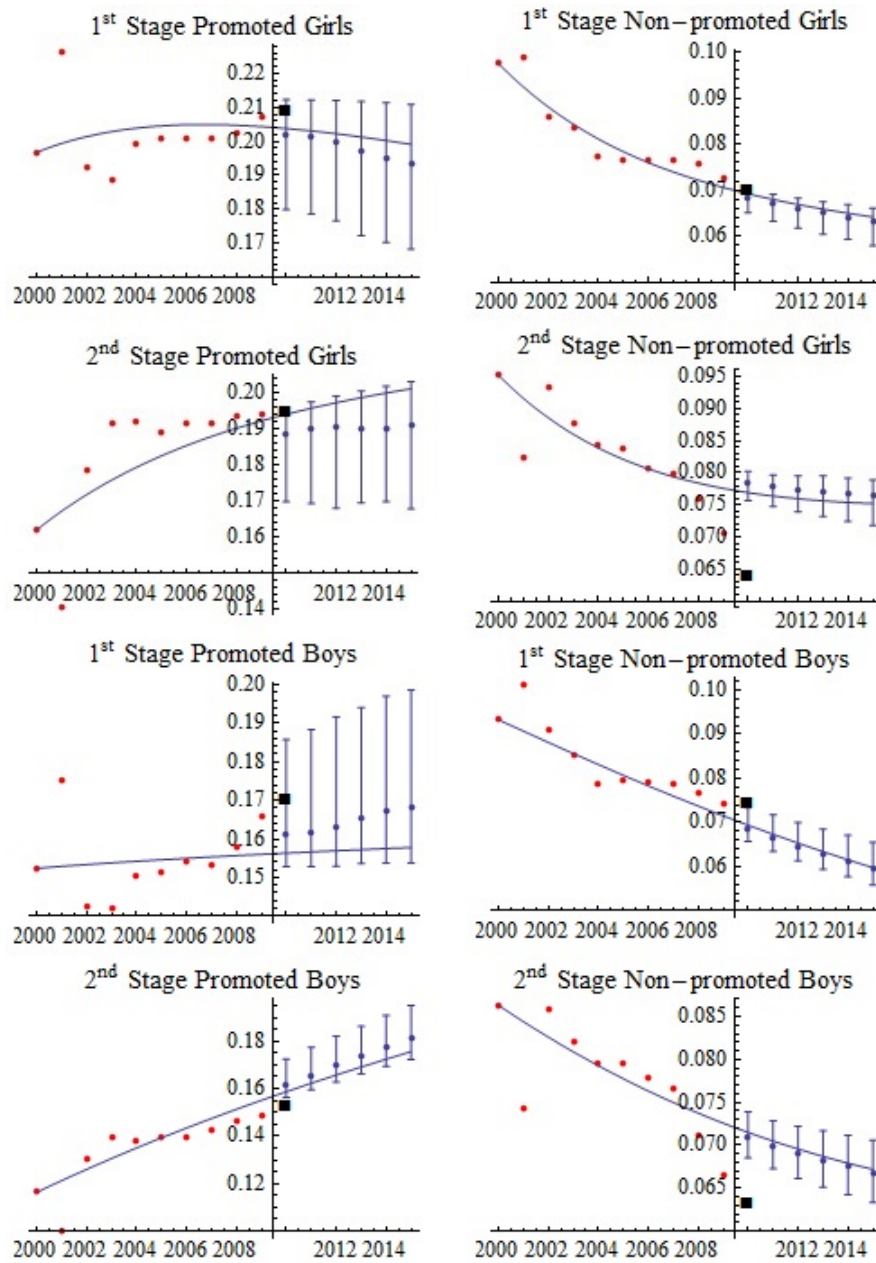


FIGURE 3.2: Real data (red points on the left side of vertical axis) and prediction (line) with confidence intervals (on the right side of vertical axis) of the academic performance of *Bachillerato* Spanish students over the academic years 1999 – 2000 to 2014 – 2015. Smaller confidence intervals, represent less uncertainty in the predictions, the points in the middle of the confidence intervals are their means. The square black point represents the last academic results published recently corresponding to the academic year 2009 – 2010. Notice that each graph has its own scale.



| Group            | Time (t)    | Real data | Confidence Interval<br>Predicted | Absolute<br>Error |
|------------------|-------------|-----------|----------------------------------|-------------------|
| $G_1$            | 2009 – 2010 | 0.20923   | [ 0.17993 , 0.21227 ]            | –                 |
| $\overline{G}_1$ | 2009 – 2010 | 0.07005   | [ 0.06512 , 0.07041 ]            | –                 |
| $G_2$            | 2009 – 2010 | 0.19525   | [ 0.16987 , 0.19554 ]            | –                 |
| $\overline{G}_2$ | 2009 – 2010 | 0.06399   | [ 0.07564 , 0.08020 ]            | 0.01165           |
| $B_1$            | 2009 – 2010 | 0.17037   | [ 0.15286 , 0.18575 ]            | –                 |
| $\overline{B}_1$ | 2009 – 2010 | 0.07485   | [ 0.06574 , 0.07340 ]            | 0.00145           |
| $B_2$            | 2009 – 2010 | 0.15286   | [ 0.15632 , 0.17254 ]            | 0.00346           |
| $\overline{B}_2$ | 2009 – 2010 | 0.06340   | [ 0.06852 , 0.07391 ]            | 0.00512           |

TABLE 3.8: The 95% confidence interval prediction, the real academic results and the distance between these real data from its corresponding confidence interval. The dash indicates that the point lies inside its confidence interval. Data corresponding to the First and Second Stage of *Bachillerato*, in both, state and private high schools all over Spain during the academic year 2009 – 2010. Each row shows the rate of girls/boys who promote ( $G_i|B_i$ ) and do not promote ( $\overline{G}_i|\overline{B}_i$ ) for each level  $i = 1, 2$ .

### 3.5 Abandon analysis

One of the most difficult aspects in academic performance is the study and analysis of the abandon, because there are not much available data and the experts still do not agree with a consensuused definition [64]. In fact, we have made a decision in order to include this issue in the model and this is to consider *abandon* when, during the academic year, the student leaves the academic system. It is an improvement over the model presented in Chapter 2. The student may resume her/his studies in the future, but our model will consider her/him as a new student. Thus, the model allows us to quantify the number of students who leave yearly the system by computing:

$$\int_t^{t+1} (\eta_1^G \overline{G}_1(s) + \eta_2^G \overline{G}_2(s) + \eta_1^B \overline{B}_1(s) + \eta_2^B \overline{B}_2(s)) ds, \quad (3.5)$$

where  $t = 2009 - 2010, \dots, 2013 - 2014$ . The results obtained by 3.5 are collected in Table 3.9.

Moreover, taking advantage of the bootstrapping analysis carried out in Section 3.4, we can predict the evolution of abandon rate in the next few years by means

| Academic year | Total Percentage of abandon |
|---------------|-----------------------------|
| 1999 – 2000   | 1.58                        |
| 2000 – 2001   | 1.51                        |
| 2001 – 2002   | 1.46                        |
| 2002 – 2003   | 1.41                        |
| 2003 – 2004   | 1.37                        |
| 2004 – 2005   | 1.33                        |
| 2005 – 2006   | 1.30                        |
| 2006 – 2007   | 1.27                        |
| 2007 – 2008   | 1.25                        |
| 2008 – 2009   | 1.22                        |
| 2009 – 2010   | 1.21                        |
| 2010 – 2011   | 1.19                        |
| 2011 – 2012   | 1.17                        |
| 2012 – 2013   | 1.16                        |
| 2013 – 2014   | 1.15                        |
| 2014 – 2015   | 1.13                        |

TABLE 3.9: Estimation of the percentage of abandon in Spanish *Bachillerato* during the academic years from 1999 – 2000 to 2014 – 2015.

of 95% confidence intervals. The obtained results are shown in Table 3.10.

| Academic Year   | 2009 – 2010 | 2010 – 2011 | 2011 – 2012 | 2012 – 2013 | 2013 – 2014 | 2014 – 2015 |
|-----------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Mean            | 1.25        | 1.23        | 1.21        | 1.19        | 1.18        | 1.16        |
| Median          | 1.25        | 1.23        | 1.21        | 1.19        | 1.18        | 1.16        |
| Percentile 2.5  | 1.20        | 1.17        | 1.15        | 1.13        | 1.12        | 1.10        |
| Percentile 97.5 | 1.30        | 1.27        | 1.25        | 1.23        | 1.23        | 1.21        |

TABLE 3.10: Descriptive analysis of the percentage of abandon in Spanish *Bachillerato* during the academic years from 2009 – 2010 to 2014 – 2015.

According to the estimated abandon rates given in Tables 3.9 and 3.10, we can see that the evolution of the number of students who leave the high school seems to decrease slightly over the next few years.

## 3.6 Conclusions

Based on the dynamical model developed in Chapter 2 in which we considered, among other factors, the transmission of bad academic habits, in this chapter, we propose an improved dynamical model to study the students' academic performance in high school in Spain. The main idea behind our approach is that academic performance depends on students' autonomous academic behavior and the influence of their classmates with both bad and good academic habits. The abandon is a crucial issue to analyze academic underachievement that has also been considered in this model. To make more realistic our approach, we have included uncertainty in the study. This fact allows us to predict the students' academic performance in the next few years through confidence intervals. In addition, the model presented in this chapter is validated verifying that the 95% confidence intervals predicted either collect or are nearby (with an error, at most, of order  $10^{-2}$ , in absolute terms) of the deterministic estimations provided by the model developed in Chapter 2. Furthermore, the model predictions for 2009–2010 have been compared with the recently available data with acceptable predictions for most of the student groups.

The results inform us that there is a slight decreasing of the percentage of students in the non-promotable groups and who leave the high school. It seems to reach a stationary situation passing, on average, from 28% in the academic year 2010 to 26.5% in 2015. See Tables 3.9 and 3.10 and, graphically, Figure 3.2. The current and predicted scenarios are still worrying as around 27% of the students have bad academic results similar to the percentage obtained in the mathematical model developed in Chapter 2 (around 30%).

Finally, we point out that in Chapter 4, we will take advantage of the ability of the proposed model to be adapted to other foreign academic systems. This will be made by applying the model to the educational system belonging to the German region of North Rhine-Westphalia.



## Chapter 4

# Modelling the dynamics of the students academic performance in the German region of North Rhine-Westphalia: an epidemiological approach with uncertainty

### 4.1 Introduction

In the previous chapter, we have applied the proposed model to the Spanish educational level of *Bachillerato*. In this chapter, we adopt the improved model developed in Chapter 3 to study the dynamics of the students academic performance in the German region of North Rhine-Westphalia. This study has been motivated by my research stay during three months (from 7th April to 29th June, 2012) in Bergische Universität Wuppertal (Germany) working under supervision of Prof. M. Ehrhardt. This study has been possible since the required data were available.

This new approach is supported by the same idea that we stated in the Spanish mathematical model in the Chapter 3, that is, both, good and bad study habits,

are susceptible to be transmitted among students in the same academic level. This model also allows us to forecast the student academic performance by means of confidence intervals over the next few years. Moreover, this application could be of great interest since it allows us to prove that our dynamic model stated in Chapter 3 is not only suitable to the Spanish *Bachillerato* with two educational levels but also to other educational systems, in particular, for the German educational system in North Rhine-Westphalia. This also allows us to compare the academic results in the last courses of high school in both countries. Following the same reasons as we presented for the Spanish academic performance, we apply our mathematical model focusing on the last courses of the high school before accessing the university which, in this case and according to the German educational system [65, 66], are made up of three educational courses (Levels 11, 12 and 13).

## 4.2 Model building

### 4.2.1 Available data

According to the data and in the same way we set out in the Spanish model developed in Chapter 3, we consider that a student *promotes* if, in case the course finishes now, she/he will pass to the next level or graduate satisfying the current legislation into force in North Rhine-Westphalia [65]. Otherwise, this student is in the *non-promote* group. In contrast to the Spanish educational system, the German legislation in North Rhine-Westphalia establishes that the grades are "very good" (1), "good" (2), "satisfactory" (3), "sufficient" (4), "bad" (5) and "very bad" (6). A student in Level 11 and 12 does not promote to the next level if she/he has in 2 or more main subjects (like Maths, Physics, German, English) or in 3 or more minor subjects (like Music, Arts, Sports), a grade of 5 or 6. In case the student is in the last level (Level 13), she/he has to pass all the subjects to obtain the grade [66, 67].

The available data that we have considered in this chapter correspond to the academic results belonging to the students of the last three courses of high schools during the academic years from 2006 – 2007 to 2010 – 2011, in both, state and private high schools all over North Rhine-Westphalia, divided by gender, level and promote/non-promote. The corresponding data can be seen in Table 4.1 [68]. In

| GIRLS    |               | 2006 – 2007 | 2007 – 2008 | 2008 – 2009 | 2009 – 2010 | 2010 – 2011 |
|----------|---------------|-------------|-------------|-------------|-------------|-------------|
| Level 11 | % Promote     | 19.37       | 19.09       | 19.1        | 19.24       | 18.27       |
|          | % Non-Promote | 0.81        | 0.67        | 0.59        | 0.53        | 0.44        |
| Level 12 | % Promote     | 18.23       | 17.96       | 18.15       | 17.77       | 18.29       |
|          | % Non-Promote | 0.75        | 0.68        | 0.58        | 0.47        | 0.47        |
| Level 13 | % Promote     | 15.34       | 15.96       | 15.94       | 16.25       | 16.44       |
|          | % Non-Promote | 0.25        | 0.25        | 0.19        | 0.19        | 0.17        |
| BOYS     |               | 2006 – 2007 | 2007 – 2008 | 2008 – 2009 | 2009 – 2010 | 2010 – 2011 |
| Level 11 | % Promote     | 16.05       | 15.92       | 15.95       | 16.3        | 15.87       |
|          | % Non-Promote | 0.96        | 0.88        | 0.81        | 0.73        | 0.6         |
| Level 12 | % Promote     | 14.7        | 14.73       | 14.77       | 14.72       | 15.21       |
|          | % Non-Promote | 0.85        | 0.81        | 0.67        | 0.67        | 0.64        |
| Level 13 | % Promote     | 12.38       | 12.77       | 13.04       | 12.94       | 13.39       |
|          | % Non-Promote | 0.31        | 0.28        | 0.21        | 0.19        | 0.21        |

TABLE 4.1: The available data corresponding to Levels 11, 12 and 13, in both, state and private high schools all over North Rhine-Westphalia from academic year 2006 – 2007 to 2010 – 2011 divided by gender, level and promote/non-promote over the total number of students in the three levels.

contrast to the Spanish academic results shown in Table 3.1. In particular, from 2006 – 2007 to 2008 – 2009, we notice a remarkable difference in relation to the percentage of students not promoted being much lower in the case of the German students with figures lower than 1%.

## 4.2.2 The type-epidemiological model

As it has been stated previously in our Spanish mathematical model, we fit our mathematical model to the German academic results based on ideas of Christakis and Fowler where individual habits may be transmitted by social contact [29, 30] considering as main idea that the academic performance of a student, Girl (G) or Boy (B), is a mixture of her/his own study habits and the study habits, good or bad, of their classmates. We have also taken into account pedagogical studies [42–45] which confirm that exist a significative difference of academic performance depending on genre and also consider that the transmission of academic habits is carried out between students of the same academic level [23, 46, 47].

The subpopulation of the model will be (time  $t$  in years and  $i = 1$  for Level 11,  $i = 2$  for Level 12 and  $i = 3$  for Level 13):

- $G_i = G_i(t)$  is the number of girls of Level  $i$  who promote at time  $t$ .

- $B_i = B_i(t)$  is the number of boys of Level  $i$  who promote at time  $t$ .
- $\overline{G}_i = \overline{G}_i(t)$  is the number of girls of Level  $i$  who do not promote at time  $t$ .
- $\overline{B}_i = \overline{B}_i(t)$  is the number of boys of Level  $i$  who do not promote at time  $t$ .

To build the German model, we will assume the same hypotheses we have considered in Chapter 3 to formulate the Spanish model. These assumptions are:

- Let us assume a homogeneous population mixing. We follow the same reasoning as we stated in the improved Spanish model (Chapter 3) and also following to other recently published works [30, 32–36, 50] in which it is considered that a particular attitude of any person is not only influenced by her/his autonomous behaviour but also by the social environment surrounded.
- *Negative autonomous decision*: For each academic level,  $i = 1, 2, 3$ , students belonging to the promotable groups  $G_i$  or  $B_i$  may change their personal study habits and, this change may lead them to obtain bad academic results, moving to  $\overline{G}_i$  or  $\overline{B}_i$ . We assume that this transition is proportional to the number of pupils in  $G_i$  and  $B_i$ , and it is modeled by the linear terms  $\alpha_i^G G_i$  and  $\alpha_i^B B_i$ . According to educational experts, as we considered in the Spanish model, it is assumed that the academic attitude is different in the same educational level depending on gender: girls are usually more responsible for their academic performance than boys [51]. This leads us to suppose the following inequality constraints:

$$\alpha_1^G < \alpha_1^B, \alpha_2^G < \alpha_2^B, \alpha_3^G < \alpha_3^B. \quad (4.1)$$

In addition, we will assume that:

$$\alpha_1^G > \alpha_2^G > \alpha_3^G, \alpha_1^B > \alpha_2^B > \alpha_3^B, \quad (4.2)$$

because students in the higher levels are more mature than their mates in the lower levels [51].

- *Negative habits transmission*: For each academic level,  $i = 1, 2, 3$ , students in  $G_i$  or  $B_i$  may move to the non-promotable group,  $\overline{G}_i$  or  $\overline{B}_i$  respectively,



due to the negative influence transmitted by encounters between students (girls and boys) in the non-promotable group in the same academic level. Hence, these transitions are modeled by the non-linear terms  $\beta_i^{G\bar{G}}G_i\bar{G}_i + \beta_i^{G\bar{B}}G_i\bar{B}_i$  and  $\beta_i^{B\bar{G}}B_i\bar{G}_i + \beta_i^{B\bar{B}}B_i\bar{B}_i$ , where  $\beta_i^{G\bar{G}}$ ,  $\beta_i^{G\bar{B}}$ ,  $\beta_i^{B\bar{G}}$  and  $\beta_i^{B\bar{B}}$  are the corresponding transmission rates where the first letter in the superindexes denotes the group susceptible to acquire bad study habits and, the second one denotes the group that transmit bad study habits. All specific factors and social encounters involved in the transmission of the bad academic habits are embedded in  $\beta$  parameters.

- *Positive autonomous decision:* Analogously to *negative autonomous decision*, students belonging to the non-promotable groups may change their personal behavior towards their study habits and, this change may lead the students to improve their academic results, moving to  $G_i$  or  $B_i$ . We assume that this transition is proportional to the number of pupils in  $\bar{G}_i$  and  $\bar{B}_i$ , and it is modeled by the linear terms  $\gamma_i^G\bar{G}_i$  and  $\gamma_i^B\bar{B}_i$ .
- *Positive habits transmission:* Students in non-promotable group may move to the promotable groups due to the positive influence transmitted in the encounters between students (girls and boys) in the promotable group in the same academic level. Hence, these transitions are modeled by the non-linear terms  $\delta_i^{G\bar{G}}\bar{G}_iG_i + \delta_i^{G\bar{B}}\bar{G}_iB_i$  and  $\delta_i^{B\bar{G}}\bar{B}_iG_i + \delta_i^{B\bar{B}}\bar{B}_iB_i$ . The interpretation of the transmission rate parameters is the same as in the *negative habits transmission*.
- *Passing courses and graduation:* The students in  $G_i$  and  $B_i$ , in September, transit automatically to next level  $G_{i+1}$  and  $B_{i+1}$ , respectively, for  $i = 1, 2$ . Students in  $G_3$  and  $B_3$  will graduate in September according to the German educational law in force [65] and it was also defined in the Spanish model. These transitions are modeled by  $\varepsilon_{G_1}$ ,  $\varepsilon_{G_2}$ ,  $\varepsilon_{G_3}$ ,  $\varepsilon_{B_1}$ ,  $\varepsilon_{B_2}$ ,  $\varepsilon_{B_3}$ , where

$$\varepsilon = \begin{cases} 1 & \text{if } \frac{9}{12} + j \leq t \leq \frac{10}{12} + j, \\ 0 & \text{otherwise,} \end{cases}$$

where  $j = 0, 1, 2, 3, 4$ , correspond to the academic years 2006–2007, . . . , 2010–2011, respectively. As in the previous model shown in Chapter 3, this parameter allows to model the transitions of students who pass successfully in

September (ninth month of the year) from Levels: 11 to 12 and 12 to 13, as well as, when the graduation takes place.

- *Abandon*: For each academic level,  $i = 1, 2, 3$ , a proportion of the students in  $\overline{G}_i$  or  $\overline{B}_i$  with bad academic results may leave their studies by autonomous decision. We also assume that these transitions are proportional to the number of pupils in  $\overline{G}_i$  and  $\overline{B}_i$ . This situation is modeled by the linear terms  $\eta_i^G \overline{G}_i$  and  $\eta_i^B \overline{B}_i$ .
- *Access*: New students enter into the Level 11 in the month of September in the promotable groups of girls and boys. It is modeled by the functions

$$\sigma^G = \begin{cases} \tau^G & \text{if } \frac{9}{12} + j \leq t \leq \frac{10}{12} + j, \\ 0 & \text{otherwise,} \end{cases} \quad \sigma^B = \begin{cases} \tau^B & \text{if } \frac{9}{12} + j \leq t \leq \frac{10}{12} + j, \\ 0 & \text{otherwise,} \end{cases}$$

where  $j = 0, 1, 2, 3, 4$ , correspond to the academic years 2006–2007, . . . , 2010–2011, respectively, and  $\tau^G$  and  $\tau^B$  to be determined. As in the improved Spanish model, this parameters allow us to model the incorporation of students into the Level 11 (see Figure 4.1).

Thus, under the above assumptions, we build the non-linear system of ordinary differential equations (4.3)-(4.5) using *epiModel* software (see Appendix A), as it was performed for the Spanish model in Chapter 3. This non-linear system is built in order to describe the dynamics of students academic performance in the German region of North Rhine-Westphalia.

$$\begin{aligned}
G_1'(t) &= \sigma^G - \varepsilon G_1(t) - \alpha_1^G G_1(t) + \gamma_1^G \overline{G}_1(t) \\
&\quad - \left[ \beta_1^{G\overline{G}} G_1(t) \frac{\overline{G}_1(t)}{T(t)} + \beta_1^{G\overline{B}} G_1(t) \frac{\overline{B}_1(t)}{T(t)} \right] + \left[ \delta_1^{\overline{G}G} \overline{G}_1(t) \frac{G_1(t)}{T(t)} + \delta_1^{\overline{G}B} \overline{G}_1(t) \frac{B_1(t)}{T(t)} \right], \\
\overline{G}_1'(t) &= \alpha_1^G G_1(t) - \gamma_1^G \overline{G}_1(t) - \eta_1^G \overline{G}_1(t) \\
&\quad + \left[ \beta_1^{G\overline{G}} G_1(t) \frac{\overline{G}_1(t)}{T(t)} + \beta_1^{G\overline{B}} G_1(t) \frac{\overline{B}_1(t)}{T(t)} \right] - \left[ \delta_1^{\overline{G}G} \overline{G}_1(t) \frac{G_1(t)}{T(t)} + \delta_1^{\overline{G}B} \overline{G}_1(t) \frac{B_1(t)}{T(t)} \right], \\
G_2'(t) &= \varepsilon G_1(t) - \varepsilon G_2(t) - \alpha_2^G G_2(t) + \gamma_2^G \overline{G}_2(t) \\
&\quad - \left[ \beta_2^{G\overline{G}} G_2(t) \frac{\overline{G}_2(t)}{T(t)} + \beta_2^{G\overline{B}} G_2(t) \frac{\overline{B}_2(t)}{T(t)} \right] + \left[ \delta_2^{\overline{G}G} \overline{G}_2(t) \frac{G_2(t)}{T(t)} + \delta_2^{\overline{G}B} \overline{G}_2(t) \frac{B_2(t)}{T(t)} \right], \\
\overline{G}_2'(t) &= \alpha_2^G G_2(t) - \gamma_2^G \overline{G}_2(t) - \eta_2^G \overline{G}_2(t) \\
&\quad + \left[ \beta_2^{G\overline{G}} G_2(t) \frac{\overline{G}_2(t)}{T(t)} + \beta_2^{G\overline{B}} G_2(t) \frac{\overline{B}_2(t)}{T(t)} \right] - \left[ \delta_2^{\overline{G}G} \overline{G}_2(t) \frac{G_2(t)}{T(t)} + \delta_2^{\overline{G}B} \overline{G}_2(t) \frac{B_2(t)}{T(t)} \right], \\
G_3'(t) &= \varepsilon G_2(t) - \varepsilon G_3(t) - \alpha_3^G G_3(t) + \gamma_3^G \overline{G}_3(t) \\
&\quad - \left[ \beta_3^{G\overline{G}} G_3(t) \frac{\overline{G}_3(t)}{T(t)} + \beta_3^{G\overline{B}} G_3(t) \frac{\overline{B}_3(t)}{T(t)} \right] + \left[ \delta_3^{\overline{G}G} \overline{G}_3(t) \frac{G_3(t)}{T(t)} + \delta_3^{\overline{G}B} \overline{G}_3(t) \frac{B_3(t)}{T(t)} \right], \\
\overline{G}_3'(t) &= \alpha_3^G G_3(t) - \gamma_3^G \overline{G}_3(t) - \eta_3^G \overline{G}_3(t) \\
&\quad + \left[ \beta_3^{G\overline{G}} G_3(t) \frac{\overline{G}_3(t)}{T(t)} + \beta_3^{G\overline{B}} G_3(t) \frac{\overline{B}_3(t)}{T(t)} \right] - \left[ \delta_3^{\overline{G}G} \overline{G}_3(t) \frac{G_3(t)}{T(t)} + \delta_3^{\overline{G}B} \overline{G}_3(t) \frac{B_3(t)}{T(t)} \right],
\end{aligned} \tag{4.3}$$

$$\begin{aligned}
B_1'(t) &= \sigma^B - \varepsilon B_1(t) - \alpha_1^B B_1(t) + \gamma_1^B \overline{B}_1(t) \\
&\quad - \left[ \beta_1^{B\overline{G}} B_1(t) \frac{\overline{G}_1(t)}{T(t)} + \beta_1^{B\overline{B}} B_1(t) \frac{\overline{B}_1(t)}{T(t)} \right] + \left[ \delta_1^{\overline{B}G} \overline{B}_1(t) \frac{G_1(t)}{T(t)} + \delta_1^{\overline{B}B} \overline{B}_1(t) \frac{B_1(t)}{T(t)} \right], \\
\overline{B}_1'(t) &= \alpha_1^B B_1(t) - \gamma_1^B \overline{B}_1(t) - \eta_1^B \overline{B}_1(t) \\
&\quad + \left[ \beta_1^{B\overline{G}} B_1(t) \frac{\overline{G}_1(t)}{T(t)} + \beta_1^{B\overline{B}} B_1(t) \frac{\overline{B}_1(t)}{T(t)} \right] - \left[ \delta_1^{\overline{B}G} \overline{B}_1(t) \frac{G_1(t)}{T(t)} + \delta_1^{\overline{B}B} \overline{B}_1(t) \frac{B_1(t)}{T(t)} \right], \\
B_2'(t) &= \varepsilon B_1(t) - \varepsilon B_2(t) - \alpha_2^B B_2(t) + \gamma_2^B \overline{B}_2(t) \\
&\quad - \left[ \beta_2^{B\overline{G}} B_2(t) \frac{\overline{G}_2(t)}{T(t)} + \beta_2^{B\overline{B}} B_2(t) \frac{\overline{B}_2(t)}{T(t)} \right] + \left[ \delta_2^{\overline{B}G} \overline{B}_2(t) \frac{G_2(t)}{T(t)} + \delta_2^{\overline{B}B} \overline{B}_2(t) \frac{B_2(t)}{T(t)} \right], \\
\overline{B}_2'(t) &= \alpha_2^B B_2(t) - \gamma_2^B \overline{B}_2(t) - \eta_2^B \overline{B}_2(t) \\
&\quad + \left[ \beta_2^{B\overline{G}} B_2(t) \frac{\overline{G}_2(t)}{T(t)} + \beta_2^{B\overline{B}} B_2(t) \frac{\overline{B}_2(t)}{T(t)} \right] - \left[ \delta_2^{\overline{B}G} \overline{B}_2(t) \frac{G_2(t)}{T(t)} + \delta_2^{\overline{B}B} \overline{B}_2(t) \frac{B_2(t)}{T(t)} \right], \\
B_3'(t) &= \varepsilon B_2(t) - \varepsilon B_3(t) - \alpha_3^B B_3(t) + \gamma_3^B \overline{B}_3(t) \\
&\quad - \left[ \beta_3^{B\overline{G}} B_3(t) \frac{\overline{G}_3(t)}{T(t)} + \beta_3^{B\overline{B}} B_3(t) \frac{\overline{B}_3(t)}{T(t)} \right] + \left[ \delta_3^{\overline{B}G} \overline{B}_3(t) \frac{G_3(t)}{T(t)} + \delta_3^{\overline{B}B} \overline{B}_3(t) \frac{B_3(t)}{T(t)} \right], \\
\overline{B}_3'(t) &= \alpha_3^B B_3(t) - \gamma_3^B \overline{B}_3(t) - \eta_3^B \overline{B}_3(t) \\
&\quad + \left[ \beta_3^{B\overline{G}} B_3(t) \frac{\overline{G}_3(t)}{T(t)} + \beta_3^{B\overline{B}} B_3(t) \frac{\overline{B}_3(t)}{T(t)} \right] - \left[ \delta_3^{\overline{B}G} \overline{B}_3(t) \frac{G_3(t)}{T(t)} + \delta_3^{\overline{B}B} \overline{B}_3(t) \frac{B_3(t)}{T(t)} \right],
\end{aligned} \tag{4.4}$$

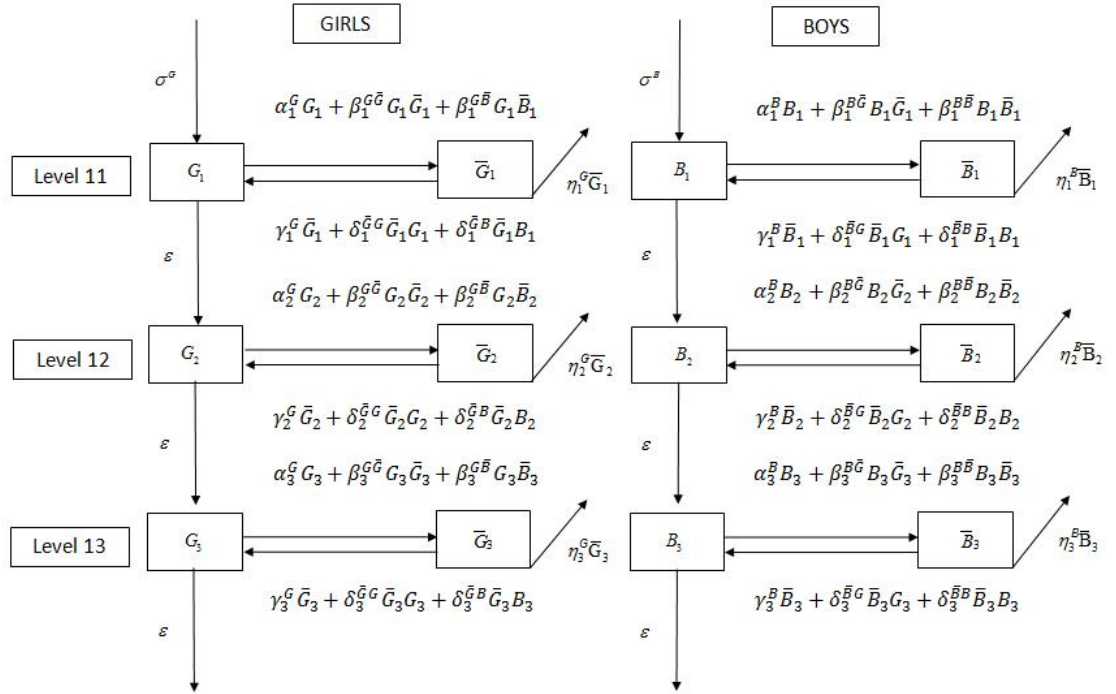


FIGURE 4.1: Flow diagram of the model (4.3)-(4.5). The boxes represent the students depending on their gender, level and academic results. The arrows denote the transit of students labelled by the cause of the flow.

$$\begin{aligned}
 T(t) &= G_1(t) + \bar{G}_1(t) + B_1(t) + \bar{B}_1(t) + G_2(t) + \bar{G}_2(t) + B_2(t) + \bar{B}_2(t) \\
 &+ G_3(t) + \bar{G}_3(t) + B_3(t) + \bar{B}_3(t).
 \end{aligned} \tag{4.5}$$

The flow diagram, associated to the above model, is plotted in Figure 4.1.

### 4.3 Scaling, fitting and predictions

Following the same procedure adopted to scale the Spanish model shown in Section 3.2.3, the German model has also been scaled. In this case, over again, in order to avoid introducing new notation, we will consider that the subpopulations  $G_1(t)$ ,  $\bar{G}_1(t)$ ,  $B_1(t)$ ,  $\bar{B}_1(t)$ ,  $G_2(t)$ ,  $\bar{G}_2(t)$ ,  $B_2(t)$ ,  $\bar{B}_2(t)$ ,  $G_3(t)$ ,  $\bar{G}_3(t)$ ,  $B_3(t)$ ,  $\bar{B}_3(t)$  correspond to the percentage of Girls and Boys in the promotable and non-promotable groups in the Levels 11, 12 and 13, respectively. As we mentioned previously, this process can be entirely seen in [57].

The system of differential equations (4.3)-(4.5), in its scaled version, is numerically solved by taking as initial conditions the data of the academic year 2006 – 2007 (corresponding to  $t = 0$ ). We also compute the model parameters that best fit (in the mean square sense) the scaled model following the same procedure shown in the Spanish model (see Section 2.3). The best estimated model parameters are collected in Tables 4.2 and 4.3.

According to the foregoing definition of the model parameters, we notice that:

- *Negative autonomous decision ( $\alpha$  parameter)* is very low for both girls and boys in the three educational levels (Level 11, 12 and 13) (see Table 4.2).
- *Positive autonomous decision ( $\gamma$  parameter)* is much higher in case of the girls in the Level 12 and boys in the Levels 11 and 12, respectively (see Table 4.2).
- *Abandon rates* seems to be higher for girls in the Levels 11 and 13 and for boys in Levels 11 and 12 (see Table 4.2).
- *Negative transmission ( $\beta$  parameter)*. Promotable girls in the Level 13 are more negatively influenced by non-promotable girls and boys in the same academic level while promotable boys in Level 11 also by the non-promotable girls. Boys in Levels 11 and 12 are also negatively influenced by non-promotable boys in their corresponding levels (see Table 4.3).
- *Positive transmission ( $\delta$  parameter)*. Non-promotable girls in Levels 11 and 12 are more positively influenced by promotable boys and also by promotable girls in case the group of non-promotable girls in Level 12. Besides, non-promotable boys in Levels 11 and 13 are more positively influenced by other promotable boys in their corresponding levels and also by other promotable girls in case of the non-promotable boys in Level 12 (see Table 4.3).

Comparing these parameters against the corresponding ones to the Spanish model (see Tables 3.2 and 3.3 in Section 3.3), some significant differences can be quoted. The rates of negative autonomous decision and the abandon are lower in the case of German model. Whereas, the rates of transmission of the positive and negative academic habits in both countries are varying depending on the different educational levels.

| Gender | Negative autonomous decision |         | Positive autonomous decision |         | Abandon rates |         |
|--------|------------------------------|---------|------------------------------|---------|---------------|---------|
|        | Parameter                    | Value   | Parameter                    | Value   | Parameter     | Value   |
| Girls  | $\alpha_1^G$                 | 0.00060 | $\gamma_1^G$                 | 0.03113 | $\eta_1^G$    | 0.12966 |
|        | $\alpha_2^G$                 | 0.00000 | $\gamma_2^G$                 | 0.14896 | $\eta_2^G$    | 0.00163 |
|        | $\alpha_3^G$                 | 0.00000 | $\gamma_3^G$                 | 0.00061 | $\eta_3^G$    | 0.11675 |
| Boys   | $\alpha_1^B$                 | 0.00590 | $\gamma_1^B$                 | 0.14688 | $\eta_1^B$    | 0.12641 |
|        | $\alpha_2^B$                 | 0.00585 | $\gamma_2^B$                 | 0.00799 | $\eta_2^B$    | 0.14986 |
|        | $\alpha_3^B$                 | 0.00004 | $\gamma_3^B$                 | 0.14933 | $\eta_3^B$    | 0.00691 |

TABLE 4.2: Estimation of positive and negative autonomous decision and abandon rates.

| Gender | Negative transmission |         | Positive transmission |         |
|--------|-----------------------|---------|-----------------------|---------|
|        | Parameter             | Value   | Parameter             | Value   |
| Girls  | $\beta_1^{G\bar{G}}$  | 0.05398 | $\delta_1^{\bar{G}G}$ | 0.06951 |
|        | $\beta_1^{G\bar{B}}$  | 0.08182 | $\delta_1^{\bar{G}B}$ | 0.11225 |
|        | $\beta_2^{G\bar{G}}$  | 0.00093 | $\delta_2^{\bar{G}G}$ | 0.14570 |
|        | $\beta_2^{G\bar{B}}$  | 0.00118 | $\delta_2^{\bar{G}B}$ | 0.14777 |
|        | $\beta_3^{G\bar{G}}$  | 0.14628 | $\delta_3^{\bar{G}G}$ | 0.00016 |
|        | $\beta_3^{G\bar{B}}$  | 0.12087 | $\delta_3^{\bar{G}B}$ | 0.00625 |
| Boys   | $\beta_1^{B\bar{G}}$  | 0.14022 | $\delta_1^{\bar{B}G}$ | 0.05547 |
|        | $\beta_1^{B\bar{B}}$  | 0.12587 | $\delta_1^{\bar{B}B}$ | 0.13882 |
|        | $\beta_2^{B\bar{G}}$  | 0.07844 | $\delta_2^{\bar{B}G}$ | 0.01273 |
|        | $\beta_2^{B\bar{B}}$  | 0.12687 | $\delta_2^{\bar{B}B}$ | 0.01189 |
|        | $\beta_3^{B\bar{G}}$  | 0.00714 | $\delta_3^{\bar{B}G}$ | 0.14485 |
|        | $\beta_3^{B\bar{B}}$  | 0.02304 | $\delta_3^{\bar{B}B}$ | 0.12842 |

TABLE 4.3: Estimation of positive and negative transmission parameters.

## 4.4 Introducing uncertainty in the model parameters and predicting the next few years

As in the previous chapter, in order to calculate these confidence intervals, let us use the technique called bootstrapping. Bootstrapping is an efficient method

for determining a non-parametric probabilistic estimation of model parameters [58, 59], which allow us to obtain predictions with confidence intervals. Specifically, the probabilistic estimation of the parameters is performed using a *residual* bootstrapping approach. In order to do it, we have applied an adaptation of the general procedure presented in [59].

As it was stated in the improved Spanish mathematical model, we are going to follow the next steps:

- Step 1** Compute the error terms for the estimated parameters (deterministic parameters) by the difference between the output of the model with the estimated parameters (deterministic parameters) at the time instants  $t = 2006 - 2007, \dots, 2010 - 2011$  and their corresponding real data collected in Table 4.1. We analyze these error terms to find out their probabilistic distribution to resample them using bootstrapping.
- Step 2** Obtain new perturbed data by adding the resampled error (obtained in Step 1) to output of the model collected in Table 4.1 for  $t = 2006 - 2007, \dots, 2010 - 2011$ , obtaining a new set of perturbed data.
- Step 3** For each new data perturbation calculated (in Step 2), we compute the parameters that best fit the model (in the mean square sense).
- Step 4** For each set of parameter values obtained by fitting the model with the perturbed data, we solve the model with these parameters and compute the outputs in the required time instants.
- Step 5** Taking 95% confidence interval (of each output) from each subpopulation by percentile 2.5 and percentile 97.5 we will be able to conclude the percentage of students who promote/do not promote.

Next, we will show the details of the procedure followed in this section.

#### 4.4.1 Error term analysis

In order to obtain their probability distribution of the error terms (residual terms), we have followed the next steps:

- We compute the output of the model with the estimated parameters (deterministic parameters) at the time instants  $t = 2006 - 2007, \dots, 2010 - 2011$  and compute their differences (errors) with the corresponding data from Table 4.1.
- We analyze if the error terms are correlated. The Pearson correlation coefficient is used. The obtained results from the matrix of Pearson correlation coefficients for the errors terms indicate that none of the test statistic values is statistically significant ( $p - value > 0.05$ ), therefore the set of all pairs of errors were not correlated in contrast to the obtained Spanish ones which are correlated.
- Taking into account the Box-Ljung test [62], we also analyze if each error term is autocorrelated in order to find out if there is correlation between error of the process at different times. The obtained results allow us to accept that the error term corresponding to the Level 13 - Promoted Boys is statistically significant ( $p - value = 0.027$ ), therefore there is autocorrelation. However, the rest of the test statistic values are not ( $p - value > 0.05$ ), that is, there is not autocorrelation in any of them.

Notice that starting from here, due to the error corresponding to the Level 13 - Promoted Boys is autocorrelated, we will try to get its probability distribution using an autoregressive (AR) time series model, not in case of the others errors that are not autocorrelated and we will try to get them by statistical tests, as we will see in Section 4.4.2.

- For all the non-autocorrelated error terms, initially, the normality of the distribution of errors is checked by the Shapiro-Wilk Normality test [64]. We have obtained the p-values corresponding to each error term and they are not statistically significant ( $p - value > 0.05$ ), except for the error of the Level 11 - Promoted Girls, whose p-value is 0.034. Therefore, we can accept that all the errors present a univariate normal distribution excluding the error corresponding to the Level 11 - Promoted Girls.



### 4.4.2 Generating new output perturbed data

In the previous subsection, we have obtained the probability distribution of each error terms (residual terms). Then, in this subsection, we generate the new perturbed output. To our approach and taking into account the limited data available (five academic years), in this case, we obtain 10 000 random error terms following different processes according to the statistical properties of each error term:

- For all the error terms, except the ones corresponding to the Level 11 - Promoted Girls and Level 13 - Promoted Boys, we sample 10 000 random error terms following the univariate normal distribution with their means and variances, respectively, obtained from the error terms.

For the autocorrelated error term corresponding to the Level 13 - Promoted Boys, we sample 10 000 random error terms using autoregressive techniques [69]. This has been carried out by fitting an autoregressive (AR) time series model to the data [70]. In this case, the obtained autoregressive function, AR(1), whose coefficient has been estimated by the *The R Project for Statistical Computing* [71] using the *Stats package*, it is given by:

$$e_t = -0.7833397e_{t-1} + r_t, \quad (4.6)$$

where  $e_t$  is the obtained error and  $r_t$  is the white noise at times  $t = 2006 - 2007, \dots, 2010 - 2011$ .

Once we have obtained the autoregressive function, AR(1), we need to analyze if the white noise generated satisfies the necessary statistical requirements, that is, white noise must be a random process of random variables that are uncorrelated, have zero mean, and a finite variance [69, 70]. Often one assumes a normal distribution for it, in which case the distribution is completely specified by the mean and variance. In our case, we check the white noise ( $r_t$ ) obtained is uncorrelated using the non-parametric Box-Ljung test [62] which indicates, with  $p$ -value = 0.9855, that autocorrelation should be rejected. We also check its probability distribution, testing previously the univariate normal distribution, using for that the Shapiro-Wilk Normality Test with  $p$ -value = 0.8088 which confirms that the white noise terms follow an univariate normal distribution with  $\mu = 0$ ,  $\sigma = 0.000367$ .

Now, we are in conditions to generate a set of 10 000 white noises. We add these generated white noises,  $r_t$ , (10 000 times) to the expression (4.6) obtaining a set of 10 000 error terms.

For sake of clarity, we point out that, until now, we have generated a set of 10 000 errors which have been obtained using different statistical techniques. Firstly, to the autocorrelated error (the corresponding one to the Level 13 - Promoted Boys) is treated by using an autoregressive (AR) time series model. Secondly, the rest of the errors have been tested by applying the Shapiro-Wilk Normality tests for each one which have confirmed that they follow a univariate normal distribution, except the error corresponding to the Level 11 - Promoted Girls. Finally, this error (corresponding to the Level 11 - Promoted Girls) has been obtained after assuming that the total sum of the errors of each instant  $t$  is 0.

- To conclude, we compute the parameters which best fit (in the mean square sense) the model with the set of perturbed data and store them, using the same procedure we used to estimate the obtained parameters in Section 4.3. Note that this procedure allows us to have 10 000 sets of values for the model parameters.

### 4.4.3 Obtaining confidence intervals for model outputs

Finally, the confidence intervals are obtained as follows:

- For each one of the 10 000 set of parameters, we solve the scaled system of differential equations in order to compute the model output for each subpopulation of students and  $t = 2011 - 2012, \dots, 2014 - 2015$ . Once the models are solved, we select the set of parameters which the resulting mean square error value is, at most, 5% greater than the best fit obtained of the model in Section 4.3. This percentage has been selected by convenience in order to remove those set of parameters that do not provide a good fit and, therefore, obtain a best estimate of the confidence intervals. Moreover, it gives us an acceptable number of set parameters to generate model output perturbed that, in this case, has been reduced to 1 000.

- For each  $t$  and each subpopulation, we have a set of 1 000 model output values. Then, we compute the mean, median and the 95% confidence interval by percentiles 2.5 and 97.5. These confidence intervals give us the non-parametric probabilistic prediction of the evolution in the next few years. The obtained results can be seen in Table 4.4.

Thus, in Figure 4.2 we can see graphically, for each subpopulation, the real data from Table 4.1 (black points) and the 95% confidence intervals (red lines). The dashed line in the middle of the confidence intervals represents the mean of the 1 000 outputs for each subpopulation of German students at each time. These mean values are the ones obtained from our model and these predicted values from the academic year 2011 – 2012 to 2014 – 2015 are collected in Table 4.4.

We can see that our predictions draw different tendencies in the plots in each subpopulation. We notice that there is also a slight decreasing in the non-promotable groups, in both, Girls and Boys, as occurs in the predictions obtained in the Spanish model shown in Chapter 3 although with the great difference that in the Spanish model seems to lie at worries rates (around 27%) while in the German one is not higher than 1%.

In both, Table 4.4 and, graphically, in Figure 4.2, we can observe that, although not all the points lie inside the red bands (95% confidence intervals), they are close to them taking into account the small scale of these graphs. This can also be seen in Table C.1 in Appendix C where the higher errors correspond to the promotable groups with associated error, at most, of the order  $10^{-2}$  (in absolute terms).

Note that there are high differences in the scale of the graphs between the promotable and non-promotable students, specially with very small values in the non-promotable groups.

| Level    | Groups             | Time ( $t$ ) | Mean    | Median  | Confidence Interval   |
|----------|--------------------|--------------|---------|---------|-----------------------|
| Level 11 | Promoted Girls     | 2012         | 0.18836 | 0.19160 | [ 0.18093 , 0.19173 ] |
|          | Non-Promoted Girls | 2012         | 0.00393 | 0.00393 | [ 0.00391 , 0.00411 ] |
|          | Promoted Boys      | 2012         | 0.16100 | 0.16005 | [ 0.15997 , 0.16325 ] |
|          | Non-Promoted Boys  | 2012         | 0.00831 | 0.00830 | [ 0.00827 , 0.00845 ] |
| Level 12 | Promoted Girls     | 2012         | 0.18039 | 0.18167 | [ 0.17719 , 0.18180 ] |
|          | Non-Promoted Girls | 2012         | 0.00505 | 0.00505 | [ 0.00501 , 0.00522 ] |
|          | Promoted Boys      | 2012         | 0.15026 | 0.14959 | [ 0.14955 , 0.15198 ] |
|          | Non-Promoted Boys  | 2012         | 0.00549 | 0.00549 | [ 0.00547 , 0.00554 ] |
| Level 13 | Promoted Girls     | 2012         | 0.16228 | 0.16089 | [ 0.16080 , 0.16517 ] |
|          | Non-Promoted Girls | 2012         | 0.00096 | 0.00096 | [ 0.00096 , 0.00109 ] |
|          | Promoted Boys      | 2012         | 0.13258 | 0.13094 | [ 0.13078 , 0.13647 ] |
|          | Non-Promoted Boys  | 2012         | 0.00127 | 0.00125 | [ 0.00125 , 0.00146 ] |
| Level 11 | Promoted Girls     | 2013         | 0.18843 | 0.18674 | [ 0.17693 , 0.19230 ] |
|          | Non-Promoted Girls | 2013         | 0.00346 | 0.00345 | [ 0.00343 , 0.00366 ] |
|          | Promoted Boys      | 2013         | 0.16061 | 0.16103 | [ 0.15942 , 0.16385 ] |
|          | Non-Promoted Boys  | 2013         | 0.00813 | 0.00812 | [ 0.00807 , 0.00828 ] |
| Level 12 | Promoted Girls     | 2013         | 0.18034 | 0.17964 | [ 0.17564 , 0.18196 ] |
|          | Non-Promoted Girls | 2013         | 0.00482 | 0.00482 | [ 0.00477 , 0.00501 ] |
|          | Promoted Boys      | 2013         | 0.15069 | 0.15091 | [ 0.14986 , 0.15356 ] |
|          | Non-Promoted Boys  | 2013         | 0.00503 | 0.00503 | [ 0.00500 , 0.00508 ] |
| Level 13 | Promoted Girls     | 2013         | 0.16312 | 0.16393 | [ 0.16144 , 0.16719 ] |
|          | Non-Promoted Girls | 2013         | 0.00079 | 0.00079 | [ 0.00079 , 0.00094 ] |
|          | Promoted Boys      | 2013         | 0.13340 | 0.13400 | [ 0.13139 , 0.13927 ] |
|          | Non-Promoted Boys  | 2013         | 0.00106 | 0.00104 | [ 0.00104 , 0.00128 ] |
| Level 11 | Promoted Girls     | 2014         | 0.18864 | 0.18724 | [ 0.17736 , 0.19283 ] |
|          | Non-Promoted Girls | 2014         | 0.00306 | 0.00305 | [ 0.00303 , 0.00327 ] |
|          | Promoted Boys      | 2014         | 0.16014 | 0.16047 | [ 0.15887 , 0.16341 ] |
|          | Non-Promoted Boys  | 2014         | 0.00796 | 0.00795 | [ 0.00790 , 0.00812 ] |
| Level 12 | Promoted Girls     | 2014         | 0.18032 | 0.17975 | [ 0.17574 , 0.18206 ] |
|          | Non-Promoted Girls | 2014         | 0.00463 | 0.00464 | [ 0.00457 , 0.00484 ] |
|          | Promoted Boys      | 2014         | 0.15107 | 0.15122 | [ 0.15017 , 0.15396 ] |
|          | Non-Promoted Boys  | 2014         | 0.00460 | 0.00460 | [ 0.00457 , 0.00466 ] |
| Level 13 | Promoted Girls     | 2014         | 0.16382 | 0.16454 | [ 0.16204 , 0.16780 ] |
|          | Non-Promoted Girls | 2014         | 0.00066 | 0.00065 | [ 0.00065 , 0.00082 ] |
|          | Promoted Boys      | 2014         | 0.13408 | 0.13455 | [ 0.13192 , 0.13991 ] |
|          | Non-Promoted Boys  | 2014         | 0.00089 | 0.00087 | [ 0.00087 , 0.00113 ] |
| Level 11 | Promoted Girls     | 2015         | 0.18877 | 0.18771 | [ 0.17768 , 0.19330 ] |
|          | Non-Promoted Girls | 2015         | 0.00272 | 0.00272 | [ 0.00269 , 0.00295 ] |
|          | Promoted Boys      | 2015         | 0.15967 | 0.15989 | [ 0.15828 , 0.16322 ] |
|          | Non-Promoted Boys  | 2015         | 0.00781 | 0.00780 | [ 0.00774 , 0.00799 ] |
| Level 12 | Promoted Girls     | 2015         | 0.18025 | 0.17981 | [ 0.17576 , 0.18213 ] |
|          | Non-Promoted Girls | 2015         | 0.00448 | 0.00449 | [ 0.00442 , 0.00470 ] |
|          | Promoted Boys      | 2015         | 0.15145 | 0.15151 | [ 0.15046 , 0.15445 ] |
|          | Non-Promoted Boys  | 2015         | 0.00420 | 0.00420 | [ 0.00417 , 0.00427 ] |
| Level 13 | Promoted Girls     | 2015         | 0.16449 | 0.16509 | [ 0.16261 , 0.16837 ] |
|          | Non-Promoted Girls | 2015         | 0.00054 | 0.00054 | [ 0.00054 , 0.00071 ] |
|          | Promoted Boys      | 2015         | 0.13473 | 0.13505 | [ 0.13240 , 0.14050 ] |
|          | Non-Promoted Boys  | 2015         | 0.00075 | 0.00073 | [ 0.00073 , 0.00100 ] |

TABLE 4.4: The 95% confidence interval predictions corresponding to the Levels 11, 12 and 13, in both, state and private high schools all over the German region of North Rhine-Westphalia during academic years 2011 – 2012 to 2014 – 2015. Each row shows the percentage of girls/boys who promote and do not promote for each academic level.

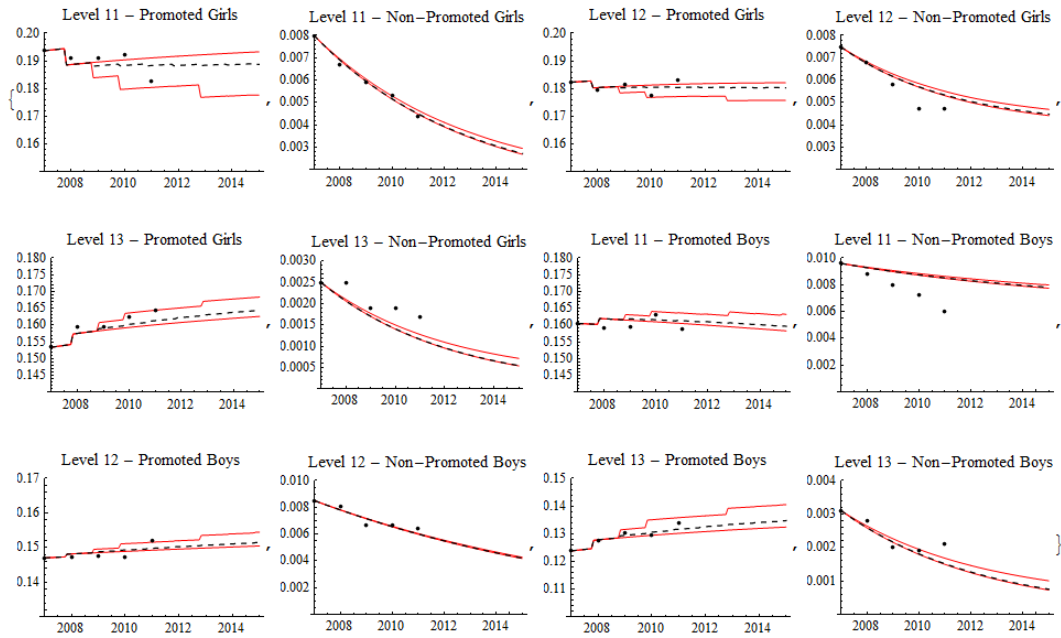


FIGURE 4.2: Real data (black points) and prediction with 95% confidence intervals (red line) of the academic performance of German students in the North Rhine-Westphalia over the academic years 2006 – 2007 to 2014 – 2015. Smaller confidence intervals, represent less uncertainty in the predictions, the dashed lines in the middle of the confidence intervals are their means. Note that there are high differences in the scale of the graphs between the promotable and non-promotable students, specially with very low rates in the non-promotable groups.

## 4.5 Conclusions

In this chapter, we have presented an application of the Spanish mathematical model developed in Chapter 3 in order to study the dynamics of the students academic performance in the German region of North Rhine-Westphalia. In this model, we have divided the students by gender and academic levels, and it is based on the assumption that both, good and bad study habits, are a mixture of personal decisions and influence on classmates. Using the academic results of German students, we have estimated the model parameters fitting the model with the data. Thus, we can predict with confidence intervals the student's academic performance in the next few years. From Figure 4.2, it is expected that the decreasing trend in all non-promotable groups remains in the next years, as occurs in the predictions obtained in the Spanish model shown in Chapter 3. However, there is a significant difference between the academic underachievement performance of both academic systems, namely, the Spanish model seems to be stabilizing at worry levels (around 27%) whereas in the German system this value is much lower. For instance, in the course 2014 – 2015 less than 2% of the students will not promote (see Table 4.4).

As it has been shown in previous chapters, we have developed a mathematical model which has been applied to both the Spanish and German educational system. Besides, this has also allowed us to confirm the alarming situation that the Spanish education system will have to face considering the predictions of bad academic results obtained from our model (see Tables 3.8 and 3.10 in Chapter 3), with average rates of academic underachievement around 27%. This motivates the study to be presented in next Chapter 5, where we will quantify the cost of these high rates of academic underachievement to be faced to both the Spanish Government and Spanish families.

# Chapter 5

## Estimation of the cost of the academic underachievement in high school in Spain over the next few years

### 5.1 Introduction

In Chapter 3, we have developed our improved Spanish mathematical model that predicts the academic results of Spanish *Bachillerato* over the next few years. The predictions are given in 95% confidence intervals. In this chapter, we will take advantage of these predictions, in particular, of the high rates of academic underachievement (around 27%) in order to quantify the economical cost that will have to support both, the Spanish Government and the Spanish families. The estimations will be performed separately for the Spanish Government (Section 5.2) and Spanish families (Section 5.3) by 95% confidence intervals for the next few years. Finally, conclusions are given in Section 5.4.

## 5.2 Estimation with 95% confidence intervals of the cost of the academic underachievement in *Bachillerato* for the next few years for the Spanish Government

In this section, as we said previously, we will pay special attention on the predictions of the percentage of Spanish *Bachillerato* students who may abandon or not promote in the next few years (see Tables 3.7 and 3.10 in Chapter 3). These predictions will allow us, with the required suitable economical data, to predict the cost to the academic underachievement in this educational level for the Spanish Government.

To perform estimations as close as possible, we will follow the next steps:

**Step 1** We obtain the average Spanish Government cost of each *Bachillerato* student during the academic years 1999 – 2000 to 2008 – 2009.

**Step 2** We predict the Spanish Government investment in each Spanish *Bachillerato* student during the academic years 2009 – 2010 to 2014 – 2015 using the cost of each *Bachillerato* student given in Step 1.

**Step 3** We predict the number of *Bachillerato* students registered during the academic years 2009 – 2010 to 2014 – 2015. This is required to obtain the number of *Bachillerato* students that will not promote and abandon at that period using the corresponding percentages estimated in Tables 3.7 and 3.10 in Chapter 3.

**Step 4** We compute the total of the Spanish Government investment in *Bachillerato* students that will not promote and abandon during the academic years 2009 – 2010 to 2014 – 2015 using the predictions given in Step 2 and 3.

First, we obtain the Spanish Government cost of each *Bachillerato* student during the academic years 1999 – 2000 to 2008 – 2009 (**Step 1**). For that, we collect the total investment in education (in euros) and calculate the percentage of the Spanish Government investment amount of money expended in *Bachillerato* educational level, in both, state and private high schools all over Spain from academic



| $t$ | Academic Year | Euros    |
|-----|---------------|----------|
| 1   | 1999 – 2000   | 2 610,70 |
| 2   | 2000 – 2001   | 2 796,50 |
| 3   | 2001 – 2002   | 2 991,48 |
| 4   | 2002 – 2003   | 3 384,28 |
| 5   | 2003 – 2004   | 3 691,93 |
| 6   | 2004 – 2005   | 3 972,37 |
| 7   | 2005 – 2006   | 4 224,20 |
| 8   | 2006 – 2007   | 4 569,65 |
| 9   | 2007 – 2008   | 5 130,38 |
| 10  | 2008 – 2009   | 5 146,88 |

TABLE 5.1: Investment per Spanish student in the First and Second Stage of *Bachillerato*, in both, state and private high schools all over Spain from academic year 1999–2000 to 2008–2009 by the Government [1].

year 1999 – 2000 to 2008 – 2009 [1]. These available data let us know the total the Spanish Government investment in *Bachillerato* at that period of time. Furthermore, we also know the number of students registered during the mentioned period given in [2].

The aforementioned data allow us to work out (dividing the Spanish Government investment in *Bachillerato* by the number of *Bachillerato* students registered in each academic year) the amount of money in euros that the Spanish Government has invested in each *Bachillerato* student in recent years. The obtained results can be seen in Table 5.1. Notice that these figures have progressively increased as time goes on ranging between 2 610,70 euros in 1999 – 2000 until 5 146,88 euros in 2008 – 2009.

Then, we need to predict the Spanish Government investment in each *Bachillerato* student during the academic years 2009 – 2010, ..., 2014 – 2015 (**Step 2**). To do that, we are going to use statistical techniques, in particular, time series analysis [70, 72, 73]. This statistical technique provides tools for selecting a model in order to forecast future events. In our case, the application of these techniques will return predictions of the investment in each *Bachillerato* student over the next

| Model   | RMSE    | MAPE    |
|---|---------|---------|
| Random walk with trend                                | 151.187 | 2.56029 |
| Linear trend  | 104.440 | 1.78572 |
| Simple moving average of 3 terms                      | 633.230 | 14.7076 |
| Simple exponential smoothing with alpha 0.999         | 315.820 | 6.49393 |
| Brown's Linear Exponential Smoothing with alpha 0.853 | 222.497 | 4.35632 |

TABLE 5.2: The indicators (RMSE and MAPE) considered for the validation of the different models in order to determine the model that best fit the data in Table 5.1. The best is the *Linear Trend Model*.

few years taking into account the known Spanish Government investment the previous years (Table 5.1). We will address our approach using *Statgraphics Plus for Windows 5.1* software [74]. This powerful statistical tool provides the user five different forecasting models: Random Walk with Trend, Linear Trend, Simple Moving Average, Simple Exponential Smoothing and Brown's Linear Exponential Smoothing. Then, the models are validated by their corresponding Root Mean Square Error (RMSE) and Percentage of the Mean Absolute Error (MAPE). Finally, it is selected the model that best fit the available data and provide us the predictions with 95% confidence intervals, both analytically and graphically. The model that best fit our data is the *Linear Trend Model* because it returns the minimum Root Mean Square Error (RMSE=104.44) whose corresponding Mean Absolute Percentage of Error is 1.79, as can be seen in Table 5.2 (see Section D.2 and D.3, in Appendix D). Therefore, the obtained equation which allows us to predict the Spanish Government investment in euros in each *Bachillerato* student over the next few years is stated as follows:

$$G_t = -601795.0 + 302.144t, \quad (5.1)$$

where  $G_t$  is the estimation of the investment at time  $t = 1, 2, 3, \dots$  where  $t = 1$  corresponds to the academic year 1999 – 2000,  $t = 2$  to the academic year 2000 – 2001 and so on.

According to the time series model stated, in Table 5.3, we show the obtained estimations with 95% confidence intervals given by *Statgraphics Plus for Windows 5.1* (See Section D.4, Appendix D) of the cost in euros that the Spanish Government would invest in each *Bachillerato* student during the academic years

from 2009–2010 to 2014–2015. Graphically, this results can be seen in Figure 5.1.

| $t$ | Academic Year | Prediction (Euros) | 95% Confidence interval (Euros) |
|-----|---------------|--------------------|---------------------------------|
| 11  | 2009 – 2010   | 5 513,62           | [5 221,95 , 5 805,29]           |
| 12  | 2010 – 2011   | 5 815,77           | [5 509,97 , 6 121,56]           |
| 13  | 2011 – 2012   | 6 117,91           | [5 796,43 , 6 439,39]           |
| 14  | 2012 – 2013   | 6 420,05           | [6 081,52 , 6 758,58]           |
| 15  | 2013 – 2014   | 6 722,20           | [6 365,47 , 7 078,92]           |
| 16  | 2014 – 2015   | 7 024,34           | [6 648,42 , 7 400,26]           |

TABLE 5.3: The prediction of euros invested by the Spanish Government in each Spanish student in the First and Second Stage of *Bachillerato*, in both, state and private high schools during the academic years from 2009 – 2010 to 2014 – 2015.

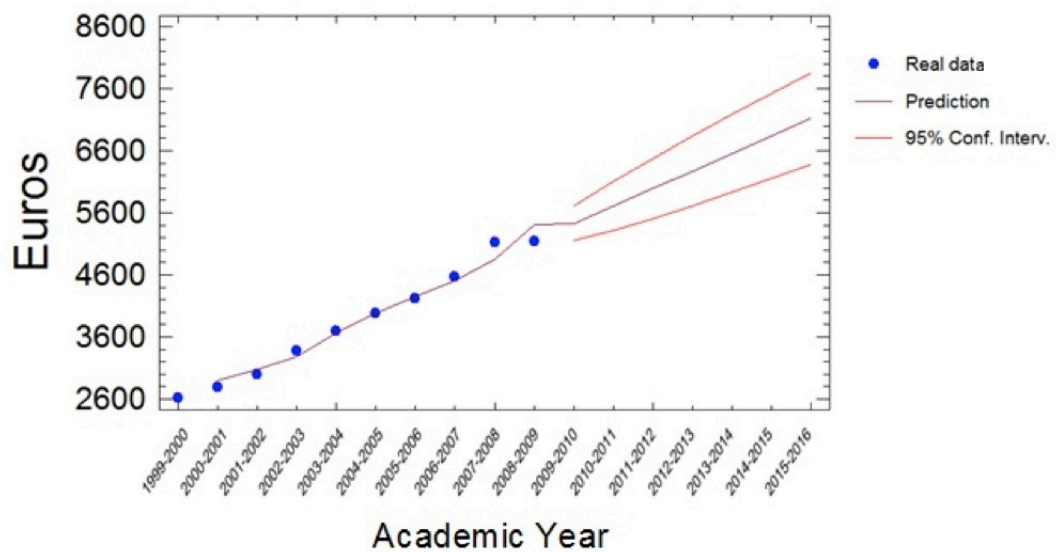


FIGURE 5.1: Graph of the prediction of euros invested by the Spanish Government in each Spanish student in the First and Second Stage of *Bachillerato*, in both, state and private high schools during the academic years from 2009 – 2010 to 2014 – 2015.

The next step is to predict the number of *Bachillerato* students registered during the academic years 2009 – 2010 to 2014 – 2015 (**Step 3**). Due to the predictions of *Bachillerato* students are given in percentages (see Tables 3.7 and 3.10, Chapter 3), we need to estimate the number of students registered in both First and Second Stage of *Bachillerato* to be able to estimate the number of them who do not promote and abandon over the next few years using our predictions. To do that,

| Academic Year | Number of <i>Bachillerato</i> Students |
|---------------|--|
| 1999 – 2000   | 766 964                                |
| 2000 – 2001   | 738 407                                |
| 2001 – 2002   | 676 107                                |
| 2002 – 2003   | 654 655                                |
| 2003 – 2004   | 626 926                                |
| 2004 – 2005   | 613 581                                |
| 2005 – 2006   | 604 806                                |
| 2006 – 2007   | 595 571                                |
| 2007 – 2008   | 584 693                                |
| 2008 – 2009   | 629 247                                |

TABLE 5.4: Number of Spanish student in the First and Second Stage of *Bachillerato* in both, state and private high schools, all over Spain from academic year 1999 – 2000 to 2008 – 2009 [2].

| Model   | RMSE    | MAPE    |
|---|---------|---------|
| Random walk with trend                                | 27978.4 | 2.70597 |
| Linear trend  | 32784.3 | 3.83946 |
| Simple moving average of 3 terms                      | 43745.2 | 6.28963 |
| Simple exponential smoothing with alpha 0.999         | 30496.7 | 3.49021 |
| Brown's Linear Exponential Smoothing with alpha 0.853 | 29404.0 | 3.209   |

TABLE 5.5: The indicators (RMSE and MAPE) considered for the validation of the different models in order to determine the model that best fit the data in Table 5.4. The best is the *Random Walk with Trend Model*

we will again use the time series models mentioned above following the same procedure as it was shown previously, applied, in this case, to the number of *Bachillerato* students in the specific period of time given in Table 5.4 [2].

In this case, using *Statgraphics Plus for Windows 5.1*, the time series model selected that best fit our data in Table 5.4 is the *Random Walk with Trend Model*. This has the least Root Mean Square Error (RMSE) and Percentage of the Mean Absolute Error (MAPE), as can be seen in Table 5.5 (see Section D.2 and D.3, in Appendix D).

As regards to the definition of the *Random Walk with Trend Model* (see Section

| Academic<br>Year | Number of estimated <i>Bachillerato</i> Students |                          |
|------------------|--|--------------------------|
|                  | Predicted  | 95% Confidence Intervals |
| 2009 – 2010      | 613 945  | [562 245 , 665 646]      |
| 2010 – 2011      | 598 643  | [525 528 , 671 759]      |
| 2011 – 2012      | 583 341  | [493 793 , 672 889]      |
| 2012 – 2013      | 568 039  | [464 638 , 671 441]      |
| 2013 – 2014      | 552 738  | [437 132 , 668 343]      |
| 2014 – 2015      | 537 436  | [410 796 , 664 076]      |

TABLE 5.6: Estimations with 95% confidence intervals of the number of Spanish students in the First and Second Stage of *Bachillerato* in both, state and private high schools, all over Spain from academic year 2009 – 2010 to 2014 – 2015.

D.2, Appendix D), we consider  $Y_t$  as the observed number of *Bachillerato* students in a specific academic year at time  $t$  and  $F_t(k)$  the obtained forecast. Despite the *Statgraphics Plus for Windows 5.1* software only gives us the predictions if all the required assumptions are fulfilling, we will also confirm them analyzing statistically if the obtained white noise in this process follows a normal distribution, as is required. In order to check this, we apply the *Shapiro-Wilks normality test* which, with a significative p-value at significance level of 0.05 (p-value=0.407). The p-value confirms that the white noise follows a univariate normal distribution. This fact is also supported by having a closed mean and median (-13 345 and -15 302, respectively) and the kurtosis is 3.198, approximately 3, value considered as a reference to data following a univariate normal distribution [75].

Then, as the model is stated, in Table 5.6 we show the estimation with 95% confidence intervals of the number of Spanish *Bachillerato* students during the academic years from 2009 – 2010 to 2014 – 2015 (see Section D.4, Appendix D).

Finally, we compute the Spanish Government total investment in *Bachillerato* students that will not promote and abandon during the academic years 2009 – 2010 to 2014 – 2015 (**Step 4**). To obtain them, we take into account the Spanish Government investment in each *Bachillerato* student given in Table 5.3 and the estimated number of *Bachillerato* students in Table 5.6. After some algebraic operations (simply multiplications of the extremes of the intervals obtained in each mentioned tables), Table 5.7 collects the estimated number of students who

will not promote and abandon and their corresponding costs that would have for the Spanish Government in the next few years.

As we can see, if expectations are fulfilled and educational measures are not taken, the Spanish Government would lose a huge amount of money in groups of *Bachillerato* students who, most of them, would not promote and abandon the year or access to the labor market without sufficient qualification to perform works requiring improved training. Notice that, for example, this investment could be ranging between 39 226 440,83 and 68 848 080,60 euros in the academic year 2012 – 2013.

| Academic year | Estimated number of <i>Bachillerato</i> students who will not promote and abandon | Estimated Spanish Government investment (in euros) |
|---------------|---|--|
| 2009 – 2010   | [8 293 , 10 636]  | [43 306 812,55 , 61 747 912,30]                    |
| 2010 – 2011   | [7 561 , 10 502]  | [41 661 939,75 , 64 294 039,41]                    |
| 2011 – 2012   | [6 978 , 10 362]  | [40 449 413,28 , 66 728 551,64]                    |
| 2012 – 2013   | [6 450 , 10 186]  | [39 226 440,83 , 68 848 080,60]                    |
| 2013 – 2014   | [6 005 , 10 121]  | [38 225 011,05 , 71 646 592,10]                    |
| 2014 – 2015   | [5 541 , 9 902]   | [36 842 317,83 , 73 278 632,94]                    |

TABLE 5.7: Estimation with 95% confidence intervals of the number of *Bachillerato* students who do not promote and abandon in the First and Second Stage of *Bachillerato*, in both, state and private high schools all over Spain from academic year 2009 – 2010 to 2014 – 2015 and their corresponding cost for the Spanish Government also given with 95% confidence intervals.

### 5.3 Estimation with 95% confidence intervals of the investment in education by Spanish families of *Bachillerato* students in the next few years

In the previous section, we have estimated the cost that would have for the Spanish Government the predicted negative academic results of *Bachillerato* students. However, Government not only has to make those educational investments but also students' families. Undoubtedly, families have a very important role on their

children's education, in fact, most of students depend heavily on their parents for their studies, parents that, with their efforts, try to support and provide them the best conditions to develop their children's knowledge. That effort is commonly shown through their understanding, their care and, of course, with financial support. This financial support that, especially in periods of economical crisis as we are suffering at this moment, is really difficult for most families.

In this section, we will show that high rates of academic underachievement (including the abandon rates) not only have negative economical consequences on the Spanish Government but also to Spanish families, in particular, of *Bachillerato* students. To address this approach, we will estimate the Spanish families investment following the same procedure as it is shown in Section 5.2. For sake of clarity, in this case, we will follow the next steps:

**Step 1** We obtain the Spanish families cost of each *Bachillerato* student during the academic years 1999 – 2000 to 2008 – 2009.

**Step 2** We predict the Spanish families investment in each *Bachillerato* student during the academic years 2009 – 2010 to 2014 – 2015 using the cost of each *Bachillerato* student given in Step 1.

**Step 3** We compute the Spanish families total investment in *Bachillerato* students that will not promote and abandon during the academic years 2009 – 2010 to 2014 – 2015 using the predictions given in the previous step (Step 2) and in the Step 3 shown in Section 5.2.

First of all, we need to obtain the Spanish families cost of *Bachillerato* student during the academic years 1999 – 2000 to 2008 – 2009 (**Step 1**). For this, we collect the Spanish families investment over the total registered students in the non-university Spanish education during the corresponding academic years 1999 – 2000 to 2008 – 2009 given in [1]. Furthermore, we know the total number of non-university Spanish students registered [2]. These available data allow us to work out (dividing the Spanish families total investment over the total of registered non-university students by the corresponding number of non-university students) the Spanish families investment on each non-university Spanish student. Unfortunately, it has not been possible to get this information corresponding only to the *Bachillerato* educational level. As a consequence, we will consider these figures as a reference to determine, on average, the cost of a Spanish *Bachillerato* student for

| Spanish families Investment per<br><i>Bachillerato</i> student |               |          |
|--|---------------|----------|
| $t$  | Academic Year | Euros    |
| 1  | 1999 – 2000   | 889,21   |
| 2  | 2000 – 2001   | 900,85   |
| 3  | 2001 – 2002   | 951,66   |
| 4  | 2002 – 2003   | 1 008,23 |
| 5  | 2003 – 2004   | 1 028,53 |
| 6  | 2004 – 2005   | 1 067,29 |
| 7  | 2005 – 2006   | 1 131,23 |
| 8  | 2006 – 2007   | 1 156,78 |
| 9  | 2007 – 2008   | 1 173,82 |
| 10   | 2008 – 2009   | 1 141,92 |

TABLE 5.8: Spanish families investment, on average, per Spanish student in the First and Second Stage of *Bachillerato* in both, state and private high schools, all over Spain from academic year from 1999 – 2000 to 2008 – 2009 [1].

their families. Thus, in Table 5.8, we show, on average, the assumed Spanish families investment in each *Bachillerato* student during the academic years 1999 – 2000 to 2008 – 2009.

Then, we predict the Spanish families investment in each *Bachillerato* student during the academic years 2009 – 2010 to 2014 – 2015 (**Step 2**) using the cost of each *Bachillerato* student given in Step 1. These predictions, as it was developed in the previous section, have also been obtained by best time series model that fit the available data in Table 5.8 using, again, *Statgraphics Plus for Windows 5.1* software. After applying them, we consider that the model that best fit our data is the *Linear Trend Model* because it returns the minimum Root Mean Square Error (RMSE=27.705) whose corresponding Mean Absolute Percentage of Error is 1.71 (see Table 5.9). As a consequence, *Statgraphics Plus for Windows 5.1* software provides, by the model selected, 95% confidence intervals predictions of the Spanish families investment in each *Bachillerato* student over the next few years (see Appendix D). The obtained results can be seen in Table 5.10 and, graphically, in Figure 5.2.

Finally, we compute the Spanish families total investment in *Bachillerato* students



| Model   | RMSE    | MAPE    |
|---|---------|---------|
| Random walk with trend                                | 29.1247 | 2.04206 |
| Linear trend  | 27.7048 | 1.70595 |
| Simple moving average of 3 terms                      | 74.3422 | 6.37513 |
| Simple exponential smoothing with alpha 0.999         | 39.2763 | 2.9957  |
| Brown's Linear Exponential Smoothing with alpha 0.999 | 29.95   | 2.31063 |

TABLE 5.9: The indicators (RMSE and MAPE) considered for the validation of the different models in order to determine the model that best fit the data in Table 5.8. The best is the *Linear Trend Model*

| $t$ | Academic Year | Prediction (Euros) | 95% Confidence interval (Euros) |
|-----|---------------|--------------------|---------------------------------|
| 11  | 2009 – 2010   | 1 232,24           | [1 154,86 , 1 309,61]           |
| 12  | 2010 – 2011   | 1 266,29           | [1 185,17 , 1 347,40]           |
| 13  | 2011 – 2012   | 1 300,34           | [1 215,06 , 1 385,62]           |
| 14  | 2012 – 2013   | 1 334,39           | [1 244,59 , 1 424,19]           |
| 15  | 2013 – 2014   | 1 368,44           | [1 273,81 , 1 463,07]           |
| 16  | 2014 – 2015   | 1 402,49           | [1 302,77 , 1 502,21]           |

TABLE 5.10: The prediction of euros Spanish families will invest in each Spanish student in the First and Second Stage of *Bachillerato*, in both, state and private high schools during the academic years from 2009 – 2010 to 2014 – 2015.

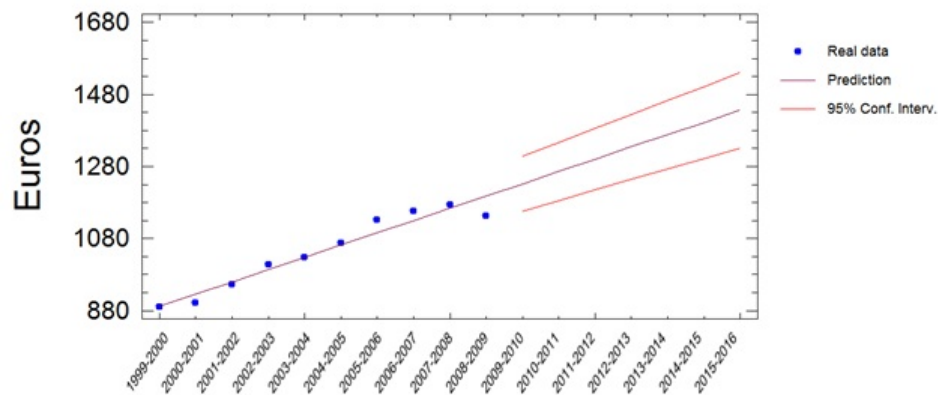


FIGURE 5.2: Graph of the prediction (in euros) the Spanish families will invest in each *Bachillerato* student during the academic years from 2009 – 2010 to 2014 – 2015.

that will not promote and abandon during the academic years 2009–2010 to 2014–2015 (**Step 3**). To obtain them, we use the estimated number of *Bachillerato*

students that will not promote and abandon (see Table 5.7) and the cost for the Spanish families of each *Bachillerato* student during the academic years 2009 – 2010 to 2014 – 2015 (see Table 5.8). After some algebraic operations (simply multiplications of the extremes of the intervals obtained in each mentioned tables), in Table 5.11, we show the estimation of the Spanish families total investment in education during the academic years from 2009 – 2010 to 2014 – 2015.

Notice that these values could be ranging between 8 027 735,83 and 14 507 891,88 euros in the current academic year. No negligible amount of money if we consider the economic difficult situation of most Spanish families as a result of the severe economic crisis in Spain is immersed.

| Academic Year | Estimated Spanish families investment (in euros) |
|---------------|--|
| 2009 – 2010   | [9 577 515,21 , 13 929 654,41]                   |
| 2010 – 2011   | [8 961 297,64 , 14 151 586,96]                   |
| 2011 – 2012   | [8 479 092,15 , 14 358 567,46]                   |
| 2012 – 2013   | [8 027 735,83 , 14 507 891,88]                   |
| 2013 – 2014   | [7 649 301,83 , 14 807 905,66]                   |
| 2014 – 2015   | [7 219 319,24 , 14 875 138,87]                   |

TABLE 5.11: 95% confidence intervals of the Spanish families cost in the group of *Bachillerato* students with academic underachievement over the next few years.

## 5.4 Conclusions

In this chapter, we quantify the important social problem of the academic underachievement, we take advantage of our predictions of the Spanish academic performance to propose an estimation of the Spanish Government and families investment in the *Bachillerato* students over the next few years, paying special attention on the groups of students who abandon and do not promote during their corresponding academic year. According to our results, notice that, for example, in the academic year 2012 – 2013, the Spanish Government will invest in students

with academic underachievement a large amount of money, ranging between 39 226 440,83 and 68 848 080,60 euros and, in case of the Spanish families, the costs will range between 8 027 735,83 and 14 507 891,88 euros. According to our predictions (the total number of Bachillerato student and the cost per *Bachillerato* student for the Spanish Government and families given in Tables 5.4, 5.3, 5.10, respectively), these amounts of money, on average, would represent around the 1.5% of the Spanish Government total investment predicted and the 1.4% Spanish families total investment predicted in the academic year 2012 – 2013.

From our expectations and if new and innovative educational measures are not taken, the Spanish Government and families would lose a huge amount of money in groups of *Bachillerato* students who, most of them, would have to repeat a year or access to the labor market without sufficient qualification to perform works requiring improved training.



# Chapter 6

## Conclusion and discussion

In this dissertation, we have proposed a non-linear system of differential equations to model the evolution of the academic performance in the educational level of *Bachillerato* in Spain over the next few years using mathematical epidemiology modeling techniques. We have focused on this educational level since it represents a milestone in the career training of students because they have to make important decisions about academic and professional future: keep studying to get a better qualification or access to the labor market [23]. We have paid special attention on the academic underachievement in the Spanish *Bachillerato* since, although the percentage of high school academic underachievement in this educational level has slightly reduced over the last years, nowadays it seems to be at a worrying steady-level (around 30%) [11]. This is becoming a major social and political concern because of the negative effects on the country's economic development [18], especially in the unemployment and its serious consequences.

The major novelty of this contribution is the treatment of academic performance as a problem that is transmitted through social contact using mathematical type-epidemiological modeling. This is based on the idea that academic habits of any student is a mixture of personal decisions and influence of classmates [29, 30].

We have developed a first mathematical model (Chapter 2) in which we have considered the academic attitude of *Bachillerato* students depends on their autonomous positive/negative academic behavior and the transmission of bad academic habits transmitted by other students in the same academic level [23, 46, 47]. Moreover, in order to reflect as truthfully as possible the attitude of students towards their studies, we have taken into account pedagogical studies [42–45] which

state that exists a significative difference of academic performance depending on genre. This model has allowed us to predict the academic underachievement evolution over the next few years. Then, this mathematical model has been improved in Chapter 3. It is based on the same ideas developed in the previous chapter but, in this case, we consider that not only the bad academic habits are socially transmitted but also the good ones. We have also decomposed the transmission parameters into good and bad academic habits in order to analyze with more detail which group of students are more susceptible to be influenced by good or bad academic students. Besides, we have introduced uncertainty in the model. This enables us to predict the evolution of the Spanish *Bachillerato* academic performance by 95% confidence intervals. The model presented in this chapter is validated verifying that the predictions given by 95% confidence intervals either collect or are nearby (with an error, at most, of order  $10^{-2}$ , in absolute terms) of the deterministic estimations provided by the model developed in Chapter 2 including new available data which were published during the development of this dissertation. Other important improvement in this model is the quantification of the abandon rates. The results inform us that there is a slight decreasing of the percentage of students in the non-promotable groups and the ones who leave the high school. It seems to reach a stationary situation passing, on average, from 28% in the academic year 2010 to 26.5% in 2015. However, the current and predicted scenarios are worrying because around 27% of the students have bad academic results.

An important characteristic of the proposed mathematical model is its ability to be adapted, although in this research we have focused on *Bachillerato* educational level, to other Spanish educational levels and also foregoing any educational systems. This is illustrated in Chapter 4 where the proposed model for Spanish *Bachillerato* has been adapted successfully to study the academic performance in the German region of North Rhine-Westphalia. The obtained results show that there is a significant difference between the academic underachievement performance of both academic systems, namely, the Spanish model seems to be stabilizing at worrying levels whereas in the German system this value is much lower. For instance, in the course 2014 – 2015 less than 2% of the students will not promote.

In the treatment of social issues as the study of education, it has been inevitable to refer to the current critical economic situation in which most of the European countries are immersed, mainly, countries like Spain where the unemployment

rates are much higher than the rest of the European Countries [21]. In order to contribute positively to this important social problem, in Chapter 5 of this dissertation, we have quantified and estimated the Spanish Government and Spanish families investment in the Spanish *Bachillerato* over the next few years, especially, focused on the groups of students who abandon and do not promote whose academic attitude could lead an increasing of the economic costs. According to our predictions and if any educational measure is not taken, the Spanish Government, families and Society in general would lose a huge amount of money in groups of *Bachillerato* students who, most of them, would have to repeat a year or access to the labor market without sufficient qualification to perform works requiring improvement training. For example, in the academic year 2012 – 2013, the Spanish Government will invest in students with academic underachievement a large amount of money, ranging between 39 226 440.83 and 68 848 080.60 euros. In case of the Spanish families, these figures could be ranging between 8 027 735,83 and 14 507 891,88 euros. No negligible and alarming amounts of money if we consider the severe economic crisis that is currently affecting to Spain.

The proposed approach will allow us to understand better the mechanisms behind the academic performance as well as to predict and quantify the evolution of the Spanish *Bachillerato* students in the coming years. As it has been stated, our model gives us information about how academic results will evolve over the next future among different groups of *Bachillerato* students according to their gender, academic level and their academic results. In this way, it could provide relevant information in order to policymakers make appropriate decisions, for instance, policies of inclusion or gather to improve the transmission of good academic habits and avoid the transmission of the bad ones. In general terms, the results inform us that there is a slight decreasing of the number of students in the non-promotable groups and who leave the high school, and it seems to reach a stationary situation. The current and predicted scenarios are very worrying because around 27% of the total of *Bachillerato* students could get bad academic results in the coming years.

To conclude, we want to point out that this contribution constitutes a first step in the modeling of academic underachievement applying a type-epidemiological approach. We think that more research following the proposed approach must be done in the network framework.





# Appendix A

## epiModel: A system to build automatically systems of differential equations of compartmental type-epidemiological models

### A.1 Introduction

In this dissertation we propose a non-linear system of differential equations to model the evolution of the academic performance in the educational level of *Bachillerato* in Spain using modelling techniques in mathematical epidemiology also applied to the German educational system. As the model is stated the building of its corresponding non-linear system of differential equations could be a tedious work because of the large number of equations could contain. For this reason, in this appendix, we present the software developed to facilitate the construction of that system of equations facilitating users to work in models as we can see through this dissertation.

From Kermarck-McKendricks seminal paper of 1927 [76] epidemiologists and mathematicians have developed mathematical models to understand the transmission dynamics of diseases. The advances in this area have led to more complex models

and, therefore, larger systems of differential equations. For instance, the model developed in [77] for the study of the spread of Human Papillomavirus (HPV) is made up of more than 7,000 equations, the model described in [78] for the dynamics of meningococcal disease with around nine hundred equations or the models described in chapters 2, 3 and 4 of this dissertation.

Compartmental diagrams have enabled the development of the epidemiological models and their expression as differential equations. However, when the models include a lot of subpopulations and take into account age and/or sex groups, the building of the system of differential equations become more complex when handling a large number of functions and parameters.

Thus, in order to facilitate researchers in the epidemiology area to build model equations for linear or quadratic epidemiological models, in this appendix we present *epiModel*, a code developed in *Mathematica* capable of automatically generating the system of differential equations and its parameters from a short and easy description of the model contained in a text file.

*epiModel* consists of three files (see scheme in Figure A.1):

- "ModelDefinition". In this file the user describes the characteristics of the model using a simple syntax explained in Section 2.
- "epiModel". This file contains the code that carries out the transformation of the data model in "ModelDefinition" into a system of differential equations.
- "ModelBuilder.nb". This *Mathematica* file loads the files "ModelDefinition" and "epiModel" and executes them in order to generate two new files:
  - "Model.data" with the system of differential equations.
  - "parameters.data" with a list of all the model parameters.

This appendix is organized as follows. In Section 2 we describe how to build the file "ModelDefinition". Once the "ModelDefinition" has been built, in Section 3 we explain how to generate the files "Model.data" and "parameters.data". In Section 4, three examples are presented: in the first, we generate the system of differential equations corresponding to a typical SIRS (Susceptible - Infected - Recovered - Susceptible) epidemiological model; in the second, the same is done

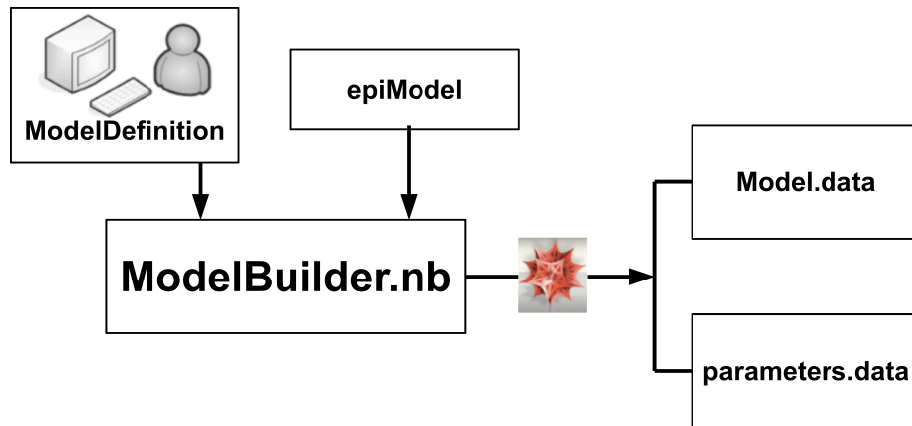


FIGURE A.1: Process of how *epiModel* works. "ModelBuilder.nb" loads data from "ModelDefinition" and "epiModel" creates "Model.data" and "parameters.data".

for a SIR (Susceptible - Infected - Recovered) epidemiological model with two age groups; finally, the equations for a SIR model with two age and two sex groups are generated. In Section 5, conclusions are given.

This appendix does not aim to explain the code line by line, it should be stated that this represents a slight improvement on the idea of Capasso [79, 80] as to how to represent an epidemiological model in matrix form. In fact, when a model is generated, the file "Model.data" contains the system of equations and the matrices corresponding to the matrix form of the model.

Thus, if the vector  $z(t) = (z_1(t), \dots, z_n(t))^T$  contains as entries the model subpopulation functions and we denote by

$$\text{diag}(z(t)) = \begin{pmatrix} z_1(t) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & z_n(t) \end{pmatrix},$$

any compartmental model (even including age and/or sex groups) can be written as

$$\frac{dz(t)}{dt} = c + Lz(t) + \sum_i A_i \text{diag}(z(t)) B_i z(t), \quad (\text{A.1})$$

where  $c$  is a vector of size  $n$  and  $L$ ,  $A_i$  and  $B_i$  are matrices of size  $n \times n$ .  $c$  contains parameters corresponding to the model independent terms,  $L$  to the linear terms and  $A_i$  and  $B_i$  to the model non-linear terms, placed in the appropriate coordinates.

The variations on Capasso's idea is the inclusion of matrices  $A_i$  and  $B_i$  that allow us to take several subpopulations (usually the same in different age/sex groups) as a part of the same transmission (non-linear) term. In Section 4, in the first and second example, we will provide the obtained matrices  $c$ ,  $L$ ,  $A_i$  and  $B_i$ .

## A.2 How to build the file "ModelDefinition"

This is a text file and is made up of three parts: a general variable; the definition of subpopulations; and the definition of the parameters. Note that the syntax of this file should fit the *Mathematica* syntax.

It is important to preserve the names of the variables defined below ( $\backslash$ [NTilde], **SP**, **TI**, **LIN**, **NOLIN**,  $x$ ) as they will be called by "epiModel".

### A.2.1 General variable

This variable indicates the structural characteristics of the model, i.e. age groups. The variable is

|    | Name                  | Value  | Description           |
|----|-----------------------|--------|-----------------------|
| 1. | $\backslash$ [NTilde] | number | Number of age groups. |

### A.2.2 Definition of the subpopulations

Data corresponding to subpopulations are stored in a list named **SP**. Each row of the list consists of the following fields

|    | Name               | Value  | Description                |
|----|--------------------|--------|----------------------------|
| 1. | <b>Number</b>      | Number | Subpopulation ID number.   |
| 2. | <b>Description</b> | String | Name of the subpopulation. |

In the following example, three subpopulations of a model are defined.

```
SP = {
  {1, "Susceptible"},
  {2, "Infected" },
  {3, "Recovered" }
};
```

### A.2.3 Defining Parameters

The parameters have two ways of being classified. The first is dependent on the part of the model to which they contribute:

- Those that are included in the independent term of the model, being in list **TI**.
- Those that are linear terms of the model, being in list **LIN**.
- Those that are part of the non-linear term of the model, being stored in list **NOLIN**.

The second classification is dependent on the type of parameter. To explain this, we should note that compartmental models are illustrated by diagrams where the boxes represent the subpopulations and the arrows represent the terms involving the model parameters. However, not all the arrows are equal: some only enter into a box; some of them only exit from a box; others exit from a box and enter into another; some of the latter are special because they connect the same box for different age groups (see Figure A.2). These four possibilities lead to the second parameter classification:

- **Type 1:** also called *birth type*, because this parameter comes from arrows that only enter into a subpopulation, e.g. newborns.
- **Type 2:** also called *death type*, because it is related to an arrow that only exits from a subpopulation, e.g. dead people leaving the system.

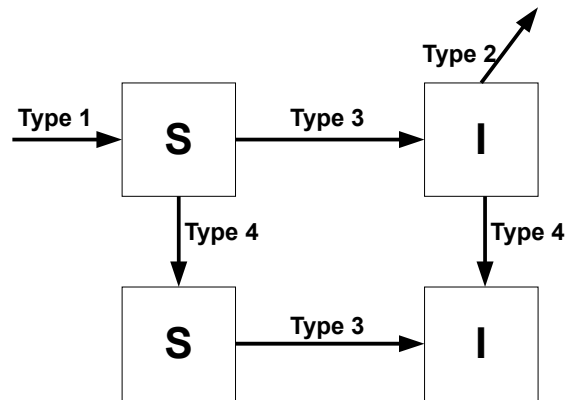


FIGURE A.2: Parameter types dependent on where the arrows enter and exit in compartmental models.

- **Type 3:** also called *input-output type*, because this parameter measures the flow from one box or subpopulation to another, for instance, disease transmission or recovering illness average time.
- **Type 4:** also called *between ages type*, because this parameter is related to population growth and connects the same box in two consecutive age groups.

### A.2.3.1 Parameters of independent term and linear term

Each independent term or linear term parameter is encoded with a list containing the following fields:

|    | Name  | Value               | Description  |
|----|---|---------------------|--|
| 1. | <b>Name</b>                                     |                     | Parameter name.  |
| 2. | <b>Type</b>                                     | 1, 2, 3, 4          | Parameter type.  |
| 3. | <b>The arrow exits from the subpopulations</b>  | $\{n1, n2, \dots\}$ | Subpopulations from which the arrow related to the current parameter, exits. The list $\{\}$ means that the arrow does not exit from any box (e.g., a birth type parameter) .  |
| 4. | <b>The arrow enters into the subpopulations</b> | $\{n1, n2, \dots\}$ | Subpopulations from which the arrow related to the current parameter, enters. The list $\{\}$ means that the arrow does not enter into any box (e.g., a death type parameter). |
| 5. | <b>Depending on the age group?</b>              | True/False          | True means that this parameter can be different depending on the age group.  |
| 6. | <b>Description</b>                              | String              | Description of the parameter.  |

The following considerations concerning with the definition of parameters should be taken into account:

- Any parameter should not be named  $x$  because this variable defines the subpopulations.
- A parameter of **Type 3** can not have common elements in the fields 3 and 4 of the above table. These situations can be avoided by defining various parameters properly.
- If a parameter does not depend on age groups, i.e. the field 5 in the table is False, it will appear only in the first age group between the boxes included in the lists of the fields 3 and/or 4.
- **Type 4** parameters cannot appear in the variable **TI**.
- The **Type 4** parameters have always to be dependent on age group.

In order to avoid confusion, it is not convenient to use the same or similar variable names for different parameters. Now, let us show some examples of how to encode parameters corresponding to independent and linear model terms:

- **Type 1:** Suppose that newborns enter directly into subpopulation 1 at a rate  $\mu$ . The model does not consider age group. Then, this term is encoded as

```
{\mu, 1, {}, {1}, False, "Birth rate"}
```

- **Type 2:** Now, we consider an age group model with three subpopulations where the death rate depends on age group and all people of any subpopulation is susceptible to death. This is encoded as:

```
{d, 2, {1,2,3}, {}, True, "Death rate"}
```

- **Type 3:** Let us suppose that, after recovering from a disease, the individuals have an average temporary immunity  $\gamma$  in a typical SIRS model with age groups. This parameter will be encoded as:

```
{\gamma, 3, {3}, {1}, True, "Average immunity time"}
```

- **Type 4:** In an age group model with three subpopulations, the growth rate  $c$  is encoded as follows:

```
{c, 4, {1,2,3}, {}, True, "Growth rate"}
```

### A.2.3.2 Parameters of non-linear terms

Each non-linear term parameter is encoded using the following fields:



|    | Name  | Value               | Description  |
|----|---|---------------------|--|
| 1. | <b>Name</b>   |                     | Parameter name.  |
| 2. | <b>Transmission is affected by the subpopulations</b> | $\{n1, n2, \dots\}$ | Those subpopulations related to any infectious state (infectious, latent, etc.). |
| 3. | <b>The arrow exits from the subpopulation</b>         | $\{n1\}$            | Subpopulation whose individuals are susceptible to be infected.                  |
| 4. | <b>The arrow enters into the subpopulation</b>        | $\{n2\}$            | Subpopulation where an infected individual enters (latent, infectious, etc.).    |
| 5. | <b>Depending on the age groups?</b>                   | True/False          | True means that this parameter can be different depending on the age group.      |
| 6. | <b>Description</b>                                    | String              | Description of the parameter.  |

The same advice given for independent and linear parameters can be applied to non-linear ones. In the following example, the list

`{\[Beta], {3,4}, {1}, {2}, True, Transmission rate}`

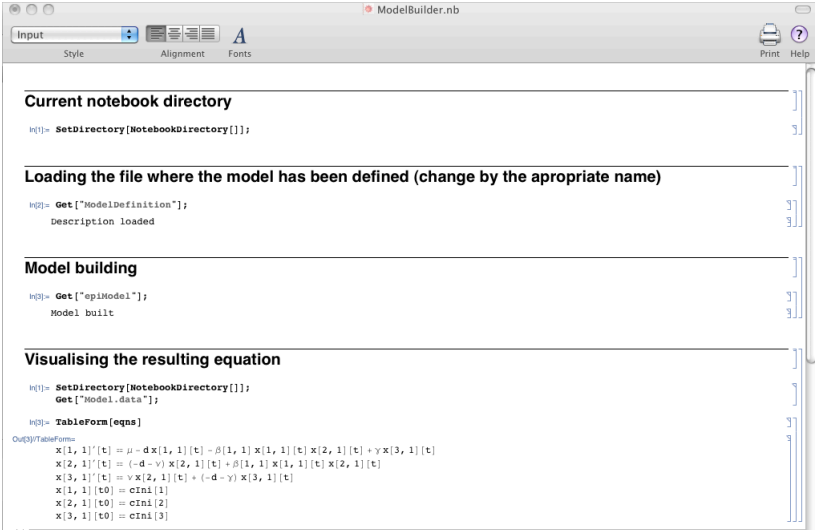
indicates that if an individual  $A$  in subpopulation 1 of any age group has successful contact (the disease is transmitted) with another individual belonging to subpopulations 3 or 4 of any age group, individual  $A$  moves to subpopulation 2 in the same age group as  $A$  was previously.

### A.3 Steps to building the system of differential equations

As we said before, the code was developed using *Mathematica 7* or higher [52]. Then, in order to build the system of differential equations, we need to have

*Mathematica* installed in the computer. Moreover, the files "ModelDefinition", "epiModel" and "ModelBuilder.nb" have to be in the same directory. Then, open the notebook "ModelBuilder.nb" using *Mathematica*. This notebook has 4 cells.

- 1st. This cell allows us to set the current directory (the directory containing the files to build the model) as the working directory.
- 2nd. The second cell loads the text file "ModelDefinition" where we have defined the model, following the rules described in Section 2.
- 3rd. The third cell loads "epiModel" and executes it. Then the files "Model.data" and "parameter.data" appear in the working directory.
- 4th. Once the system of equations is in the "Model.data" file, this cell loads this file and displays the resulting system of equations. This is useful to verify that no errors have occurred and check the correctness of the equations.



```

Input
Style Alignment Fonts Print Help

Current notebook directory
In[1]:= SetDirectory[NotebookDirectory[]];

Loading the file where the model has been defined (change by the appropriate name)
In[2]:= Get["ModelDefinition"];
Description loaded

Model building
In[3]:= Get["epiModel"];
Model built

Visualising the resulting equation
In[4]:= SetDirectory[NotebookDirectory[]];
Get["Model.data"];
In[5]:= TableForm[eqns]

Out[5]:TableForm:
x[1, 1][t] =  $\mu - d x[1, 1][t] - \beta[1, 1] x[1, 1][t] x[2, 1][t] + \gamma x[3, 1][t]$ 
x[2, 1][t] =  $(-d - \gamma) x[2, 1][t] + \beta[1, 1] x[1, 1][t] x[2, 1][t]$ 
x[3, 1][t] =  $\nu x[2, 1][t] + (-d - \gamma) x[3, 1][t]$ 
x[1, 1][t0] = cIni[1]
x[2, 1][t0] = cIni[2]
x[3, 1][t0] = cIni[3]

```

FIGURE A.3: Screenshot of "ModelBuilder.nb" in *Mathematica*.

Given that no errors appeared, two files will be generated: "Model.data" and "parameters.data".

### A.3.1 The file "Model.data"

This file contains the variables  $ti$ ,  $mc$ ,  $mcNoL1$ ,  $mcNoL2$ ,  $Fvars$ ,  $eqns$ ,  $vars$ .

- $ti$  corresponds to the vector independent term of the model, i.e. vector  $c$  in expression (A.1).
- $mc$  corresponds to the coefficient matrix of the linear model, i.e matrix  $L$  in (A.1).
- $mcNoL1$  and  $mcNoL2$  are the coefficient matrices which enable the construction of the non-linear part of the model, i.e. matrices  $A_i$  and  $B_i$  in (A.1), respectively.
- $Fvars$  is a vector function where each entry corresponds to each subpopulation in the model.
- $eqns$  is a list with the system of differential equations. It is built computing the expression (A.1) using all the above matrices.
- $vars$  is the same as  $Fvars$  but removing  $t$  in the functions.

Thus, if we want numerically to solve a model in *Mathematica* (system of differential equations) we execute

```
sol = NDSolve[ eqns, vars, {t, t0, tEnd} ]
```

and in order to evaluate and draw the solutions we can execute

```
Plot[ Evaluate[ Fvars /. sol ], {t, t0, tEnd} ]
```

### A.3.2 The file "parameters.data"

This file has a complete list of the parameters appearing in the model. This is useful because some of them can be replaced by known values and the ones that can not, can be included in a procedure to be estimated.

## A.4 Examples

In this section let us show three examples with different options to build type-epidemiological models.

### A.4.1 SIRS model

The first example is the classical SIRS model, where we have three subpopulations: Susceptible (S), Infectious (I) and Recovered (R). The transmission is carried out with effective contacts between a susceptible individual and an infectious individual. Once an individual has been infected he/she recovers and acquires temporal immunity. When this finishes, the individual again becomes susceptible. This description has been depicted in Figure A.4.

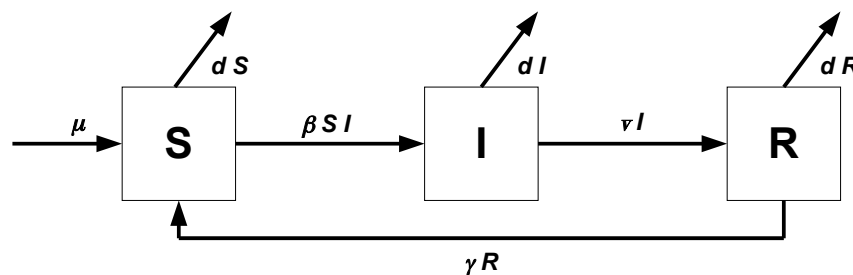


FIGURE A.4: Diagram of a Susceptible-Infectious-Recovered-Susceptible model.

To build the "ModelDefinition" file, we should take into account that:

- It is a model without age groups.
- $\mu$  is a Type 1 parameter belonging to the independent term model, because this represents newborns that enter directly into the susceptible subpopulation.
- Parameter  $d$  is the death rate (Type 2) and exits from all the subpopulations.
- $\beta$  is the transmission parameter belonging to the non-linear term.
- The Type 3 parameters  $\gamma$  and  $\nu$  belong to linear terms. They are the average time of infection and immunity respectively.

Note that the number of age groups indicated is 1 and there are no parameters depending on the age groups. In this way, the "ModelDefinition" file will be as follows:

```
(* Number of age groups *)
\[NTilde] = 1;
```

```

(* Subpopulations *)
SP = {
{1, "Susceptible"},
{2, "Infected" },
{3, "Recovered" }
};

(* INDEPENDENT TERM *)
TI = {
{\[Mu], 1, {}, {1}, False, "Birth rate"}
};

(* LINEAR TERM *)
LIN = {
{ \[Nu], 3, {2}, {3}, False, "Average time of infection"},
{\[Gamma], 3, {3}, {1}, False, "Average time of immunity"},
{ d, 2, {1,2,3}, {}, False, "Death rate"}
};

(* NON LINEAR TERM *)
NOLIN = {
{\[Beta], {2}, {1}, {2}, False, "Transmission rate"}
};

```

After running "epiModel" and building the model from the above data in the "ModelDefinition" file, *Mathematica* returns the following system of differential equations ( $x[1, 1]$  is susceptible,  $x[2, 1]$  infectious and  $x[3, 1]$  recovered subpopulations):

$$\begin{aligned}
x[1, 1]'[t] &== \mu - dx[1, 1][t] - \beta x[1, 1][t]x[2, 1][t] + \gamma x[3, 1][t] \\
x[2, 1]'[t] &== (-d - \nu)x[2, 1][t] + \beta[1, 1]x[1, 1][t]x[2, 1][t] \\
x[3, 1]'[t] &== \nu x[2, 1][t] + (-d - \gamma)x[3, 1][t] \\
x[1, 1][t_0] &== cIni[1] \\
x[2, 1][t_0] &== cIni[2] \\
x[3, 1][t_0] &== cIni[3]
\end{aligned}$$

and the matrices that, computing the expression (A.1), enable the construction of the above system are:

$$c = \begin{pmatrix} \mu \\ 0 \\ 0 \end{pmatrix}, \quad L = \begin{pmatrix} -d & 0 & \gamma \\ 0 & -d - \nu & 0 \\ 0 & \nu & -d - \gamma \end{pmatrix},$$

$$A_1 = \begin{pmatrix} -1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 0 & \beta & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

### A.4.2 SIR model with two age groups

This is a typical SIR (Susceptible-Infectious-Recovered) model with two age groups. We have two susceptible groups,  $S_1$  and  $S_2$ , one for each age group, the same for infectious,  $I_1$  and  $I_2$  and for recovered  $R_1$  and  $R_2$ . An individual in  $S_1$ ,  $I_1$  or  $R_1$  grows up and can enter in the corresponding box of the 2nd age group,  $S_2$ ,  $I_2$  or  $R_2$ , respectively. Transmission is carried out with effective contacts between a susceptible individual and an infectious individual of any age group. Once an individual has been infected, after set time, he or she recovers. This description has been depicted in Figure A.5.

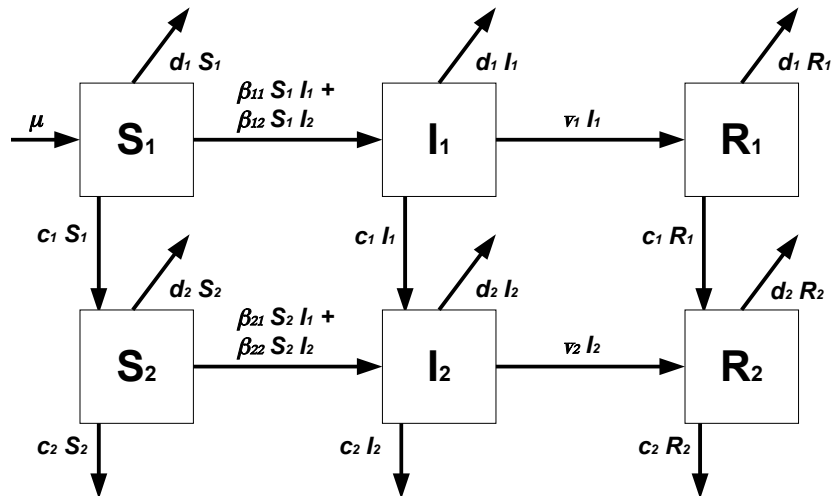


FIGURE A.5: Diagram of a Susceptible-Infectious-Recovered model with two age groups.

To build the "ModelDefinition" file, we take into account that:

- It is a model with 2 age groups.
- $\mu$  is a Type 1 parameter belonging to the independent term model that enters directly into the susceptible subpopulation of the first age group, i.e. it does not depend on age group.
- $\beta$  is the transmission parameter belonging to the non-linear term. "epi-Model" will generate four different parameters ( $\beta[1,1]$ ,  $\beta[1,2]$ ,  $\beta[2,1]$  and  $\beta[2,2]$ ) depending on the crossed products between susceptible and infectious subpopulations.
- The Type 2 death parameter  $d$  depends on the age group. "epiModel" will generate  $d[1]$  parameter for the first age group and  $d[2]$  for the second one.
- The Type 3 parameter  $\nu$  belongs to linear term. It is the average recovery time and also depends on age group.
- Parameter  $c$  is Type 4 and is the population growth rate between these age groups.

Thus, the "ModelDefinition" file will be as follows:

```
(* Number of age groups *)
\NTilde = 2;

(* Subpopulations *)
SP = {
{1, "Susceptible"},
{2, "Infected" },
{3, "Recovered" }
};

(* INDEPENDENT TERM *)
TI = {
{\Mu}, 1, {}, {1}, False, "Birth rate"}
};
```

```
(* LINEAR TERM *)
LIN = {
{ \[Nu], 3, {2}, {3}, True, "Average time of infection"},
{ d, 2, {1,2,3}, {}, True, "Death rate"},
{ c, 4, {1,2,3}, {}, True, "Growth rate"}
};

(* NON LINEAR TERM *)
NOLIN = {
{\[Beta], {2}, {1}, {2}, False, "Transmission rate"}
};
```

Then, running "epiModel" and building the model from the above data in the "ModelDefinition" file, *Mathematica* returns the following system of differential equations ( $x[1, 1]$ ,  $x[2, 1]$  and  $x[3, 1]$  which are susceptible, infectious and recovered subpopulations, respectively, of the first age group, and  $x[1, 2]$ ,  $x[2, 2]$ ,  $x[3, 2]$  for the second one):

$$\begin{aligned}
x[1, 1]'[t] &== \mu + (-c[1] - d[1])x[1, 1][t] \\
&\quad -\beta[1, 1]x[1, 1][t]x[2, 1][t] - \beta[1, 2]x[1, 1][t]x[2, 2][t] \\
x[2, 1]'[t] &== (-c[1] - d[1] - \nu[1])x[2, 1][t] \\
&\quad +\beta[1, 1]x[1, 1][t]x[2, 1][t] + \beta[1, 2]x[1, 1][t]x[2, 2][t] \\
x[3, 1]'[t] &== \nu[1]x[2, 1][t] + (-c[1] - d[1])x[3, 1][t] \\
x[1, 2]'[t] &== c[1]x[1, 1][t] + (-c[2] - d[2])x[1, 2][t] \\
x[2, 2]'[t] &== c[1]x[2, 1][t] + (-c[2] - d[2] - \nu[2])x[2, 2][t] \\
x[3, 2]'[t] &== \nu[2]x[2, 2][t] + c[1]x[3, 1][t] + (-c[2] - d[2])x[3, 2][t] \\
x[1, 1][t_0] &== cIni[1] \\
x[2, 1][t_0] &== cIni[2] \\
x[3, 1][t_0] &== cIni[3] \\
x[1, 2][t_0] &== cIni[4] \\
x[2, 2][t_0] &== cIni[5] \\
x[3, 2][t_0] &== cIni[6]
\end{aligned}$$

and the matrices that permit the construction of the above system, computing the expression (A.1), are:



$$L = \begin{pmatrix} -c[1] - d[1] & 0 & 0 & 0 & 0 & 0 \\ 0 & -c[1] - d[1] - \nu[1] & 0 & 0 & 0 & 0 \\ 0 & \nu[1] & -c[1] - d[1] & 0 & 0 & 0 \\ c[1] & 0 & 0 & -c[2] - d[2] & 0 & 0 \\ 0 & c[1] & 0 & 0 & -c[2] - d[2] - \nu[2] & 0 \\ 0 & 0 & c[1] & 0 & \nu[2] & -c[2] - d[2] \end{pmatrix},$$

$$c = \begin{pmatrix} \mu \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 0 & \beta[1, 1] & 0 & 0 & \beta[1, 2] & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \beta[2, 1] & 0 & 0 & \beta[2, 2] & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Note that the model includes parameter  $c[2]$ . This parameter makes sense if people in the second age group leave the system in a form different from death, otherwise,  $c[2]$  would be redundant because  $d[2]$  plays the same role and then,  $c[2]$  should be zero.

### A.4.3 SIR model with two age groups and two sexes

"epiModel" is not designed to build systems of differential equations from gender models, however, considering some "tricks", we can transform a typical age group model into an age group and gender model. These "tricks" are:

- Consider first the age groups for females and then for males. Then, if there are  $n$  age groups for each sex, we should consider a model with  $2n$  age groups ( $\backslash[\text{NTilde}] = 2n$ ).
- The growth parameter  $c[n]$  connecting the last female age group (group  $n$ ) with the first male age group (group  $n + 1$ ) is zero.
- Birth rate should be considered age dependent in order to take into account different birth rates for females and males. Then, Birth rate, say  $\mu[i]$  is a Type 1 parameter depending on age group. However, newborns enter into

the first age group, group 1 for females ( $\mu[1]$ ) and group  $n + 1$  for males ( $\mu[n + 1]$ ). The remainder  $\mu[i]$  are zero.

This example consists of a typical SIR (Susceptible-Infectious-Recovered) model with two age groups and two sexes for each age group. We have two susceptible female groups,  $S_1$  and  $S_2$ , and two susceptible male groups,  $S_3$  and  $S_4$ , one for each age group. The same for infectious,  $I_1, I_2, I_3$  and  $I_4$ , and for recovered,  $R_1, R_2, R_3$  and  $R_4$ . An individual in  $S_1, I_1, R_1, S_3, I_3, R_3$  grows up and can enter into the box  $S_2, I_2, R_2, S_4, I_4, R_4$  of the 2nd age group, respectively. People leave the system by death.

For this example, let us suppose that the disease considered is heterosexually transmitted. The transmission is carried out with effective contacts between a susceptible male or female and an infectious individual of any age group of the other sex. Once an individual has been infected, after some time, recovers. To build the "ModelDefinition" file, we take into account that:

- It is a model with 4 age groups, two for females and two for males.
- $\mu$  is a Type 1 parameter belonging to the independent term model that enters directly into the susceptible subpopulations of the first age group for males and females. This requires that  $\mu$  depends on age group and  $\mu[2] = \mu[4] = 0$ .
- $\beta$  is the transmission parameter belonging to the non-linear term. "epi-Model" will generate  $\beta[i, j]$  for  $i, j = 1, 2, 3, 4$ , one for each type of contact. Taking into account that this disease is heterosexually transmitted, parameters  $\beta[1, 1], \beta[1, 2], \beta[2, 1], \beta[2, 2], \beta[3, 3], \beta[3, 4], \beta[4, 3]$  and  $\beta[4, 4]$  are zero.
- The Type 2 death parameter  $d$  depends on age group. "epiModel" will generate  $d[i]$  parameters for  $i = 1, 2, 3, 4$ , the two first for the female age groups and the remainder for male age groups.
- The Type 3 parameter  $\nu$  belongs to linear term. It is the average recovery time and also depends on age group and gender.
- Parameter  $c$  is of Type 4 and denotes the population growth rate between these age groups. As we mentioned before,  $c[2] = 0$ . Moreover, as people leave the system by death,  $c[4] = 0$  because it plays the same role as the death parameter  $d[4]$ .

Thus, the "ModelDefinition" file will be as follows:

```
(* Number of age groups *)
\[NTilde] = 4;

(* Subpopulations *)
SP = {
{1, "Susceptible"},
{2, "Infected" },
{3, "Recovered" }
};

(* INDEPENDENT TERM *)
TI = {
{\[Mu], 1, {}, {1}, True, "Birth rate"}
};

(* LINEAR TERM *)
LIN = {
{ \[Nu], 3, {2}, {3}, True, "Average time of infection"},
{ d, 2, {1,2,3}, {}, True, "Death rate"},
{ c, 4, {1,2,3}, {}, True, "Growth rate"}
};

(* NON LINEAR TERM *)
NOLIN = {
{\[Beta], {2}, {1}, {2}, False, "Transmision rate"}
};
```

Then, running "epiModel", files "Model.data" and "parameters.data" appear. Now, in order to obtain the desired system of differential equations, we have to assign the following values to parameters:  $c[2] = c[4] = 0$ ,  $\mu[2] = \mu[4] = 0$ ,  $\beta[1, 1] = \beta[1, 2] = 0$ ,  $\beta[2, 1] = \beta[2, 2] = 0$ ,  $\beta[3, 3] = \beta[3, 4] = 0$  and  $\beta[4, 3] = \beta[4, 4] = 0$ . Thus, we obtain the system:

$$\begin{aligned}
x[1, 1]'[t] &== \mu[1] + (-c[1] - d[1])x[1, 1][t] \\
&\quad -\beta[1, 3]x[1, 1][t]x[2, 3][t] - \beta[1, 4]x[1, 1][t]x[2, 4][t] \\
x[2, 1]'[t] &== (-c[1] - d[1] - \nu[1])x[2, 1][t] \\
&\quad +\beta[1, 3]x[1, 1][t]x[2, 3][t] + \beta[1, 4]x[1, 1][t]x[2, 4][t] \\
x[3, 1]'[t] &== \nu[1]x[2, 1][t] + (-c[1] - d[1])x[3, 1][t] \\
x[1, 2]'[t] &== c[1]x[1, 1][t] - d[2]x[1, 2][t] \\
x[2, 2]'[t] &== c[1]x[2, 1][t] + (-d[2] - \nu[2])x[2, 2][t] \\
x[3, 2]'[t] &== \nu[2]x[2, 2][t] + c[1]x[3, 1][t] - d[2]x[3, 2][t] \\
x[1, 3]'[t] &== \mu[3] + (-c[3] - d[3])x[1, 3][t] \\
x[2, 3]'[t] &== (-c[3] - d[3] - \nu[3])x[2, 3][t] \\
x[3, 3]'[t] &== \nu[3]x[2, 3][t] + (-c[3] - d[3])x[3, 3][t] \\
x[1, 4]'[t] &== c[3]x[1, 3][t] - d[4]x[1, 4][t] \\
x[2, 4]'[t] &== c[3]x[2, 3][t] + (-d[4] - \nu[4])x[2, 4][t] \\
x[3, 4]'[t] &== \nu[4]x[2, 4][t] + c[3]x[3, 3][t] - d[4]x[3, 4][t] \\
x[1, 1][t_0] &== cIni[1] \\
x[2, 1][t_0] &== cIni[2] \\
x[3, 1][t_0] &== cIni[3] \\
x[1, 2][t_0] &== cIni[4] \\
x[2, 2][t_0] &== cIni[5] \\
x[3, 2][t_0] &== cIni[6] \\
x[1, 3][t_0] &== cIni[7] \\
x[2, 3][t_0] &== cIni[8] \\
x[3, 3][t_0] &== cIni[9] \\
x[1, 4][t_0] &== cIni[10] \\
x[2, 4][t_0] &== cIni[11] \\
x[3, 4][t_0] &== cIni[12]
\end{aligned}$$

where  $x[i, j][t]$  is the susceptible subpopulation for  $i = 1$ , infectious for  $i = 2$  and recovered for  $i = 3$ , and age group 1 females for  $j = 1$ , age group 2 females for  $j = 2$ , age group 1 males for  $j = 3$  and age group 2 males for  $j = 4$ .

## A.5 Conclusions

In this appendix, we have presented a *Mathematica* code that translates the description of a type-epidemiological linear or quadratic compartmental model in a

simple syntax into a system of differential equations. The obtained system can be used to estimate parameters, simulate different scenarios or predict short and long-term behavior, as we used to model the academic performance of Spanish and German students in high school in Chapters 3 and 4.

This code is easy to use, saves time building the systems and avoids errors. Moreover, it can be applied to models involving age groups and/or gender. It is particularly interesting when we have to handle a large number of groups. You can test it, changing the variable `\[NTilde]` by 100 (one hundred one-year age groups), in any of the shown examples in Section 4, thus generating 300 equations.

*epiModel* is available at <http://epimodel.imm.upv.es>.



# Appendix B

## Validation of our Spanish mathematical model results

| Academic year  | First Stage of <i>Bachillerato</i><br>(Girls   Boys) |   | Second Stage of <i>Bachillerato</i><br>(Girls   Boys) |   |
|--|--|---|---|---|
|  | % Promote<br>( $G_1$   $B_1$ )                       | % Do not promote<br>( $\bar{G}_1$   $\bar{B}_1$ ) | % Promote<br>( $G_2$   $B_2$ )                        | % Do not promote<br>( $\bar{G}_2$   $\bar{B}_2$ ) |
| Predictions mathematical model (Chapter 2)                             |  |   |   |   |
| 2009 – 2010  | 0.21197   0.16033                                    | 0.08233   0.08537                                 | 0.18945   0.13692                                     | 0.06785   0.06578                                 |
| 2010 – 2011  | 0.21246   0.16063                                    | 0.08184   0.08508                                 | 0.19073   0.13783                                     | 0.06657   0.06487                                 |
| 2011 – 2012  | 0.21287   0.16087                                    | 0.08143   0.08483                                 | 0.19187   0.13865                                     | 0.06543   0.06406                                 |
| 2012 – 2013  | 0.21324   0.16109                                    | 0.08106   0.08461                                 | 0.19289   0.13939                                     | 0.06441   0.06331                                 |
| 2013 – 2014  | 0.21360   0.16129                                    | 0.08070   0.08441                                 | 0.19386   0.14010                                     | 0.06344   0.06260                                 |
| 2014 – 2015  | 0.21397   0.16150                                    | 0.08033   0.08420                                 | 0.19480   0.14080                                     | 0.06250   0.06190                                 |
| Predictions mathematical model (Chapter 3)                             |  |   |   |   |
| 2009 – 2010  | 0.20380   0.15630                                    | 0.06920   0.06940                                 | 0.19400   0.15880                                     | 0.07690   0.07150                                 |
| 2010 – 2011  | 0.20310   0.15670                                    | 0.06800   0.06730                                 | 0.19570   0.16430                                     | 0.07640   0.07050                                 |
| 2011 – 2012  | 0.20220   0.15700                                    | 0.06690   0.06530                                 | 0.19720   0.16580                                     | 0.07600   0.06960                                 |
| 2012 – 2013  | 0.20130   0.15720                                    | 0.06590   0.06340                                 | 0.19860   0.16920                                     | 0.07560   0.06870                                 |
| 2013 – 2014  | 0.20030   0.15750                                    | 0.06500   0.06160                                 | 0.19990   0.17250                                     | 0.07540   0.06790                                 |
| 2014 – 2015  | 0.19910   0.15780                                    | 0.06420   0.05980                                 | 0.20100   0.17570                                     | 0.07520   0.06720                                 |
| 95% confidence intervals of predictions mathematical model (Chapter 3) |  |   |   |   |
| 2009 – 2010  | [ 0.17993 , 0.21227 ] [ 0.15286 , 0.18575 ]          | [ 0.06512 , 0.07041 ] [ 0.06574 , 0.07340 ]       | [ 0.16987 , 0.19554 ] [ 0.15632 , 0.17254 ]           | [ 0.07564 , 0.08020 ] [ 0.06852 , 0.07391 ]       |
| 2010 – 2011  | [ 0.17870 , 0.21222 ] [ 0.15283 , 0.18840 ]          | [ 0.06333 , 0.06920 ] [ 0.06346 , 0.07173 ]       | [ 0.16938 , 0.19734 ] [ 0.15965 , 0.17752 ]           | [ 0.07471 , 0.07969 ] [ 0.06728 , 0.07287 ]       |
| 2011 – 2012  | [ 0.17673 , 0.21202 ] [ 0.15291 , 0.19165 ]          | [ 0.06179 , 0.06830 ] [ 0.06137 , 0.07001 ]       | [ 0.16803 , 0.19898 ] [ 0.16287 , 0.18219 ]           | [ 0.07392 , 0.07949 ] [ 0.06613 , 0.07221 ]       |
| 2012 – 2013  | [ 0.17223 , 0.21172 ] [ 0.15361 , 0.19406 ]          | [ 0.06044 , 0.06749 ] [ 0.05941 , 0.06848 ]       | [ 0.16952 , 0.20041 ] [ 0.16616 , 0.18630 ]           | [ 0.07322 , 0.07948 ] [ 0.06513 , 0.07170 ]       |
| 2013 – 2014  | [ 0.17036 , 0.21135 ] [ 0.15372 , 0.19705 ]          | [ 0.05934 , 0.06691 ] [ 0.05772 , 0.06713 ]       | [ 0.16989 , 0.20172 ] [ 0.16943 , 0.19088 ]           | [ 0.07242 , 0.07916 ] [ 0.06426 , 0.07116 ]       |
| 2014 – 2015  | [ 0.16837 , 0.21089 ] [ 0.15377 , 0.19860 ]          | [ 0.05798 , 0.06610 ] [ 0.05586 , 0.06554 ]       | [ 0.16786 , 0.20304 ] [ 0.17236 , 0.19516 ]           | [ 0.07180 , 0.07888 ] [ 0.06333 , 0.07060 ]       |

TABLE B.1: The model output corresponding to the mathematical model shown in both, Chapter 2 and Chapter 3 and the predictions with corresponding 95% confidence intervals obtained in Chapter 3 of the First and Second Stage of *Bachillerato*, in both, state and private high schools all over Spain during academic years 2008 – 2009 to 2014 – 2015.



| Academic year   | First Stage of <i>Bachillerato</i><br>(Girls   Boys) |   | Second Stage of <i>Bachillerato</i><br>(Girls   Boys) |                   |   |       |
|---|--|---|---|-------------------|---|-------|
|   | % Promote<br>( $G_1$   $B_1$ )                       | % Do not promote<br>( $\bar{G}_1$   $\bar{B}_1$ ) | % Promote<br>( $G_2$   $B_2$ )                        |                   | % Do not promote<br>( $\bar{G}_2$   $\bar{B}_2$ ) |       |
| Absolute Error. Deterministic predictions (Chapter 2) vs 95% confidence intervals (Chapter 3) |  |   |   |                   |   |       |
| 2009 – 2010   | –   –  | 0.01192   0.01197                                 | –   0.01940   | 0.00780   0.00274 |   |       |
| 2010 – 2011   | 0.00024   –  | 0.01264   0.01335                                 | –   0.02182   | 0.00814   0.00242 |   |       |
| 2011 – 2012   | 0.00085   –  | 0.01313   0.01482                                 | –   0.02423   | 0.00849   0.00208 |   |       |
| 2012 – 2013   | 0.00152   –  | 0.01357   0.01613                                 | –   0.02677   | 0.00882   0.00182 |   |       |
| 2013 – 2014   | 0.00225   –  | 0.01380   0.01728                                 | –   0.02933   | 0.00898   0.00166 |   |       |
| 2014 – 2015   | 0.00308   –  | 0.01423   0.01865                                 | –   0.03157   | 0.00930   0.00143 |   |       |
| Absolute Error. Deterministic predictions (Chapter 3) vs 95% confidence intervals (Chapter 3) |  |   |   |                   |   |       |
| 2009 – 2010   | –   –  | –   –   | –   –   | –   –             | –   –   | –   – |
| 2010 – 2011   | –   –  | –   –   | –   –   | –   –             | –   –   | –   – |
| 2011 – 2012   | –   –  | –   –   | –   –   | –   –             | –   –   | –   – |
| 2012 – 2013   | –   –  | –   –   | –   –   | –   –             | –   –   | –   – |
| 2013 – 2014   | –   –  | –   –   | –   –   | –   –             | –   –   | –   – |
| 2014 – 2015   | –   –  | –   –   | –   –   | –   –             | –   –   | –   – |

TABLE B.2: Absolute errors corresponding to the distance between the deterministic predictions given in Chapter 2 and 3 (also shown in Table B.1) and the low or high 95% confidence interval extremes stated in Chapter 3 of the First and Second Stage of *Bachillerato*, in both, state and private high schools all over Spain during academic years 2008 – 2009 to 2014 – 2015. The dashes indicate that the deterministic prediction lies inside its corresponding 95% confidence interval.



# Appendix C

## Validation of our German mathematical model results

| Time ( $t$ )  | Level    | Groups             | Model output | Real data | Absolute Error<br> Model output-Real data |
|---------------|----------|--------------------|--------------|-----------|---|
| 2006 – 2007   | Level 11 | Promoted Girls     | 19.37        | 19.37     | 0.00                                      |
|               |          | Non-Promoted Girls | 0.80         | 0.81      | 0.01                                      |
|               |          | Promoted Boys      | 18.23        | 16.05     | 2.18                                      |
|               |          | Non-Promoted Boys  | 0.75         | 0.96      | 0.21                                      |
|               | Level 12 | Promoted Girls     | 15.34        | 18.23     | 2.89                                      |
|               |          | Non-Promoted Girls | 0.25         | 0.75      | 0.50                                      |
|               |          | Promoted Boys      | 16.05        | 14.7      | 1.35                                      |
|               |          | Non-Promoted Boys  | 0.96         | 0.85      | 0.11                                      |
|               | Level 13 | Promoted Girls     | 14.70        | 15.34     | 0.64                                      |
|               |          | Non-Promoted Girls | 0.85         | 0.25      | 0.60                                      |
|               |          | Promoted Boys      | 12.38        | 12.38     | 0.00                                      |
|               |          | Non-Promoted Boys  | 0.31         | 0.31      | 0.00                                      |
| 2007 – 2008   | Level 11 | Promoted Girls     | 18.87        | 19.09     | 0.22                                      |
|               |          | Non-Promoted Girls | 0.69         | 0.67      | 0.02                                      |
|               |          | Promoted Boys      | 18.04        | 15.92     | 2.12                                      |
|               |          | Non-Promoted Boys  | 0.68         | 0.88      | 0.20                                      |
|               | Level 12 | Promoted Girls     | 15.75        | 17.96     | 2.21                                      |
|               |          | Non-Promoted Girls | 0.21         | 0.68      | 0.47                                      |
|               |          | Promoted Boys      | 16.19        | 14.73     | 1.46                                      |
|               |          | Non-Promoted Boys  | 0.93         | 0.81      | 0.12                                      |
|               | Level 13 | Promoted Girls     | 14.82        | 15.96     | 1.14                                      |
|               |          | Non-Promoted Girls | 0.78         | 0.25      | 0.53                                      |
|               |          | Promoted Boys      | 12.78        | 12.77     | 0.01                                      |
|               |          | Non-Promoted Boys  | 0.26         | 0.28      | 0.02                                      |
| 2008 – 2009   | Level 11 | Promoted Girls     | 18.41        | 19.1      | 0.69                                      |
|               |          | Non-Promoted Girls | 0.60         | 0.59      | 0.01                                      |
|               |          | Promoted Boys      | 17.86        | 15.95     | 1.91                                      |
|               |          | Non-Promoted Boys  | 0.62         | 0.81      | 0.19                                      |
|               | Level 12 | Promoted Girls     | 16.09        | 18.15     | 2.06                                      |
|               |          | Non-Promoted Girls | 0.17         | 0.58      | 0.41                                      |
| Promoted Boys | 16.31    | 14.77              | 1.54         |           |   |

|             |          |                    |       |       |      |
|-------------|----------|--------------------|-------|-------|------|
|             |          | Non-Promoted Boys  | 0.90  | 0.67  | 0.23 |
|             | Level 13 | Promoted Girls     | 14.96 | 15.94 | 0.98 |
|             |          | Non-Promoted Girls | 0.72  | 0.19  | 0.53 |
|             |          | Promoted Boys      | 13.16 | 13.04 | 0.12 |
|             |          | Non-Promoted Boys  | 0.22  | 0.21  | 0.01 |
| 2009 – 2010 | Level 11 | Promoted Girls     | 18.49 | 19.24 | 0.75 |
|             |          | Non-Promoted Girls | 0.52  | 0.53  | 0.01 |
|             |          | Promoted Boys      | 17.90 | 16.3  | 1.60 |
|             |          | Non-Promoted Boys  | 0.57  | 0.73  | 0.16 |
|             | Level 12 | Promoted Girls     | 16.18 | 17.77 | 1.59 |
|             |          | Non-Promoted Girls | 0.14  | 0.47  | 0.33 |
|             |          | Promoted Boys      | 16.26 | 14.72 | 1.54 |
|             |          | Non-Promoted Boys  | 0.88  | 0.67  | 0.21 |
|             | Level 13 | Promoted Girls     | 14.99 | 16.25 | 1.26 |
|             |          | Non-Promoted Girls | 0.66  | 0.19  | 0.47 |
|             |          | Promoted Boys      | 13.24 | 12.94 | 0.30 |
|             |          | Non-Promoted Boys  | 0.18  | 0.19  | 0.01 |
| 2010 – 2011 | Level 11 | Promoted Girls     | 18.04 | 18.27 | 0.23 |
|             |          | Non-Promoted Girls | 0.45  | 0.44  | 0.01 |
|             |          | Promoted Boys      | 17.71 | 15.87 | 1.84 |
|             |          | Non-Promoted Boys  | 0.53  | 0.6   | 0.07 |
|             | Level 12 | Promoted Girls     | 16.44 | 18.29 | 1.85 |
|             |          | Non-Promoted Girls | 0.12  | 0.47  | 0.35 |
|             |          | Promoted Boys      | 16.36 | 15.21 | 1.15 |
|             |          | Non-Promoted Boys  | 0.85  | 0.64  | 0.21 |
|             | Level 13 | Promoted Girls     | 15.15 | 16.44 | 1.29 |
|             |          | Non-Promoted Girls | 0.60  | 0.17  | 0.43 |
|             |          | Promoted Boys      | 13.57 | 13.39 | 0.18 |
|             |          | Non-Promoted Boys  | 0.15  | 0.21  | 0.06 |

TABLE C.1: The model output obtained with the estimated parameters (Tables 4.2 and 4.3) in our German model, the real data and their associated absolute errors corresponding the Levels 11, 12 and 13, in both, state and private high schools all over the German region of North Rhine-Westphalia during academic years 2006 – 2007 to 2010 – 2011. Each row shows the percentage of girls/boys who promote and do not promote for each academic level.

# Appendix D

## Time series analysis: Forecasting models in Statgraphics Plus 5.1.

### D.1 Introduction

In this appendix, we show the procedure followed by the *Statgraphics Plus for Windows 5.1* software to obtain the time series models which allow us to predict the necessary information to achieve the aims proposed, in particular, in Chapter 5 of this dissertation.

This powerful statistical tool provides the user five different forecasting models that best fit the available data. Each one of the forecasted models takes a different approach to predict future values with 95% confidence intervals [74]. In the discussion below, the following notation will be used:

$$\begin{aligned} Y_t &= \text{Observed value at time } t, t = 1, \dots, n. \\ n &= \text{Sample size (number of observations used to fit the model)}. \\ m &= \text{Observations have been used to validate the model}. \\ n + m &= \text{Total sample size}. \\ F_t(k) &= \text{Forecast for time } t + k \text{ done at time } t. \\ e_t &= \text{Prediction errors calculated by :} \end{aligned} \tag{D.1}$$

$$e_t = Y_t - F_{t-1}(1). \quad (\text{D.2})$$

Moreover, this software only states the time series model and its corresponding predictions with confidence intervals if the following assumptions are satisfied:

- The proper model have been selected.
- The selected model is valid for all historical data.
- The selected model continues being valid in the future.
- The errors follow a normal distribution [75].

This appendix is organized as follows. In Section D.2, we present a brief explanation of the five forecasting models used by the *Statgraphics Plus for Windows 5.1* software to obtain the best model that fit the available data. The validation of the mentioned models is given in Section D.3. Finally, in Section D.4 we show how the *Statgraphics Plus for Windows 5.1* software obtain the 95% confidence intervals of the corresponding predictions.

## D.2 Forecasting models

In this section, we will present a brief explanation about the five forecasting models shown by this software to select the best fit.

The mentioned forecasting models are the following:

- **Random Walk with Trend.** Randomly forecasts the next observation based on the current observation and the mean and standard deviation of the difference of the values. This model, included a constant, uses the current value of the series to forecast all the future values [73, 74]. This forecast is given by:

$$F_t(k) = Y_t + k\hat{\Delta} \quad (\text{D.3})$$

where  $\widehat{\Delta}$  estimates the average change from one period to another.

- **Linear Trend.** Fits a straight line through the data and into the forecasting periods. This model estimates a regression model to the available data, using time as a independent variable. It is fitted by least squares [74, 81, 82]. The forecasts of the model are obtained by:

$$F_t(k) = \widehat{a} + \widehat{b}(t + k) \quad (\text{D.4})$$

where  $\widehat{a}$  and  $\widehat{b}$  are estimated constants.

- **Simple Moving Average.** Uses the moving average to smooth the data and to predict future values. This model uses the average of the most recent  $m$  observations to predict future values. The forecasts are given by:

$$F_t(k) = \frac{\sum_{i=0}^{m-1} Y_{t-i}}{m} \quad (\text{D.5})$$

- **Simple Exponential Smoothing.** Smooths the data and predicts future values by exponentially weighting the values in the time series [74, 83]. Let  $S'$  denote the singly-smoothed series obtained by applying simple exponential smoothing to series  $Y$ . That is, the value of  $S'$  at period  $t$  is given by:

$$S'_t = \alpha Y_t + (1 - \alpha)S'_{t-1}, \quad 0 < \alpha < 1 \quad (\text{D.6})$$

where  $\alpha$  is the "smoothing constant" ( $\alpha$  number between 0 and 1).

Therefore, the forecasts are given by:

$$F_t(k) = S'_t \quad (\text{D.7})$$

- **Brown's Linear Exponential Smoothing.** This model is similar to the *Simple Exponential Smoothing*, although in this case, smooths the data and

predicts future values by applying a double-smoothing formula to the data using one parameter,  $\alpha$  [74, 83]. Now, let  $S''$  denote the doubly-smoothed series obtained by applying simple exponential smoothing (equation D.6), that is,

$$S_t'' = \alpha S_t'(k) + (1 - \alpha)S_{t-1}'', \quad 0 < \alpha < 1 \quad (\text{D.8})$$

where  $\alpha$  is the "smoothing constant" ( $\alpha$  number between 0 and 1).

Therefore, the forecasts are given by:

$$F_t(k) = 2S_t' - S_t'' + k \frac{\alpha}{1 - \alpha} (S_t' - S_t'') \quad (\text{D.9})$$

### D.3 Validation of the model

Once the models are stated, the software considers  $m$  observations of the available data to validate the model (see Section D.1). This process is addressed taking as a reference the minimum error generated by the forecasted model according to the available data. The indicators considered for this validation are:

- RMSE: Root Mean Square Error over the validation period, given by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^m e_{n+i}^2}{m}} \quad (\text{D.10})$$

- MAPE: Percentage of the mean absolute error on the validation period, given by:

$$MAPE = 100 \frac{\sum_{i=1}^m \frac{|e_{n+i}|}{Y_{t+i}}}{m} \% \quad (\text{D.11})$$



The RMSE estimated standard deviation forecast errors a step forward. The MAPE estimates the average percentage forecast error one step ahead. The small values are desirable for RMSE and MAPE, respectively.

## D.4 Obtaining 95% confidence intervals

In this section, we show how the *Statgraphics Plus for Windows 5.1* software provides the  $100(1 - \alpha)\%$  confidence intervals for each forecast model.

They are computed assuming that the errors in the model follow a normal distribution [74]. The confidence intervals are given by:

$$F_t(k) \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{V}(k)}, \quad (\text{D.12})$$

where  $\widehat{V}(k)$  is the estimated variance of the forecast of  $k$  periods before the end of the data.



# Bibliography

- [1] Instituto Nacional de Evaluación Educativa. Gobierno de España. Sistema estatal de indicadores de la educación. [Education indicator of the Spanish Government], 2012. URL [http://www.mecd.gob.es/inee/publicaciones/indicadores-educativos/Sistema-Estatal.html#SEIE\\_2011\\_2](http://www.mecd.gob.es/inee/publicaciones/indicadores-educativos/Sistema-Estatal.html#SEIE_2011_2).
- [2] Ministerio de Educación. Gobierno de España. Enseñanzas no universitarias. Alumnado matriculado. [Non-university education. Registered students], 2013. URL <http://www.mecd.gob.es/horizontales/estadisticas/no-universitaria/alumnado/matriculado.html>.
- [3] Scientific United Nations Educational and Cultural Organization. *Youth and Skills: Putting Education to Work*. EFA Global Monitoring Report. UNESCO Publishing, 2012. ISBN 9789231042409. URL <http://www.unesco.org/new/en/education/themes/leading-the-international-agenda/efareport/reports/2012-skills>.
- [4] EuroStat. Comision European. Education statistics at regional level, 2013. URL [http://epp.eurostat.ec.europa.eu/statistics\\_explained/index.php/Education\\_statistics\\_at\\_regional\\_level](http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Education_statistics_at_regional_level).
- [5] A. Marchesi and C.H. Gil. *El fracaso escolar. [The academic underachievement]*, chapter Significado del fracaso escolar en España. [Meaning of the academic underachievement in Spain], pages 29–54. Fundación por la Modernización de España, 2000. ISBN 9788489796218.
- [6] J.M. Puig. *El Fracaso Escolar: Una Perspectiva Internacional. [Academic Underachievement: An International Perspective]*, chapter Educación en valores y fracaso escolar. [Education in values and academic underachievement]. In *Alianza Ensayo* Marchesi and Gil [15], 2003. ISBN 9788420629551.

- 
- [7] V. Martínez-Otero. Diversos condicionantes del fracaso escolar en la educación secundaria. [Different conditions of academic underachievement in secondary education]. *Revista Iberoamericana de Educación.*, (51):67–85, 2009. URL <http://dialnet.unirioja.es/servlet/articulo?codigo=3157591>.
- [8] M.J. San Segundo and A. Vaquero. Descentralización educativa y programas nacionales de mejora. [Decentralization of education and national improvement programs]. In *Proceedings of the XVI Jornadas de la Asociación de la Economía de la Educación*, pages 1–17, Spain, 2008. URL [http://www.congresos.ulpgc.es/aeet\\_aede/Descargas/Sesion3Sala5/SanSegundo-Vaquero.pdf](http://www.congresos.ulpgc.es/aeet_aede/Descargas/Sesion3Sala5/SanSegundo-Vaquero.pdf).
- [9] Ministerio de Educación. Gobierno de España. Ley Orgánica 1/1990, de 3 de Octubre, de Ordenación General del Sistema Educativo. Boletín Oficial de Estado de 4 de Octubre de 1990. [Organic Law 1/1990 of October 3, on the General Educational System. Official State Bulletin on October 4, 1990], 1990. URL [http://www.boe.es/aeboe/consultas/bases\\_datos/doc.php?id=BOE-A-1990-24172](http://www.boe.es/aeboe/consultas/bases_datos/doc.php?id=BOE-A-1990-24172).
- [10] Ministerio de Educación. Gobierno de España. Ley Orgánica 2/2006, de 3 de Mayo, de Educación. Boletín Oficial de Estado, 106, de 4 de Mayo de 2006. [Organic Law 2/2006 of May 3, on Education. Official State Bulletin, 106, of May 4, 2006], 2006. URL [http://www.boe.es/aeboe/consultas/bases\\_datos/doc.php?id=BOE-A-2006-7899](http://www.boe.es/aeboe/consultas/bases_datos/doc.php?id=BOE-A-2006-7899).
- [11] A. Bolívar Botía and L. López Calvo. Las grandes cifras del fracaso y los riesgos de exclusión educativa. [Large numbers of failure and risk of educational exclusion]. *Profesorado, Revista de Currículum y Formación de Profesorado*, 13(3):291–304, 2009. URL <http://www.ugr.es/~recfpro/rev133ART2.pdf>.
- [12] Forum Libertas. Récorde europeos de fracaso escolar en España. [European records of academic underachievement in Spain], 2011. URL <http://www.rlp.com.ni/noticias/93005>.
- [13] M. Fernández Enguita. La Educación como servicio público: Estado, Mercado y Profesión. [Education as a public service: State, Market and Occupation]. In *Jornadas educativas. Alternativas para la calidad de la enseñanza. FUHEM*, pages 17–23, Madrid, Spain, 2000. URL <http://www.fuhem.es/educacion/>.

- 
- [14] R.B. Reich. *The Work of Nations: Preparing Ourselves for 21st Century Capitalis*. Vintage. Knopf Doubleday Publishing Group, 2010. ISBN 9780307772992.
- [15] A. Marchesi and C.H. Gil. *El Fracaso Escolar: Una Perspectiva Internacional. [Academic Underachievement: An International Perspective]*. Alianza Ensayo. Alianza Editorial, 2003. ISBN 9788420629551.
- [16] EurActiv. EU youth job strategy under fire, 2011. URL <http://www.euractiv.com/innovation-enterprise/eu-youth-job-strategy-fire-news-497858>.
- [17] J. Casal, M. García, and J. Planas. Las reformas en los dispositivos de formación para combatir el fracaso escolar en Europa: paradojas de un éxito. [Reforms in training measures to face academic underachievement in Europe: paradoxes of a success]. *Revista de Educación*, (317):301–317, 1998. URL <http://www.doredin.mec.es/documentos/00820073004064.pdf>.
- [18] H. Eckert. Entre el fracaso escolar y las dificultades de inserción profesional: la vulnerabilidad de los jóvenes sin formación en el inicio de la sociedad del conocimiento. [Between academic underachievement and employability difficulties: the vulnerability of young people without training in the beginning of the knowledge society]. *Revista de Educación*, (341):35–55, 2006. URL [http://www.oei.es/etp/fracaso\\_escolar\\_dificultades\\_insercion\\_profesional\\_eckert.pdf](http://www.oei.es/etp/fracaso_escolar_dificultades_insercion_profesional_eckert.pdf).
- [19] G. Psacharopoulos. The costs of school failure: A feasibility study. *European Expert Network on Economics of Education (EENEE)*, 2007. URL [http://progetto14.itipacinotti.it/uploads/2.1\)%20The%20costs%20of%20%20School%20Failure%20.pdf](http://progetto14.itipacinotti.it/uploads/2.1)%20The%20costs%20of%20%20School%20Failure%20.pdf).
- [20] J. Calero Martínez, M. Gil Izquierdo, and M. Fernández Gutiérrez. *Los costes del abandono escolar prematuro (Recurso electrónico): una aproximación a las pérdidas monetarias y no monetarias causadas por el abandono prematuro en España. [The costs of early school abandon (Electronic resource): an approach to the monetary and nonmonetary losses caused by early abandon in Spain]*. Investigación. IFIIE (Instituto de Formación del Profesorado, Investigación e Innovación Educativa). Gobierno de España); 191. Ministerio de Educación, Subdirección General de Documentación y Publicaciones, 2011.

- ISBN 9788436952087. URL [http://biblioteca-central.educacion.gob.es/search~S0\\*spi?/XCALERO+MART{%u00CD}NEZ&SORT=D/XCALERO+MART{%u00CD}NEZ&SORT=D&SUBKEY=CALERO+MART%C3%8DNEZ/1%2C15%2C15%2CB/frameset&FF=XCALERO+MART{%u00CD}NEZ&SORT=D&2%2C2%2C](http://biblioteca-central.educacion.gob.es/search~S0*spi?/XCALERO+MART{%u00CD}NEZ&SORT=D/XCALERO+MART{%u00CD}NEZ&SORT=D&SUBKEY=CALERO+MART%C3%8DNEZ/1%2C15%2C15%2CB/frameset&FF=XCALERO+MART{%u00CD}NEZ&SORT=D&2%2C2%2C).
- [21] Instituto Nacional de Estadística. Mujeres y hombres en españa. [Women and men in Spain], 2010. URL <http://www.ine.es>.
- [22] Instituto Nacional de Evaluación Educativa. Ministerio de Educación. Gobierno de España. Panorama de la educación. indicadores de la OCDE 2012. [Education at a Glance. OECD Indicators 2012], 2012. URL <http://www.mecd.gob.es/dctm/inee/internacional/panorama2012.pdf?documentId=0901e72b81415d28>.
- [23] A. Marchesi and R. Lucena. *El Fracaso Escolar: Una Perspectiva Internacional*. [Academic Underachievement: An International Perspective], chapter La Representación Social del Fracaso Escolar. [The Social Representation of Academic Underachievement]. In *Alianza Ensayo* Marchesi and Gil [15], 2003. ISBN 9788420629551.
- [24] J.I. Pozo. *Teorías Cognitivas Del Aprendizaje*. [Cognitive Theories of Learning]. Colección Pedagogía. Ediciones Morata, 1989. ISBN 9788471123350.
- [25] A. Tryphon and J.J. Vonèche. *Piaget-Vygotsky: La Génesis Social del Pensamiento*. [Piaget-Vygotsky: The Social Genesis of the Thought]. Paidós Educador. Ediciones Paidós Iberica, S.A., 2001. ISBN 9789501221503.
- [26] L.S. Vygotski and M. Cole. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, 1978. ISBN 9780674576292.
- [27] M. Lucci. La propuesta de Vygotsky: La psicología socio-histórica. [The proposal of Vygotsky: Socio-historical psychology]. *Revista de Currículum y Formación del Profesorado*, 10(2):7–11, 2006. ISSN 1138-414X. URL <http://hdl.handle.net/10481/17420>.
- [28] H. Daniels, M. Cole, and J.V. Wertsch. *The Cambridge Companion to Vygotsky*. Cambridge Collections misc. Cambridge University Press, 2007. ISBN 9780521831048.

- 
- [29] N.A. Christakis and J.H. Fowler. *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. Little Brown and Company, 2009. ISBN 9780316072588.
- [30] N.A. Christakis and J.H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379, 2007. doi: 10.1056/NEJMsa066082. URL <http://www.nejm.org/doi/full/10.1056/NEJMsa066082>. PMID: 17652652.
- [31] M.Á.M. Zabal, P.F. Berrocal, and C. Coll. *La Interacción Social en Contextos Educativos. [Social Interaction in Educational Contexts]*. Psicología. Siglo XXI de España, 1995. ISBN 9788432308642.
- [32] F.J. Santonja, A. Morales, R.J. Villanueva, and J.C. Cortés. Analysing the effect of public health campaigns on reducing excess weight: A modelling approach for the Spanish Autonomous Region of the Community of Valencia. *Plann Programa Eval*, 35.
- [33] F.J. Santonja, Sánchez E., M. Rubio, and J. Morera. Alcohol consumption in Spain and its economic cost: A mathematical modeling approach. *Mathematical and Computer Modelling*, 52(7–8):999–1003, 2010. ISSN 0895–7177. doi: 10.1016/j.mcm.2010.02.029. URL <http://www.sciencedirect.com/science/article/pii/S0895717710000932>.
- [34] 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. doi: 10.3109/09687630903443299. URL <http://informahealthcare.com/doi/abs/10.3109/09687630903443299>.
- [35] I. García, L. Jódar, P. Merello, and F.J. Santonja. A discrete mathematical model for addictive buying: Predicting the affected population evolution. *Mathematical and Computer Modelling*, 54(7–8):1634–1637, 2011. ISSN 0895–7177. doi: 10.1016/j.mcm.2010.12.012. URL <http://www.sciencedirect.com/science/article/pii/S0895717710005947>.
- [36] L.M.A. Bettencourt, A. Cintrón-Arias, D.I. Kaiser, and C. Castillo-Chávez. The power of a good idea: quantitative modeling of the spread of ideas from epidemiological models. *Physica A: Statistical Mechanics and its Applications*,

- (364):513–536, 2006. doi: 10.1016/j.physa.2005.08.083. URL <http://dx.doi.org/10.1016/j.physa.2005.08.083>.
- [37] 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*, 57(7–8):1679–1685, 2013. ISSN 0895–7177. doi: 10.1016/j.mcm.2011.11.007. URL <http://www.sciencedirect.com/science/article/pii/S0895717711006923>.
- [38] M.V. García Jiménez, A. Jiménez Blanco, and J.M. Alvarado Izquierdo. La predicción del rendimiento académico: regresión lineal versus regresión logística. [The prediction of the academic underachievement: linear regression vs. logistic regression]. *Psicothema*, 12(2):248–252, 2000.
- [39] C. Inglés, G. Benavides, J. Redondo, J.M. García-Fernández, C. Ruiz-Esteban, C. Estévez, and E. Huescar. Conducta prosocial y rendimiento académico en estudiantes españoles de Educación Secundaria Obligatoria. [Pro-social behavior and academic performance in Spanish students of ESO]. *Anales de Psicología*, 25(1):93–101, 2009. ISSN 1695-2294. URL <http://revistas.um.es/analesps/article/view/71541>.
- [40] C. Cunchillos and F. Rodríguez. El fracaso escolar, su cuantificación y su distribución social en la Comunidad de Madrid. [Academic underachievement, its quantification and social distribution in the Community of Madrid]. 2004. URL <http://www.fracasoescolar.com/conclusions2004/cunchillos.pdf>.
- [41] Ministerio de Educación. Gobierno de España. Enseñanzas no universitarias. Alumnado. Resultados académicos. [Non-university education. Registered students. Academic results], 2013. URL <http://www.mecd.gob.es/horizontales/estadisticas/no-universitaria/alumnado/resultados.html>.
- [42] M.F. Enguita, J. Rivière, L.M. Martínez, and J.R. Gómez. *Fracaso y abandono escolar en España. [Academic underachievement and abandon in Spain]*. Estudios Sociales. Fundación La Caixa, 2010. ISBN 9788469331415.
- [43] M. Alcaide Risotto. Autoconcepto y rendimiento académico en alumnos de 1º de Bachillerato según el género. [Self-concept and academic achievement in



- students of 1st Bachillerato by gender]. *Revista Electrónica de Investigación y Docencia (REID)*, (2):27–44, 2009.
- [44] C.G. Bueno. *Identidades de género y feminización del éxito académico*. [Gender identity and feminization of academic success]. Investigación. CIDE (Centro de Investigación y Documentación Educativa). Ministerio de Educación, Cultura y Deporte, Subdirección General de Información y Publicaciones, 2001. ISBN 9788436935004. URL <http://www.mec.es/cide/index.htm>.
- [45] F. Requena Santos. Género, redes de amistad y rendimiento académico. [Gender, friendship networks and academic performance]. *Revista de sociología*, (56):233–242, 1998. ISSN 0210-2862.
- [46] J.H. González. Las relaciones afectivas en el Bachillerato como parte de la identidad estudiantil. [Emotional relationships in High School as part of the student identity]. URL <http://www.comie.org.mx/congreso/memoriaelectronica/v09/ponencias/at16/PRE1178725641.pdf>.
- [47] L.C. Rigsby and E.L. McDill. Adolescent peer influence processes: Conceptualization and measurement. *Social Science Research*, 1(3):305–321, 1972. ISSN 0049-089X. doi: 10.1016/0049-089X(72)90079-8. URL <http://www.sciencedirect.com/science/article/pii/0049089X72900798>.
- [48] Instituto Nacional de Estadística. Encuesta sobre Gasto de los Hogares en Educación. (Módulo Piloto de la Encuesta de Presupuestos Familiares 2007). [Survey of Household Spending on Education. (Module Pilot Household Budget Survey 2007)], 2009. URL <http://www.ine.es/prensa/np541.pdf>.
- [49] Instituto Nacional de Estadística. Encuesta sobre Gasto de los Hogares en Educación. (Módulo Piloto de la Encuesta de Presupuestos Familiares (Curso 2011/2012)). [Survey of Household Spending on Education. (Module Pilot Household Budget Survey (Course 2011/2012))], 2012. URL <http://www.ine.es/prensa/np763.pdf>.
- [50] J.D. Murray. *Mathematical Biology: I. An Introduction*. Interdisciplinary Applied Mathematics. Springer, 2002. ISBN 9780387952239.
- [51] C. Fierro-Hernández. Patrón de rasgos personales y comportamiento escolar en jóvenes. [Personal trait patterns and scholar behaviour]. *Revista de Educación*, (332):291–304, 2000. ISSN 0034-8082. URL

- [http://www.uma.es/psicologia/docs/eudemon/investigacion/patron\\_de\\_rasgos\\_en\\_escolares.pdf](http://www.uma.es/psicologia/docs/eudemon/investigacion/patron_de_rasgos_en_escolares.pdf).
- [52] Wolfram. URL <http://www.wolfram.com/mathematica/>.
- [53] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965. doi: 10.1093/comjnl/7.4.308.
- [54] M. Martcheva and C. Castillo-Chavez. Diseases with chronic stage in a population with varying size. *Mathematical Biosciences*, 182(1):1–25, 2003. ISSN 0025-5564. doi: 10.1016/S0025-5564(02)00184-0. URL <http://www.sciencedirect.com/science/article/pii/S0025556402001840>.
- [55] J. Mena-Lorcat and H.W. Hethcote. Dynamic models of infectious diseases as regulators of population sizes. *Journal of Mathematical Biology*, 30:693–716, 1992. ISSN 0303–6812. doi: 10.1007/BF00173264. URL <http://dx.doi.org/10.1007/BF00173264>.
- [56] 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 Applications*, 56(3):836 – 846, 2008. ISSN 0898-1221. doi: 10.1016/j.camwa.2008.01.001. URL <http://www.sciencedirect.com/science/article/pii/S0898122108000497>.
- [57] J.C. Cortés, M. Ehrhardt, A. Sánchez-Sánchez, F.J. Santonja, and R.J. Villanueva. Scaling type-epidemiological models, 2013. URL <http://scaling.imm.upv.es>.
- [58] B. Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 7(1):1–26, 1979.
- [59] G. Dogan. Bootstrapping for confidence interval estimation and hypothesis testing for parameters of system dynamics models. *System Dynamics Review*, 23(4):415–436, 2007. ISSN 1099-1727. doi: 10.1002/sdr.362. URL <http://dx.doi.org/10.1002/sdr.362>.
- [60] M. Joshi, A. Seidel-Morgenstern, and A. Kremling. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metabolic Engineering*, 8(5):447–455, 2006. ISSN 1096-7176. doi: 10.1016/j.ymben.2006.04.003. URL <http://www.sciencedirect.com/science/article/pii/S1096717606000243>.

- [61] W.C.M. Van Beers and J.P.C. Kleijnen. Customized sequential designs for random simulation experiments: Kriging metamodeling and bootstrapping. *European Journal of Operational Research*, 186(3):1099–1113, 2008. ISSN 0377-2217. doi: 10.1016/j.ejor.2007.02.035. URL <http://www.sciencedirect.com/science/article/pii/S0377221707002895>.
- [62] G.M. Ljung and G.E.P. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978. doi: 10.1093/biomet/65.2.297. URL <http://biomet.oxfordjournals.org/content/65/2/297.abstract>.
- [63] G.J. Szekely and M.L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005. ISSN 0047-259X. doi: 10.1016/j.jmva.2003.12.002. URL <http://www.sciencedirect.com/science/article/pii/S0047259X03002124>.
- [64] J. Calero, A. Choi, and S. Waisgrais. Determinantes del riesgo de fracaso escolar en España: una aproximación a través de un análisis logístico multinivel aplicado a PISA-2006. [Risk factors of academic underachievement in Spain: an approach through a multilevel logistic analysis applied to PISA-2006]. *Revista de Educación*, (Número extraordinario 2010):225–256, 2010. URL [http://www.revistaeducacion.mec.es/re2010/re2010\\_09.pdf](http://www.revistaeducacion.mec.es/re2010/re2010_09.pdf).
- [65] Ministerium für Inneres und Kommunales des Landes Nordrhein-Westfalen. Geltende gesetze und verordnungen. [Applicable Laws and Regulations], 2012. URL [https://recht.nrw.de/lmi/owa/br\\_bes\\_text?anw\\_nr=2&gld\\_nr=2&ugl\\_nr=223&bes\\_id=7345&aufgehoben=N&menu=1&sg=0](https://recht.nrw.de/lmi/owa/br_bes_text?anw_nr=2&gld_nr=2&ugl_nr=223&bes_id=7345&aufgehoben=N&menu=1&sg=0).
- [66] B. Landner. Bildungsreport Nordrhein-Westfalen, 2010. [Education Report North Rhine-Westphalia, 2010]. *Statistische Analysen und Studien*, (68), 2011. ISSN 1619-506X. URL <https://webshop.it.nrw.de/gratis/Z08920201054.pdf>.
- [67] Regionalverband Ruhr. *Bildungsbericht Ruhr*. Waxmann, 2012. ISBN 9783830926313.
- [68] Kommunales Bildungsmonitoring: Tab. D13.2 Bestand an Schülerinnen und Schülern sowie Klassenwiederholungen. [Municipal Education Monitoring: Table D13.2 Inventory of students and class repetition], 2012. URL <http://www.dipf.de/de/projekte/pdf/steufi/kbm-\handreichung-wie-erstellt-man-einen-kommunalen-bildungsbericht>.

- 
- [69] H. Akaike. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1):243–247, 1969. URL <http://ideas.repec.org/a/spr/aistmt/v21y1969i1p243-247.html>.
- [70] P.J. Brockwell and R.A. Davis. *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer, 2002. ISBN 9780387953519.
- [71] The R Project for Statistical Computing. URL <http://www.r-project.org/index.html>.
- [72] P. J. Brockwell. *Time Series Analysis*. John Wiley and Sons, Ltd, 2008. ISBN 9780470061572. doi: 10.1002/9780470061572.eqr229. URL <http://dx.doi.org/10.1002/9780470061572.eqr229>.
- [73] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics. Wiley, 2008. ISBN 9780470272848.
- [74] Statgraphics.Net. Statgraphics tutorials, 2013. URL <http://www.statgraphics.net/wp-content/uploads/2011/12/tutoriales/Pronosticos.pdf>.
- [75] J.F. Hair and R.E. Anderson. *Multivariate data analysis*. Prentice Hall, 2010. ISBN 9780138132637.
- [76] W.O. Kermack and A.G. McKendrick. Contributions to the mathematical theory of epidemics-i. *Bulletin of Mathematical Biology*, 53(115A):33–55, 1991. ISSN 0092-8240. doi: 10.1016/S0092-8240(05)80040-0. URL <http://www.sciencedirect.com/science/article/pii/S0092824005800400>.
- [77] E.H. Elbasha, E.J. Dasbach, and R.P. Insinga. Model for assessing human papillomavirus vaccination strategies. *Emerging infectious diseases*, 13(1):28–41, 2007. URL <http://europepmc.org/abstract/MED/17370513>.
- [78] C.L. Trotter, N.J. Gay, and W.J. Edmunds. The natural history of meningococcal carriage and disease. *Epidemiol Infect.*, 134(3):556–66, 2006.
- [79] E. Beretta and V. Capasso. On the general structure of epidemic systems. Global asymptotic stability. *Computers and Mathematics with Applications*, (12A):677–694, 1986.

- [80] V. Capasso. *Mathematical Structures of Epidemic Systems*. Lecture Notes in Biomathematics. Springer, 2008. ISBN 9783540565260.
- [81] S.M. Fernández. *Guía completa de Statgraphics: Desde MS-DOS a Statgraphic Plus*. [Statgraphics Complete Guide: From MS-DOS to Statgraphic Plus]. Díaz de Santos, 2001. ISBN 9788479784980.
- [82] J. Nyblom. Testing for deterministic linear trend in time series. *Journal of the American Statistical Association*, 81(394):545–549, 1986. doi: 10.1080/01621459.1986.10478302. URL <http://amstat.tandfmisc.com/doi/abs/10.1080/01621459.1986.10478302>.
- [83] J.F. Muth. Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, 55(290):299–306, 1960. doi: 10.1080/01621459.1960.10482064. URL <http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1960.10482064>.