Bergische Universität Wuppertal

Fachbereich Mathematik und Naturwissenschaften

Institute of Mathematical Modelling, Analysis and Computational Mathematics (IMACM)

Preprint BUW-IMACM 12/26

J.-C. Cortés, M. Ehrhardt, A. Sánchez-Sánchez, 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

October 2012

http://www.math.uni-wuppertal.de

International Journal of Computer Mathematics
Vol. 00, No. 00, January 2008, 1–10

RESEARCH ARTICLE

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

J.-C. Cortés[a], M. Ehrhardt[b], A. Sánchez-Sánchez[a] * , F.-J. Santonja[c] and R.-J. Villanueva[a]

[a]*Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València, Spain;* [b]*Fachbereich C - Mathematik und Naturwissenschaften, Angewandte Mathematik und Numerische Analysis, Bergische Universität Wuppertal, Germany;* [c]*Departamento de Estadística e Investigación Operativa, Universitat de València, Spain*

(*v3.6 released October 2012*)

Academic underachievement is a concern of paramount importance in Europe, where around 15% of the students in the last courses in high school do not achieve the minimum knowledge academic requirement. In this paper, we propose a model based on a system of differential equations to study the dynamics of the students academic performance in the German region of North Rhine-Westphalia. This approach is supported by the idea that both, good and bad study habits, are a mixture of personal decisions and influence of classmates. This model allows us to forecast the academic performance by means of confidence intervals over the next few years.

Keywords: Academic Performance; Modelling; Non-linear System of Differential Equations; Forecasting in Social Sciences; Bootstrapping.

AMS Subject Classification: 97M10; 34K60.

1. Introduction

In many countries of the European Union, in the last courses of high school, the rates of academic underachievement are at very worrying levels [1–3]. The concern about the high level of academic underachievement is completely justified, not only by the high rates but also by the negative effects on the country's economic development, especially in the unemployment and its serious consequences.

According to the Vygotsky learning theories [4] and the recent studies published by Christakis and Fowler [5], habits and behavior may be socially transmitted, in particular, academic and study habits.

Taking into account this approach, in this paper, we focus on the German region of North Rhine-Westphalia and propose a model to study the evolution of the students academic performance in the last three courses of the high school (levels 11, 12 and 13) before accessing to the university. To do that, we use mathematical epidemiology and statistical techniques. This approach could be of great

*Corresponding author. Email: alsncsnc@posgrado.upv.es

ISSN: 0020-7160 print/ISSN 1029-0265 online

DOI: 0020716YYxxxxxxxx
http://www.informaworld.com

Table 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.

	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

interest because a new plan of study will come into force next year in North Rhine-Westphalia. The predictions of the academic results using confidence intervals could be compared with the real ones corresponding to the new curriculum in order to evaluate if the change has been as good as expected.

Some examples of social problems using type-epidemiological mathematical models are encountered in obesity [6], alcoholism [7], drug abuse [8], shopaholism [9], spread of ideas [10], and so on.

2. Model building

2.1 *Available data*

According to the data, we say that a student *promotes* if, in case the course finishes now, he or she will pass to the next level or graduate satisfying the current legislation into force in North Rhine-Westphalia. Otherwise, this student is in *non-promote* group. The legislation establishes that the grades in North Rhine-Westphalia 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 next level if he/she 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), he/she has to pass all the subjects to obtain the grade [11, 12].

The available data that we have considered in this paper 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 1 [13].

2.2 *The type-epidemiological model*

Our mathematical model is built following an epidemiological approach considering that the academic performance of a student, Girl (G) or Boy (B), is a mixture of her/his own study habits and his/her classmates study habits, good or bad. In our model, we assume that the transmission of good and bad academic habits is caused by the social contact among students who belong to the same academic level [4, 5, 14].

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

Furthermore, we consider the following assumptions to design the model:

- Let us assume a homogeneous population mixing, i.e., each student can contact with any other student in the same educational level [15].
- *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 modelled 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 [16]. This leads us to suppose the following restrictions:

$$\alpha_1^G < \alpha_1^B,\ \alpha_2^G < \alpha_2^B,\ \alpha_3^G < \alpha_3^B. \tag{1}$$

In addition, we will assume that:

$$\alpha_1^G > \alpha_2^G > \alpha_3^G,\ \alpha_1^B > \alpha_2^B > \alpha_3^B, \tag{2}$$

because students in the higher levels are more mature than their mates in the lower levels [16].
- *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 modelled by the nonlinear 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 superindexes denotes the group susceptible to acquire bad study habits and, the second one denotes the group that transmit those bad study habits. All specific factors and social encounters involved in the transmission of the bad academic habits are embedded in β 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 $\overline{G}_i$ and $\overline{B}_i$, and it is modelled by the linear terms $\gamma_i^G \overline{G}_i$ and $\gamma_i^B \overline{B}_i$.
- *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 modelled by the nonlinear terms $\delta_i^{\overline{G}G} \overline{G}_i G_i +$

$\delta_i^{\overline{G}B}\overline{G}_i B_i$ and $\delta_i^{\overline{B}G}\overline{B}_i G_i + \delta_i^{\overline{B}B}\overline{B}_i B_i$. 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. These transitions are modelled by εG_1, εG_2, εG_3, εB_1, εB_2, εB_3, where

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

where $j = 0, 1, 2, 3, 4$, correspond to the academic years $2006 - 2007, \ldots, 2010 - 2011$, respectively.

- *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. This situation is modelled by the linear terms $\eta_i^G \overline{G}_i$ and $\eta_i^B \overline{B}_i$. We also assume that these transitions are proportional to the number of pupils in $\overline{G}_i$ and $\overline{B}_i$.
- *Access*: New students enter into the level 11 in the month of September in the promotable groups of girls and boys. It is modelled by the functions

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

where $j = 0, 1, 2, 3, 4$, correspond to the academic years $2006 - 2007, \ldots, 2010 - 2011$, respectively, and τ^G and τ^B to be determined.

Thus, under the above assumptions we build the nonlinear system of ordinary differential equations (3)-(5) 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) \tfrac{\overline{G}_1(t)}{T(t)} + \beta_1^{G\overline{B}} G_1(t) \tfrac{\overline{B}_1(t)}{T(t)}\right] + \left[\delta_1^{\overline{G}G} \overline{G}_1(t) \tfrac{G_1(t)}{T(t)} + \delta_1^{\overline{G}B} \overline{G}_1(t) \tfrac{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) \tfrac{\overline{G}_1(t)}{T(t)} + \beta_1^{G\overline{B}} G_1(t) \tfrac{\overline{B}_1(t)}{T(t)}\right] - \left[\delta_1^{\overline{G}G} \overline{G}_1(t) \tfrac{G_1(t)}{T(t)} + \delta_1^{\overline{G}B} \overline{G}_1(t) \tfrac{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) \tfrac{\overline{G}_2(t)}{T(t)} + \beta_2^{G\overline{B}} G_2(t) \tfrac{\overline{B}_2(t)}{T(t)}\right] + \left[\delta_2^{\overline{G}G} \overline{G}_2(t) \tfrac{G_2(t)}{T(t)} + \delta_2^{\overline{G}B} \overline{G}_2(t) \tfrac{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) \tfrac{\overline{G}_2(t)}{T(t)} + \beta_2^{G\overline{B}} G_2(t) \tfrac{\overline{B}_2(t)}{T(t)}\right] - \left[\delta_2^{\overline{G}G} \overline{G}_2(t) \tfrac{G_2(t)}{T(t)} + \delta_2^{\overline{G}B} \overline{G}_2(t) \tfrac{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) \tfrac{\overline{G}_3(t)}{T(t)} + \beta_3^{G\overline{B}} G_3(t) \tfrac{\overline{B}_3(t)}{T(t)}\right] + \left[\delta_3^{\overline{G}G} \overline{G}_3(t) \tfrac{G_3(t)}{T(t)} + \delta_3^{\overline{G}B} \overline{G}_3(t) \tfrac{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) \tfrac{\overline{G}_3(t)}{T(t)} + \beta_3^{G\overline{B}} G_3(t) \tfrac{\overline{B}_3(t)}{T(t)}\right] - \left[\delta_3^{\overline{G}G} \overline{G}_3(t) \tfrac{G_3(t)}{T(t)} + \delta_3^{\overline{G}B} \overline{G}_3(t) \tfrac{B_3(t)}{T(t)}\right],
\end{aligned} \tag{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) \tfrac{\overline{G}_1(t)}{T(t)} + \beta_1^{B\overline{B}} B_1(t) \tfrac{\overline{B}_1(t)}{T(t)}\right] + \left[\delta_1^{\overline{B}G} \overline{B}_1(t) \tfrac{G_1(t)}{T(t)} + \delta_1^{\overline{B}B} \overline{B}_1(t) \tfrac{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) \tfrac{\overline{G}_1(t)}{T(t)} + \beta_1^{B\overline{B}} B_1(t) \tfrac{\overline{B}_1(t)}{T(t)}\right] - \left[\delta_1^{\overline{B}G} \overline{B}_1(t) \tfrac{G_1(t)}{T(t)} + \delta_1^{\overline{B}B} \overline{B}_1(t) \tfrac{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) \tfrac{\overline{G}_2(t)}{T(t)} + \beta_2^{B\overline{B}} B_2(t) \tfrac{\overline{B}_2(t)}{T(t)}\right] + \left[\delta_2^{\overline{B}G} \overline{B}_2(t) \tfrac{G_2(t)}{T(t)} + \delta_2^{\overline{B}B} \overline{B}_2(t) \tfrac{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) \tfrac{\overline{G}_2(t)}{T(t)} + \beta_2^{B\overline{B}} B_2(t) \tfrac{\overline{B}_2(t)}{T(t)}\right] - \left[\delta_2^{\overline{B}G} \overline{B}_2(t) \tfrac{G_2(t)}{T(t)} + \delta_2^{\overline{B}B} \overline{B}_2(t) \tfrac{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) \tfrac{\overline{G}_3(t)}{T(t)} + \beta_3^{B\overline{B}} B_3(t) \tfrac{\overline{B}_3(t)}{T(t)}\right] + \left[\delta_3^{\overline{B}G} \overline{B}_3(t) \tfrac{G_3(t)}{T(t)} + \delta_3^{\overline{B}B} \overline{B}_3(t) \tfrac{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) \tfrac{\overline{G}_3(t)}{T(t)} + \beta_3^{B\overline{B}} B_3(t) \tfrac{\overline{B}_3(t)}{T(t)}\right] - \left[\delta_3^{\overline{B}G} \overline{B}_3(t) \tfrac{G_3(t)}{T(t)} + \delta_3^{\overline{B}B} \overline{B}_3(t) \tfrac{B_3(t)}{T(t)}\right],
\end{aligned} \tag{4}$$

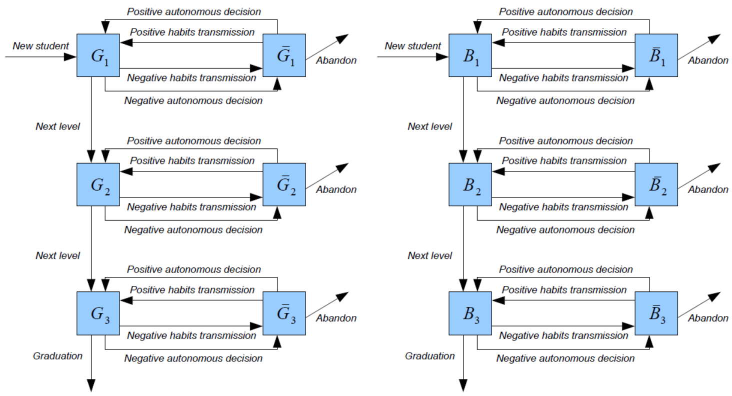

Figure 1. Flow diagram of the model (3)-(5). The boxes represent the students depending on their gender, level and academic results. The arrows denote the transit of students labelled by the cause of the flow.

$$\begin{aligned} T(t) = {} & G_1(t) + \overline{G}_1(t) + B_1(t) + \overline{B}_1(t) + G_2(t) + \overline{G}_2(t) + B_2(t) + \overline{B}_2(t) \\ & + G_3(t) + \overline{G}_3(t) + B_3(t) + \overline{B}_3(t). \end{aligned} \tag{5}$$

The flow diagram, associated to the above model, is plotted in Figure 1.

3. Scaling, fitting and predictions

Data in Table 1 are in percentages meanwhile model (3)-(5) is referred to number of students. It leads us to transform (scaling) the model into the same units as data in order to fit the model with the data. To do that, we follow the techniques developed in [17] about how to scale models where the population is varying in size. Here, we will not show the process and the scaled model because it is a technical transformation, the resulting equations are more complex and longer and do not provide extra information about the model. Moreover, the scaled model has the same parameters as the non-scaled model with the same meaning. In order to avoid introducing new notation, we will consider that the subpopulations $G_1(t)$, $\overline{G}_1(t)$, $B_1(t)$, $\overline{B}_1(t)$, $G_2(t)$, $\overline{G}_2(t)$, $B_2(t)$, $\overline{B}_2(t)$, $G_3(t)$, $\overline{G}_3(t)$, $B_3(t)$, $\overline{B}_3(t)$ correspond to the percentage of Girls and Boys in the promotable and non–promotable groups in the levels 11, 12 and 13.

Now, we compute the model parameters that best fit (in the mean square sense) the scaled model with the available data collected in Table 1. Computations have been carried out with *Mathematica 8.0®* [18]. The estimated model parameters are:

- *Negative autonomous decision*:
  - Girls per level: $\alpha_1^G = 0.00060$, $\alpha_2^G = 0.00000$, $\alpha_3^G = 0.00000$.
  - Boys per level: $\alpha_1^B = 0.00590$, $\alpha_2^B = 0.00585$, $\alpha_3^B = 0.00004$.
- *Negative habits transmission*:
  - Girls per level: $\beta_1^{G\overline{G}} = 0.05398$, $\beta_1^{G\overline{B}} = 0.08182$, $\beta_2^{G\overline{G}} = 0.00093$, $\beta_2^{G\overline{B}} = 0.00118$, $\beta_3^{G\overline{G}} = 0.14628$, $\beta_3^{G\overline{B}} = 0.12087$.
  - Boys per level: $\beta_1^{B\overline{G}} = 0.14022$, $\beta_1^{B\overline{B}} = 0.12587$, $\beta_2^{B\overline{G}} = 0.07844$, $\beta_2^{B\overline{B}} = 0.12687$, $\beta_3^{B\overline{G}} = 0.00714$, $\beta_3^{B\overline{B}} = 0.02304$.
- *Positive autonomous decision*:
  - Girls per level: $\gamma_1^G = 0.03113$, $\gamma_2^G = 0.14896$, $\gamma_3^G = 0.00061$.
  - Boys per level: $\gamma_1^B = 0.14688$, $\gamma_2^B = 0.00799$, $\gamma_3^B = 0.14933$.

Table 2. Prediction for the next four courses of the percentage of non-promoted students per gender and level, and the total. Note that there is a decreasing trend over the time in all levels with independence of gender. There are minor differences between boys and girls figures.

Groups	2011–2012	2012–2013	2013–2014	2014–2015
Level 11, Non–Promote girls	0,393%	0,345%	0,306%	0,272%
Level 12, Non–Promote girls	0,505%	0,482%	0,464%	0,449%
Level 13, Non–Promote girls	0,096%	0,079%	0,065%	0,054%
Level 11, Non–Promote boys	0,829%	0,810%	0,794%	0,781%
Level 12, Non–Promote boys	0,549%	0,502%	0,459%	0,421%
Level 13, Non–Promote boys	0,125%	0,104%	0,087%	0,073%
TOTAL	2,497 %	2,322 %	2,175 %	2,05 %

- *Positive habits transmission*:
  - Girls per level: $\delta_1^{\overline{G}G} = 0.06951$, $\delta_1^{\overline{G}B} = 0.11225$, $\delta_2^{\overline{G}G} = 0.14570$, $\delta_2^{\overline{G}B} = 0.14777$, $\delta_3^{\overline{G}G} = 0.00016$, $\delta_3^{\overline{G}B} = 0.00625$.
  - Boys per level: $\delta_1^{\overline{B}G} = 0.05547$, $\delta_1^{\overline{B}B} = 0.13882$, $\delta_2^{\overline{B}G} = 0.01273$, $\delta_2^{\overline{B}B} = 0.01189$, $\delta_3^{\overline{B}G} = 0.14485$, $\delta_3^{\overline{B}B} = 0.12842$.
- *Abandon*:
  - Girls per level: $\eta_1^G = 0.12966$, $\eta_2^G = 0.00163$, $\eta_3^G = 0.11675$.
  - Boys per level: $\eta_1^B = 0.12641$, $\eta_2^B = 0.14986$, $\eta_3^B = 0.00691$.
- *Access*:
  - Girls: $\tau^G = 0.11867$.
  - Boys: $\tau^B = 0.14977$.

Once the parameters are estimated, we are able to give predictions of each group and level over the next few years by computing the solutions of the model for values of time t in the forthcoming future.

In Table 2, we present the prediction of percentage of non–promote students for the next four courses.

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 modelling, to be precise, in the data and the model parameters. 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 a sophisticated and efficient method for determining a non–parametric probabilistic estimation of model parameters [19]. Specifically, the probabilistic estimation of the parameters is performed using a *residual* bootstrapping approach. In order to do it, and considering the general procedure published [19].

4.1 *Error term analysis*

In order to analyse the error terms, we have followed the next steps:

- We compute the output of the model with the obtained parameters at the time instants $t = 2006 - 2007, \ldots, 2010 - 2011$ and compute their differences (errors) with the corresponding data from Table 1.
- We analyse if the error terms are correlated. Pearson correlation coefficient was

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.

- Taking into account the Box-Ljung test [20], we also analyse 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. The obtained results allow us to accept that the error term corresponding to the Level 13 - Promoted Boys is statistically significant ($p - value = 0,02$), 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.
- For all the error terms which are not autocorrelated, the normality of the distribution of errors is checked by the Shapiro-Wilk Normality test [21]. 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 not ($p - value = 0.02$). Therefore, we can accept that all the errors present a univariate normal distribution excluding the error corresponding to the Level 11 - Promoted Girls.

4.2 *Generating new output perturbed data*

In order to generate the new perturbed output, we obtain the 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 obtain for each one of the 10 000 random error terms following the univariate normal distribution with their means and variances, respectively, obtained from the error terms. We add these error terms (10 000 times) to the output data for $t = 2006-2007, \ldots, 2010-2011$, obtaining a new set of perturbed data.

  For the autocorrelated error term corresponding to the Level 13 - Promoted Boys, we obtain the 10 000 random error terms using autoregressive techniques [22]. This has been realized by fitting an autoregressive time series model to the data [23]. After fitting the autoregressive model (AR) of the error term, we obtain the following model:

$$e_t = -0.7833397 * e_{t-1} + r_t, \tag{6}$$

  where e_t is the obtained error and r_t is the white noise at time t. The white noise is checked by the Shapiro-Wilk Normality Test that confirm, with a $p - value = 0,8088$, that the white noise follows the univariate normal distribution. Taking into account this, we have generated a set of 10 000 white noises following the normal distribution with its corresponding mean and variance. We add these white noises, r_t, (10 000 times) to (6) obtaining a set of 10 000 error terms.

  For the last error corresponding to the Level 11 - Promoted Girls, we assumed that the total sum of the error of each instant t is 1, this assumption allows us to obtain it by means arithmetic operations.
- Then, we compute the parameters which best fit (in the least mean square) the model with the set of perturbed data and store them, using the same procedure we used to estimate the obtained parameters. Note that this procedure allows us to have 10 000 sets of values for the parameters of the model.

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 system of differential equations (3)-(5) 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 that has obtained a mean square error whose value is in the 5% confidence interval around the best fit obtained of the model. In this case, the number of set of parameters 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 3.

Thus, in Figure 2 we can see graphically, for each subpopulation, the real data from Table 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 1000 outputs for each subpopulation at each time instant, where we have data about the academic results of German students in the North Rhine-Westphalia. These mean values are the ones obtained from our model and these predicted values from the academic year $2011 - 2012$ to $2014 - 2015$ appear in Table 3. Furthermore, we can observe that the obtained predictions in our model fit the real data and draw the different tendencies of the plots in each subpopulation, except for the students group in Level 11 - Non-Promoted Boys. Also, there is a slight decreasing in the non-promotable groups, in both, Girls and Boys. 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.

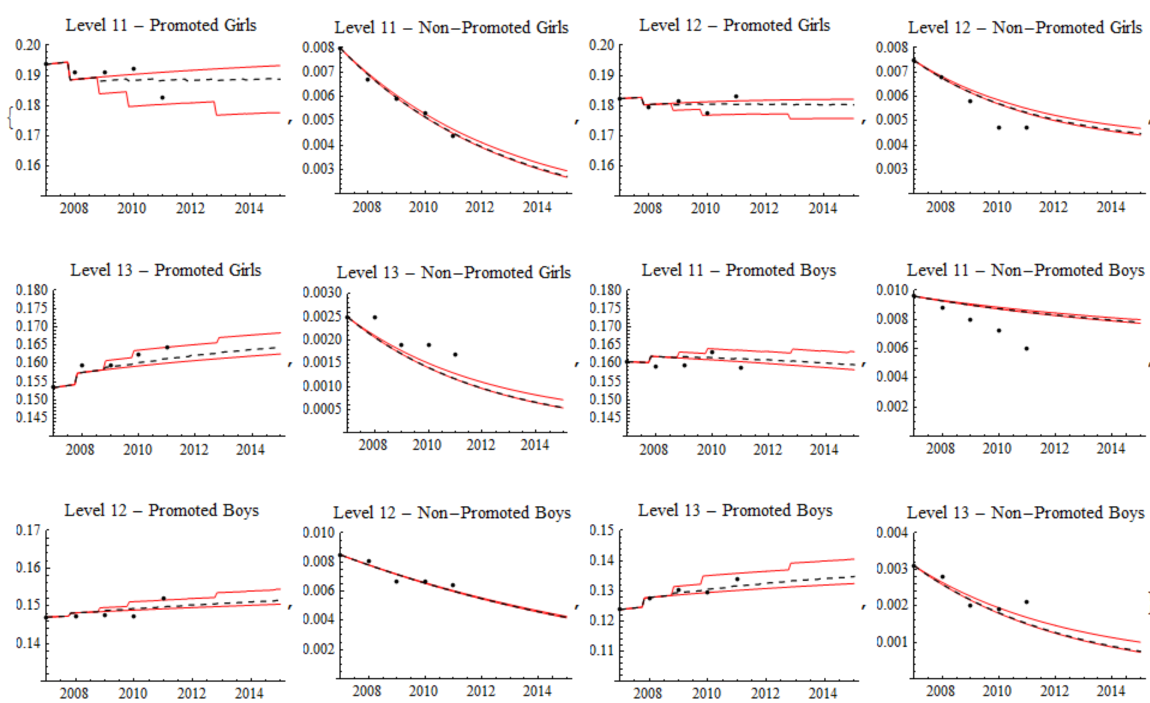

Figure 2. Real data (black points) and prediction with confidence intervals (red line) of the academic performance of German students in the North Rhine-Westphalia over the academic years $2006 - 2007$ to $2014 - 2015$. Smaller confidence intervals, represent less uncertainty in the predictions, the dash lines in the middle of the confidence intervals are their means.

Table 3. The 95% confidence interval prediction 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.

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]

5. Conclusion

In this paper, we present a model to study the dynamics of the students academic performance in the German region of North Rhine-Westphalia. In this model, we divide 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 data of the students academic performance, we estimate the model parameters fitting the model with the data. Thus, we can predict with confidence intervals the students academic performance in the next few years. In Figure 2, it is expected that the decreasing trend in all non–promotable groups continues in the next years. For instance, in the course $2014-2015$ around 2% of the students will not promote (see Table 2). Moreover, the level of promoted groups seems to the stabilization. This model will allow us to compare the performance of the coming new plan of study to the current one in order to evaluate if the change is as good as expected.

Acknowledgements

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness grant MTM2009-08587 and Universitat Politècnica de València grant PAID06-11-2070.

References

[1] *EU youth job strategy under fire*. Available at: $http://www.euractiv.com/en/enterprise-jobs/eu-youth-job-strategy-under-fire-news-497858$. (Accessed September 20, 2012).
[2] *Overcoming school failure: Policies that work. OECD Proyect description.* Available at: $http://www.oecd.org/education/preschoolandschool/45171670.pdf$. (Accessed September 20, 2012).
[3] *National system overview on education systems in Europe. European Comision.* Available at: $http://eacea.ec.europa.eu/education/eurydice/documents/eurybase/national_summary_sheets/047_BD_EN.pdf$. (Accessed September 20, 2012).
[4] H. Daniels, M. Cole and J. Wertsch, *Cambridge Companion to Vygotski*, Cambridge University Press, Cambridge, 2007.
[5] N.A. Christakis and J.H. Fowler, *Connected: The Surprising Power Of Our Social Networks And How They Shape Our Lives*, Brown and Company, 2009.
[6] 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*, Evaluation and Program Planning, 35(1) (2012), pp. 34–39.
[7] F.J. Santonja, E. Sánchez, M. Rubio and J.L. Morera, *Alcohol consumption in Spain and its economic cost: A mathematical modelling approach*, Mathematical and Computer Modelling, 52 (7-8) (2010), pp. 999–1003.
[8] 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) (2011), pp. 108–115.
[9] 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 (78) (2011), pp. 1634-1637.
[10] L.M.A. Bettencourt, A. Cintron-Arias, D.I. Kaiser and C. Castillo-Chavez, *The power of a good idea: Quantitative modelling of the spread of ideas from epidemiological models*, Physica A, 364 (2006), pp. 513–536.
[11] B. Landner, *Bildungsreport Nordrhein-Westfalen 2010 (in German) [Education Report North Rhine-Westphalia 2010]*, Statistische Analysen und Studien, Band 68. 2011.
[12] *Regionalverband Ruhr (in German) [Education Report Ruhr Region]*, Bildungsbericht Ruhr, Waxmann, 2012.
[13] *Kommunales Bildungsmonitoring: Tab. D13.2 Bestand an Schülerinnen und Schülern sowie Klassenwiederholungen (in German) [Municipal Education Monitoring: Table D13.2 Inventory of students and class repetition]*, Landesbetrieb Information und Technik Nordrhein-Westfalen(IT.NRW), Düsseldorf, 2012.
[14] K.R. Wentzel and D.E. Watkins, *Peer relationships and collaborative learning as contexts for academic enablers*, School Psychology Review, 31(3) (2002), pp. 366–377.
[15] J.D. Murray, *Mathematical Biology*, Springer, 2002.
[16] C. Fierro-Hernández, *Patrón de rasgos personales y comportamiento escolar en jóvenes (In Spanish) [Personal characteristic patterns and scholar behavior in young people]*. Revista de Educación, 332 (2000), pp. 291–304.
[17] M. Martcheva and C. Castillo-Chavez, *Diseases with chronic stage in a population with varying size*, Mathematical Biosciences, 182(1) (2003), pp. 1–25.
[18] Wolfram Mathematica 8®. Software available at $http://www.wolfram.com/products/mathematica$.
[19] G. Dogan, *Bootstrapping for confidence interval estimation and hypothesis for parameters of system dynamics models*, System Dynamics Review, 23 (2007), pp. 415–436.
[20] G. Ljung and G. Box, *On a mesasure of lack of fit in time series models*, Biometrica, 66 (1979), pp. 265–270.
[21] B. Yazici and S. Yolacan, *A comparison of various tests of normality*, Journal of Statistical Computation and Simulation, 77(2) (2007), pp. 175–183.
[22] H. Akaike, *Fitting autoregressive models for prediction*, Annals of the Institute of Statistical Mathematics, 21(1) (1969), pp. 243-247.
[23] Brockwell, P. J. and Davis, R. A. *(1996) Introduction to Time Series and Forecasting*, Springer, New York, 1996.