Analysis of Portugal's national exam grades in the period of 2010-2017 from a macro socioeconomic perspective.
Processing Big Data, Spring 2018

João Diogo Pinto
joao.pinto@tecnico.ulisboa.pt

Keywords: Data Mining, Portugal, Linear Regression, School, Student performance

Abstract: The Portuguese education system requires all high-school students who want to enter university to take a standardized exam on the same day at the same hour for one same subject. These exams are one of the main requirements for entering in a Portuguese university, and can be seen as the final result of three years of hard work in a student's life. In this study we perform a data mining task to assess the effects that socioeconomic factors have on a students national exam performance. Using linear regression as a model to predict a student's exam score with student's anonymous information and socioeconomic indicators associated with the school's municipality, the models scored around 0.25 in terms of R-squared. This was somewhat expected since current literature puts the emphasis on the student's familiar and social environments, data which was not used for this study. Despite this, the socioeconomic indicators add predictive value to the regression, so a socioeconomically adjusted school ranking was developed (SEAR) in order to compare school exam performane taking into account different socioeconomic context at the municipality level. All the data and code used to produce this article can be accessed at https://github.com/jonydog/examspt .

1 INTRODUCTION

1.1 Goals of this work

The main goal of this study is to assess the effects of socioeconomic factors over students grades on the high-school national exams and take them into account when doing a school ranking. Every year when the exam results are available a school ranking is done based on that year's average exam grades per school. This ranking is given much praise in the media, and is taken into consideration when parents decide the school for their children. Yet the typical ranking does not account for different socioeconomic contexts among the schools, it is simply the result of averaging the school's grades and order them from highest to lowest. If the goal of the ranking is to evaluate the quality of teaching, it is a fairly poor ranking, as known in several studies, the most important factors regarding school achievment point in the direction of the level of education in the family, and social environment as the most significant factors (DGEEC, 2016), socioeconomic variables are also known to be decisive in regional performance gaps (Pereira and Reis, 2012) Thus a ranking that would adjust the observed school ranking to its socioeconomic context should be very useful. Adjusted rankings are fairly common, for example, when comparing hospital treatment quality a very common indicator is the risk-adjusted mortality rate (Peterson et al., 2000), this indicator takes into account the different kinds of population treated on the hospitals under comparison.

In this work we propose a new ranking method called SEAR standing for socioeconomically adjusted school ranking. The steps to get to its definition are the following:

1. Assess regional variability at the municipality level;
2. try to predict exam scores using socioeconomic variables at student and school levels;
3. formulate socioeconomic adjusted school ranking (SEAR) and compare them with existing rankings.

1.2 State of the art

School performance has been since very long, a much prolific scenario of social sciences studies. One important driver for this work, is the notion that maybe contrary to popular belief, school plays a somewhat more limited role than a students socioeconomic environment (Hanushek, 1997), the authors in this famous paper, reviewed 400 papers on the school performance subject and demonstrated that there is no strong relationship between school resources and academic success, which is paradoxical. More recent studies like (de Oliveira et al., 2013) also favor socioeconomic factors to explain school performance. Within external factors even contingent phenomena like a child's month of birth (Kinard and Reinherz, 1986) seems to play a role in school performance. On the other hand more Portuguese studies indicate that there is some variability explained by school's quality of teaching (Pereira and Reis, 2012).

2 Dataset

The dataset used for this work has two main sources. The first is the Portuguese government's national exams official database, made publicly available by (DGEC,). This source provides us with exam data, and its associated dimensions that are represented in the ER diagram in Figure 1, by the tables that link directly to the table **Exam**. The other source of data provides us with socioeconomic indicators at the municipality level, represented by the tables with the prefix **PORDATA** since their source is (PORDATA,), an electronic database of official government data. Both the data from DGEC and PORDATA were in the form of .xslx files that were loaded to a MySQL database organized in relational tables as shown in Figure 1. The main link between students exams and socioeconomic data is the school at which the exam is taken, it links the exam to the municipality. The table **Exam** contains approximately 3.5 million rows of data, each represents an exam performance by a given student on a given subject, at a given date. The data is anonymous, the existing variables and its description are provided in Table 1.

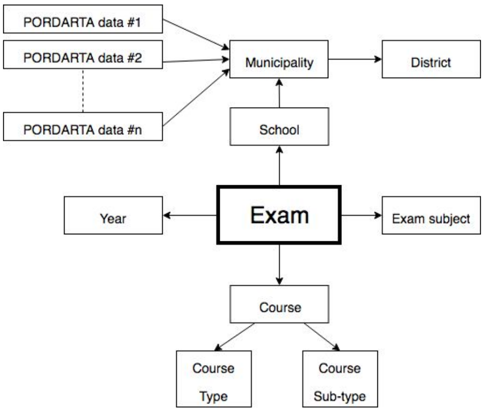

Figure 1: Entity-relationship model of the dataset, a star-schema.

Table 1: Variables for each exam performance

Variable	Description
year	between 2010 and 2017
school_id	id of the school
phase	1 or 2
exam_type_id	the exam's discipline
for_approval	if the student needs it to pass the year
internal	student enrolled in the school
to_improve	if student is improving previous grade
to_ingress	if student is doing the exam to ingress in the university
has_internal	if the student is enrolled in the school
sex	the sex of the student
age	the age of the student
course_id	the high-school course in which the student is enrolled
exam_score	numeric score between 0 and 200

Regarding the PORDATA dimensions, their description

Table 2: Socioeconomic variables

Variable	Description
births_out	birth rate oustide of wedlock or couples union
child_mort_rate	child mortality rate
crime_rate	number of crimes per 1000 residents
divorce_rate	divorce rate
ilet_rate	fraction of analphabet population
purchase_power	average purchase power of the population
university_degree	fraction of resident with a university degree

Regarding the economic indicators, some of them were not available annually, e.g. the rates of analphabet population and university degree holders. For this reason, all of these indicators are fixed along the years from 2010 to 2017, for each indicator the most recent data available was taken.

3 Assessing municipality variability

This is the first step of the analysis, mainly because given the fact that our datasets only provides socioeconomic indicators at the municipality level, if there are no significant differences on exam scores among municipalities there is no point in continuing this analysis, since regarding any exam if they belong to the same municipality the will share the same values for the available socioeconomic indicators. To assess this differences an ANOVA test was for the mean exam score among 308 different groups, one group representing each existing municipality. The ANOVA tests were run for each pair $\{$year,exam_type$\}$, every inference in this work was made cutting data by one specific year and one particular exam discipline, the reason for this decision was to assure that there were no correlated data rows in any analysis, and also to mitigate confounding factors regarding possible fluctuations in exam difficulties along the years. The p-values were saved for each test, and its summary can be seen in Table 3 below.

By inspecting the table it is fairly reasonable to

Table 3: Summary of the p-values for the ANOVA tests

Min	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	0.00	0.00	0.007	0.00	0.11

conclude that regional differences at the municipality level are indeed registered.

4 Predicting exam scores using socioeconomic variables

This step is required because, our final adjusted ranking will compare the real observed exam score, with an artificial score produced by a linear regression that takes socioeconomic variables into account in its training, and intentionally ignores regional or school variables. We approached this prediction task in two different ways. The first one formulates the problem as if each schools took an exam, with its score being its mean exam score among its students for the given year and discipline. The second approach tries to capture the variability associated with student level information, such as sex, age, and others. After predicting for each student, the average of the school is taken, and that is the final prediction. Both of the approaches were experimented and results are described in the next subsections.

4.1 Predicting exam grades by school

After several tests, a very typical fit is presented in Table 4. The fitted models consistently presents both variables births_out and ilet_rate as statistically significant (p-value < 0.05), and variable university_rate is about 50% of the times significant, the remaining ones attained high p-values virtually every time. The signals of the significant coefficients remained always the same. Interpreting the coefficients, we see that having a negative sign with births out of wedlock makes sense, since this indicator is also associated with familiar instability, the signs of university_degree ilet_rate are also intuitive. The problem with this fits was that the typical value for the R-squared metric was between 0.04 and 0.09, which is very low, even for the standards in this field which usually never exceed 0.30 (Pereira and Reis, 2012). Regarding the predictive value of the model, we assessed it using the mean absolute error with a 70-30 cross-validation scheme, the mean value of this metric was 16 points, which means that the model on average predicts a grade 16 points below or above the observed value of the mean exam score for a given school.

Table 4: Linear regression fit for year 2011 and discipline 'Matemática A'. R-squared =0.09

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	91.9370	5.7830	15.90	0.0000
births_ou	-0.4049	0.1700	-2.38	0.0177
child_mort_rate	0.1142	0.2417	0.47	0.6367
crime_rate	-0.1723	0.1121	-1.54	0.1249
divorce_rate	-0.0103	0.0335	-0.31	0.7582
ilet_rate	-71.8458	31.5753	-2.28	0.0234
purchase_power	-0.0095	0.0842	-0.11	0.9099
university_rate	93.4203	38.9976	2.40	0.0170

4.2 Predicting exam grades by student

Using the second approach, as mentioned we can add each student specific variables, which in principle adds more information to the model. The application of this approach, for the same year and discipline used previously, is presented in Table 5. The student variables are all significant, and their coefficients signal is systematically the same. Their values make sense in terms of experience, females perform better overall, second phase exams get worse grades, internal students behave better, and older ones tend to get worse grades, this last is explained by the fact that older students probably did not pass some of the previous years. The significant socioeconomic variables are the same as in the school approach. By using the same cross-validation scheme, an average mean absolute error of 320 was measured, in the other hand these models attain on average a much better R-squared, typically between 0.19 and 0.25.

Table 5: Linear regression fit for year 2011 and discipline 'Matemática A'. R-squared=0.22

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	163.2342	2.5445	64.15	0.0000
sexm	-2.5100	0.3421	-7.34	0.0000
phase	-14.8164	0.3825	-38.74	0.0000
for_approvalS	-26.0515	0.4407	-59.12	0.0000
internalS	38.2110	0.3896	98.07	0.0000
age	-3.6272	0.1083	-33.51	0.0000
births_out	-0.2441	0.0369	-6.61	0.0000
child_mort_rate	-0.0175	0.0595	-0.29	0.7686
crime_rate	-0.0096	0.0228	-0.42	0.6736
divorce_rate	0.0076	0.0077	0.98	0.3248
ilet_rate	-76.1964	8.1565	-9.34	0.0000
purchase_power	-0.0092	0.0167	-0.55	0.5812
university_rate	98.8468	7.3380	13.47	0.0000

This by student approach was the chosen one, the much higher R-squared is the main reason, and with this approach we also take into account variables that are not responsibility of the school quality, such as the ratio of boys and girls, or their age. In addition to these reasons, when computing all the predictions for a given school, year and discipline and averaging them, the typical values for the mean absolute error of these predicted averages and observed school averages were between 10 and 14 for the tests done, which indicates a smaller error comparing with tbe by-school approach.

5 Socioeconomically adjusted school ranking (SEAR)

5.1 Calculating the SEAR ranking

In order to construct this new school ranking we need to fit a linear regression to the exam data regarding one specific year *year* and one given discipline *disc*. The typical used ranking score *RS* is the school's average grade, for a school i with $N_{i,year,disc}$ students is given by:

$$RS_{typical} = \frac{\sum_{i=1}^{N_{i,year,disc}} exam\_grade(year, disc, i)}{N_{i,year,disc}} \quad (1)$$

with $exam\_grade(year, disc, i)$ being the exam score of student i on that specific year for a given discipline.

The SEAR ranking score is given by the following formula:

$$RS_{SEAR} = \frac{RS_{typical}}{\frac{\sum_{i=1}^{N_{i,year,disc}} M_{year,disc}(i)}{N_{i,year,disc}}} \quad (2)$$

where $M_{year,disc}(i)$ is the prediction of the corresponding fitted model to the students i data, including its associated socioeconomic variables inherited by its school municipality. So this ranking gives us a ratio, as to how much the school is behaving above or below the value predicted by the linear regression taking as explanatory variables only its students individual characteristics and the inherited socioeconomic ones.

5.2 Comparing rankings

In this subsection, some examples of the SEAR ranking are calculated for some pairs {year,exam_type} and compared with the usual average grade non-adjusted ranking.

In Tables 6 and 7 (see Appendix) , we have the non-adjusted and adjusted top-10 ranked schools for the year 2011 on the exam of 'Matemática A'. We see that between the rankings, the majority of the top-10 schools remains in the two, but four schools from less favourable socioeconomic context now appear on the top-10 SEAR ranking. The top school is now one from Loulé, one locations with high value of births out of wedlock or couple unions and also one of the highest in terms of analphabet peope and less university degrees per capita. In Tables 8 and 9 we have these two ranking now for the Português exam of 2017. Again the top-10 remains very similar, but for example a school from Povóa do Varzim climbs one spot given its less favourable socioeconomic indicators.

6 Future work

I think the more severe limitations of this work have to due with the lack of fine graining of the data, it is common sense that within the same municipality, contrasting socioeconomic contexts exist. The predictive task can be further refined, the mean absolute error of around 15, it is still fairly high, it should be a good idea to experiment with other models besides linear regression. One thing that was not accounted for when doing the regression was feature selection and outlier removal, this should be the next direction of improvement, if we want a more accurate ranking.

7 Conclusions

We observed that socioeconomic factors do influence overall students performance on the national exams, although as seen in previous studies the large majority of the variability cannot be explained by these macro socioeconomic factors contained in our dataset. In this work the best linear regression using the available socioeconomic data to predict the exam score only achieved an R^2 of around 0.20. The new developed ranking can be very useful when comparing schools independently of their socioeconomic context, and with this better reflect the school overall teaching quality. Regarding the dataset, a problem was found with the data, when compared with the official annual rankings provided by Ministerio da Educação, the observed averages by school are slightly different, one possible cause may be from the fact that our dataset is anonymous and so the dataset may contain first phase and second phase exams from the same student. This hypothesis is strengthened by the fact that the official values are always a bit higher than the ones calculated using this dataset.

REFERENCES

de Oliveira, P. R., Belluzzo, W., and Pazello, E. T. (2013). The public–private test score gap in brazil. *Economics of Education Review*, 35:120 – 133.

DGEC. Base de dados dos exames finais nacionais do ensino secundário do período 2011-2017.

DGEEC (2016). Desigualdades socioeconómicas e resultados escolares.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis*, 19(2):141–164.

Kinard, E. M. and Reinherz, H. (1986). Birthdate effects on school performance and adjustment: A longitudinal study. *The Journal of Educational Research*, 79(6):366–372.

Pereira, M. and Reis, H. (2012). What accounts for portuguese regional differences in students' performance? evidence from oecd pisa. *Economic Bulletin and Financial Stability Report Articles and Banco de Portugal Economic Studies*.

Peterson, E. D., DeLong, E. R., Muhlbaier, L. H., Rosen, A. B., Buell, H. E., Kiefe, C. I., and Kresowik, T. F. (2000). Challenges in comparing risk-adjusted bypass surgery mortality results: Results from the cooperative cardiovascular project. *Journal of the American College of Cardiology*, 36(7):2174 – 2184.

PORDATA. Dados estatísticos de municípios.

8 Appendix

Table 6: Matemática A, 2011 top-10 ranked school by the typical ranking score

	School	Municipality	AvgGrade	SEAR
1	Instituto de Educação e Desenvolvimento - INED	Maia	159.00	1.70
2	Colégio Nossa Senhora do Rosário	Porto	153.33	1.46
3	Colégio Rainha Santa Isabel	Coimbra	147.43	1.33
4	Externato Marista de Lisboa	Lisboa	143.30	1.31
5	Colégio D. Diogo de Sousa	Braga	142.37	1.34
6	Colégio São Teotónio	Coimbra	141.64	1.33
7	Colégio Luso-Francês	Porto	139.14	1.25
8	Escola Técnica e Liceal Salesiana Santo António (Estoril)	Cascais	136.94	1.42
9	Colégio dos Cedros	Vila Nova de Gaia	136.93	1.35
10	Externato Ribadouro	Porto	136.45	1.27

Table 7: Matemática A, 2011 top-10 ranked school by the SEAR ranking

	School	Municipality	AvgGrade	SEAR
1	Colégio Internacional de Vilamoura	Loulé	122.00	1.88
2	Instituto de Educação e Desenvolvimento - INED	Maia	159.00	1.70
3	Externato Paulo VI	Gondomar	136.17	1.48
4	Escola Básica e Secundária Cunha Rivara, Arraiolos	Arraiolos	129.42	1.46
5	Colégio Nossa Senhora do Rosário	Porto	153.33	1.46
6	Escola Técnica e Liceal Salesiana Santo António (Estoril)	Cascais	136.94	1.42
7	Escola Secundária D. Pedro I	Alcobaça	97.00	1.42
8	Colégio Valsassina	Lisboa	132.42	1.40
9	Instituto D. João V	Pombal	102.53	1.40
10	Escola Básica e Secundária de Albufeira	Albufeira	112.29	1.39

Table 8: Português, 2017 top-10 ranked school by the typical ranking

#	School	Municipality	AvgGrade	SEAR
1	Academia de Música de Santa Ceilia	Lisboa	146.06	1.33
2	Colégio São João de Brito	Lisboa	145.16	1.32
3	Colégio Nossa Senhora do Rosário	Porto	143.48	1.29
4	Colégio de Amorim	Póvoa de Varzim	142.84	1.31
5	Colégio de Santa Doroteia	Lisboa	139.98	1.24
6	Colégio Bartolomeu Dias	Loures	136.84	1.27
7	Colégio Luso-Francês	Porto	136.64	1.23
8	Externato Ribadouro	Porto	135.61	1.21
9	Colégio da Associação Cultural e Recreativa de Fornelos	Fafe	135.41	1.25
10	Colégio Horizonte	Porto	134.56	1.15

Table 9: Português, 2017 top-10 ranked school by the SEAR ranking

	School	Municipality	AvgGrade	SEAR
1	Academia de Música de Santa Cecilia	Lisboa	146.06	1.33
2	Colégio São João de Brito	Lisboa	145.16	1.32
3	Colégio de Amorim	Póvoa de Varzim	142.84	1.31
4	Colégio Nossa Senhora do Rosário	Porto	143.48	1.29
5	Colégio Bartolomeu Dias	Loures	136.84	1.27
6	Colégio dos Cedros	Vila Nova de Gaia	133.36	1.27
7	Escola Secundária de Castro Daire	Castro Daire	125.92	1.26
8	Colégio da Associação Cultural e Recreativa de Fornelos	Fafe	135.41	1.25
9	Externato Senhora do Carmo	Lousada	132.00	1.24
10	Colégio de Santa Doroteia	Lisboa	139.98	1.24