Overview description of the Gambian GABECE Educational Data and associated algorithms and unsupervised learning process

Ousman Saine (1), Soumaila Dembele (2), Gane Samb Lo (3) and Mohamad Cheikh Haidara (4)

1 University of The Gambia, Banjul, The Gambia & Policy Planning, Analyses, Research and Budgeting Directorate, Ministry of Basic and secondary Education, Willy Thorpe Palace, Banjul, The Gambia
2 Université des Sciences Sociale et de Gestion de Bamako (USGB) Faculté des Sciences Économiques et de Gestion (FSEG)
3 LERSTAD, Gaston Berger University, Saint-Louis, Sénégal (main affiliation).
4 AUST - African University of Sciences and Technology, Abuja, Nigeria

Received on January 1, 2020; Accepted on June 10, 2020.

Abstract. As the first paper of a series of exploratory analysis and statistical investigation works on the Gambian GABECE data based on a variety of statistical tools, we wish to begin with a thorough unsupervised learning process through descriptive and exploratory methods. This will lead to a variety of discoveries and hypotheses that will direct future research works related to this data.

Keywords. GABECE; educational data, unsupervised learning; data analysis; exploratory methods; R programs; The Gambia.

*Corresponding author: Ousman Saine (ousmansaine@edugambia.gm)
Soumaila Dembele : soumailadembeleussgb@gmail.com
Gane Samb Lo : gane-samb.lo@edu.ugb.sn
Cheikh Mohamad Haidara: chheikhh@yahoo.fr
1. Introduction

In this paper, the GABECE data is first presented and next statistically studied for the first time. We proceed to an unsupervised learning process. We will provide the description of the variables and consider the geographical and temporal sub-data. For each sub-data, different techniques will be applied: cleaning, outliers detection, frequency and related tables, spatial (geographical) or gender patterns, proportions of grades and evolution in time, etc.

A series of statistical analyzes will be undergone on the data by ourselves or other researchers. To prepare that, we wish to devote a note on the general presentation of the data and an unsupervised learning process with respect to the gender, the area and the time. Also R codes (R Core Team (2020) and Crawley (2009)) aiming at reading the data and getting the variables presented in Section 2 will be provided in Appendix II in page 848.
The rest of the paper is organized as follows. We will complete this introduction by the main findings in this paper and related recommendations. The data is fully described in Section 2. In Section 3, the GABECE data is compared with national statistics concerning population sizes of regions and genders to see whether or not access to education is fair in the Gambia. In section 4, the completeness of the the data is studied and it will be shown that the data present un non-negligible num-ber of incomplete data everywhere. In Section 5, a cleaning study is undergone by the detection of the outliers and it will appear that eliminating outliers, as usually recommended, would lead to a serious error since excellent students would simply ignored in the study. Instead of rejecting, we will propose regrouping the grades. In Section 6, we undergo a general investigation of the performances. We close the paper by conclusions in Section 7. Appendices are put after the bibliography. The main conclusions are summarized below.

(C1) Neither the regions nor the two genders are well represented in the educational GABECE data. The western regions, in general, have more students at grade 13 when compared with these real weighs in the country.

Although being more numerous in the country, females are under represented in the data. Even worse, the number of females in the data behaves chaotically where the number of males steadily increases with the year.

(R1) As a consequence, we strongly recommend to use sampling methods with sample sizes proportional to the real weighs of sub-groups in the global population.

(C2) There is a big number of incomplete data in all years, in all regions, and in each gender. After dropping them, we still have enough data to proceed to a significant analysis data.

(R2) On this side, we recommend statistical studies based on missing data estimation techniques. There is a lot to do on this, using nearest neighborhood, regression, classification, etc.

(C3) The statistical studies are about the grades in the core disciplines: Mathematics, Sciences, English, and SES. The grades range from 1 (excellent) to 9 (fail). Unfortunately, the level of students is not very fairly good since the fail grade is dominant almost everywhere. As well, there is a very few number of excellent students.

(R3) The result of that situation is that excellent grades 1, 2 or 3 are detected as outliers. So, in the statistical studies, such grades should be excluded. But, fortunately, a regrouping of all grades into three levels seems to be the best way to overcome that problem. Those three levels are: credit (regrouping credit grades 1 to 6), pass (regrouping 8 and 8) and fail.
The paper is destined to be a basis for a series of statistical studies on GABECE data and beyond on similar data for countries of the West African Examinations Council, whose The Gambia is member of, and on educational data in general.

2. Overall presentation of the GABECE data

The West African Examinations Council was established in 1952. This Council has the responsibility of ascertaining examinations required in the public interest in the English-speaking West African countries, comparable to those of equivalent examining authorities internationally, and to develop syllabuses, arranges and administers examinations and awards certificates. It also conducts a number international examinations.

The data we are going to use in this document is from the above aforementioned body. The data covers the period from the 2012 academic year to 2017. The mode of the GABECE examination is that each student is required to choose 7 subjects at least and 9 subjects at most. Out of the subjects taken, four (4) are compulsory and are called core subjects, which are English (ENG), Mathematics (MATH), Sciences (SCIENCES) and Social and Environmental Studies (SES).

Individual performance grants the student admission to Senior Secondary Education. The grading system in this result is categorized as follows: Credit (grades from 1 to 6), Pass (7 to 8), Fail (9). The result of a student is determined by the grades of the four core subjects plus the two best grades from the student’s choice. The best result a student could have is the aggregated grade from 6 (six), one in all the core subjects, plus one in the two choices of the students and the worst result is aggregated to 54 with failure in all subjects with grade 9.

In this study, we focus on the core grades (G1), (G2), (G3) and (G4) corresponding to English, Mathematics, Sciences and SES respectively. Here are the conventions we will be using to know whether we use a grade for the whole population or not, for all the years or not, for both genders or not, or whether we use a grade for one particular gender, or a year or an area.

The corresponding variables of the core grades are denoted as $G1$, $G2$, $G3$ and $G4$ before the cleaning of the data and cover all areas, all time and all gender.

We also have the variables year (from 2012 to 2017), gender (coded 1 for males and 2 for females), and region (coded for 1 to six regions).

The missing data, which correspond to absences of students or submission of blank sheets by student, represent 4.54% of the data (5792 out of 121838). The new variables become:

Journal home page: http://www.jafristatap.net, www.projecteuclid.org/euclid.ajas
G1A, G2A, G3A, G4A, yearA, genderA, regionA.

When we restrict our analysis on a specific year for all gender, the variables are:

G1AY, G2AY, G3AY, G4AY, yearAY, genderAY, regionAY.

When we restrict our analysis on a specific region for all genders and years, the variables are:

G1AR, G2AR, G3AR, G4AR, yearAR, genderAR, regionAR.

When we restrict our analysis on a specific region and for a specific year, for all genders, the variables are:

G1AYR, G2AYR, G3AYR, G4AYR, yearAYR, genderAYR, regionAYR.

In each of the previous cases, the specified year and region are given by the variables yearOfStudy and regionOfStudy.

If we want to specify and fix a gender, we use the filtering codes: Variable[gender=="1"] for males (and Variable[gender=="2"] for females). For example, the variables G1AR[genderAR=="1"], G2AYR[genderAYR=="2"], G3A[genderA=="1"], respectively contain:

a) the grades for males in English for the region determined by the value of regionOfStudy,

b) the grades in Mathematics for females for the year and the region determined by the values assigned to yearOfStudy and to regionOfStudy,

c) the grades for males in Sciences for all regions and all years.

These notation are fixed for once and will be used in similar papers. More notation related to transformed variables may come later.

We post a complete R code in Appendix II in page 848 for the reading the data and forming the variable described above.

In this first release of studies on GABECE, we are going to give the Gambian national context of the data and compare the representation of specific population subgroups in the data. Next we describe the main characteristics of the data under the following points:

1. facts and trends on the population;
2. incomplete data;
3. outliers;

Journal home page: http://www.jafristatap.net, www.projecteuclid.org/euclid.ajas
## 4. General Trends of the Performances

### 3. In what context the GABECE data represents the whole population

The data concerns young students at age around 13 passing the grade 9 exam over the period 2012-2017. The Gambia has six administrative regions which are: (1) Kanifing (including Banjul), (2) Brikama, (3) Kerewan, (4) Mansa Konko, (5) Janjanbureh-Kuntaur, (6) Basse. That order corresponds to covering the country from East to West. In particular Basse is the extreme eastern area in the west from the Capital Banjul in Kanifing.

In total, 127,630 students took part to the GABECE exam and 5,792 missed at least one discipline exam or handed a blank copy.

The number of declared students by year, region and gender are given in Table 1. The number of students by year, region and gender in absolute values and in percentages are posted in Table 2.

We may have expected that the number of pupils is increasing from year 2012 to 2017. However, as shown in Fig. 1, we see drops in 2013 and 2014 for the total number of students before a tendency of increase which also is broken in 2016. For boys, the positive trend is remarkable after a drop in 2013. The evolution of the number of female is somehow chaotic, suggesting a lack of a...
Overview description of the Gambian GABECE Educational Data and Associated Algorithms and Unsupervised Learning Process.

---

Table 2. Number of Students by year, region and gender in absolute values and in percentages. CO: counted Objects.

| Region | CO 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | All time |
|--------|---------|------|------|------|------|------|----------|
| Male   | 3651    | 3627 | 3617 | 3698 | 3722 | 3856 | 22,171   |
| Female | 3164    | 3090 | 3012 | 3145 | 3157 | 2976 | 18,544   |
| Male (%) | 53.57  | 54.00 | 54.56 | 54.04 | 54.11 | 56.44 | 54.45    |
| Female (%) | 46.43 | 46.00 | 45.44 | 45.96 | 45.89 | 43.56 | 45.55    |

Region 2: Brikama

| Male   | 4002    | 3897 | 4139 | 4896 | 4968 | 5303 | 27,205   |
| Female | 4180    | 4104 | 4124 | 4856 | 4551 | 4681 | 26,469   |
| Male (%) | 48.91  | 48.71 | 50.09 | 50.21 | 52.19 | 53.11 | 50.66    |
| Female (%) | 51.09 | 51.29 | 49.91 | 49.79 | 47.81 | 46.89 | 49.34    |

Region 3: Mansa Konko

| Male   | 1171    | 1065 | 1161 | 1019 | 1151 | 1052 | 6,583   |
| Female | 1089    | 1084 | 1030 | 1071 | 1069 | 960  | 6,303   |
| Male (%) | 51.81  | 49.56 | 52.99 | 48.76 | 51.05 | 52.29 | 51.09    |
| Female (%) | 48.19 | 50.44 | 47.01 | 51.24 | 48.95 | 47.71 | 48.91    |

Region 4: Janjanbureh

| Male   | 526     | 516  | 420  | 384  | 449  | 483  | 2,778   |
| Female | 516     | 563  | 427  | 428  | 405  | 434  | 2,773   |
| Male (%) | 50.48  | 47.82 | 49.59 | 47.29 | 52.58 | 52.67 | 50.05    |
| Female (%) | 49.52 | 52.18 | 50.41 | 52.71 | 47.42 | 47.33 | 49.95    |

Region 5: Kerewan

| Male   | 1048    | 920  | 804  | 799  | 872  | 832  | 5,273   |
| Female | 742     | 683  | 600  | 668  | 627  | 556  | 3,876   |
| Male (%) | 58.55  | 57.39 | 57.26 | 54.46 | 58.17 | 59.94 | 57.64    |
| Female (%) | 41.45 | 42.61 | 42.74 | 45.54 | 41.83 | 40.06 | 42.36    |

Region 6: Basse

| Male   | 274     | 353  | 412  | 458  | 491  | 506  | 2,494   |
| Female | 487     | 486  | 498  | 565  | 517  | 579  | 3,132   |
| Male (%) | 36.01  | 42.07 | 45.27 | 44.77 | 48.71 | 46.64 | 44.33    |
| Female (%) | 63.99 | 57.93 | 54.73 | 55.23 | 51.29 | 53.36 | 55.67    |

---

successful gender policy to get girls into the educational system or to keep them in.

We will compare the distribution of the regions and the gender with the national statistics. In Table 3, we have the distribution of the population over eight (8) regions (Banjul, Kanifing, Brikama, Kerewan, Kuntaur, Janjanbureh, Basse). In Fig. 2, we have the map of the Gambia. The country, which is entirely surrounded by Sénégal except at the Atlantic ocean coast, seems to be caught by a nearly rectangle horizontally ranging from East (Atlantic Ocean) to West in the interior of Sénégal.

Journal home page: http://www.jafristatap.net, www.projecteuclid.org/euclid.ajas
Overview description of the Gambian GABECE Educational Data and Associated Algorithms and Unsupervised Learning Process.

Fig. 1. Evolution of number of students over the years (globally at left, for boys at the center, for girls at right)

Fig. 2. Map of the Gambia. The region of Janjanbureh and Kuntaur are regrouped and named after Kuntaur

In the GABECE data, the regions of Banjul and Kanifing are regrouped in one region named after the later (Kanifing) and the regions of Janjanbureh and Kautur are regrouped in one region named after the first. In Table 4, we finally have the distribution in percentages of the population over the six geographical areas used in the GABECE data. That table also displays the areas in $km^2$ of each region.

Fig. 3 shows the weights of the regions in the country. We can see that the Region 1 has the greatest population density followed by the three extreme western areas.

As to the gender repartition, females are slightly more dominant of males over the study period as shown in Fig. 4
Fig. 3. Populations in % of regions nationally (left) and in the data (right)

Fig. 4. Domination of the female gender over of male gender over years. Source: Statistica.com (2020)
Table 3. Population by region in 2003 and 2013 (provisional). Abbreviations:
pop.C.03 = population Census 2003; pop.C.13 = population Census 2013; (C) Capital of the country; N.D= number of districts. Source: Wikipedia (2020)

| Regions   | Area (km²) | Pop.C.03 | Pop.C.13 | Capital | N.D |
|-----------|------------|----------|----------|---------|-----|
| Banjul (C)| 12.2       | 35,061   | 31,301   | Banjul  | 3   |
| Kanifing  | 75.6       | 322,735  | 382,096  | Kanifing| 1   |
| Brikama   | 1,764.3    | 389,594  | 699,704  | Brikama | 9   |
| Mansa Konko | 1,628.0  | 72,167   | 82,381   | Mansa Konko | 6 |
| Kerewan   | 2,255.5    | 172,835  | 221,054  | Kerewan | 7   |
| Kuntaur   | 1,466.5    | 78,491   | 99,108   | Kuntaur | 5   |
| Janjanbureh | 1,427.8  | 107,212  | 126,910  | Janjanbureh | 5 |
| Basse     | 2,069.5    | 182,586  | 239,916  | Basse   | 7   |

Total Gambia: 10,689, 1,360,681, 1,882,450, Banjul 43

Table 4. Populations of regions in % in Census 2003 and census 2013

| Num | Regions   | Area (km²) | Pop.C.03 | %   | Pop.C.13 | %   |
|-----|-----------|------------|----------|-----|----------|-----|
| 1   | Kanifing  | 87.4       | 357796   | 26.30% | 413397  | 21.96% |
| 2   | Brikama   | 1764.3     | 389594   | 28.63% | 699704  | 37.17% |
| 3   | Mansa Konko | 1628     | 72167    | 5.30%  | 82381   | 4.38%  |
| 4   | Kerewan   | 2255.5     | 172835   | 12.70% | 221054  | 11.74% |
| 5   | Janjanbureh | 2894.3  | 185703   | 13.65% | 226018  | 12.01% |
| 6   | Basse     | 2069.5     | 182586   | 13.42% | 239916  | 12.74% |

Total Gambia: 10689, 1360681, 1882450, 100.00%

Now, let us see how the distribution of the GABECE compare with the national data with respect to the region and the gender.

a. The regions are not equally represented.

The comparison of the percentages in Table 3 with those in Table 5 are summarized in Table 5. The plot of the ratios of region percentage in the data in the region percentage in the GABECE data, in Fig. 5, shows that the three western region are over represented by far in the education data, the extreme western is not almost equally represented and the three middle regions are seriously under represented. Indeed an almost equal presence of the regions in the data should have led to points very near the bisector line. The more a point is above the bisector line, the more the corresponding region is over represented and the more a point is below the bisector line, the more the corresponding region is under represented.

b. The female gender is poorly represented in the education data.

Journal home page: http://www.jafristatap.net, www.projecteuclid.org/euclid.ajas
From Tables 1 and 2 and Fig. 5, it is very clear that female Gambian are badly represented in the education system although they constitute the majority in the period of the study. At the exception of the Basse area, more boys arrive at Level 13 than girls from 2014 to 2018. The biggest gaps happen in Kerewan.

How to explain why the contrary happens in the rural zone of Basse in opposite to Kerewan area which is adjacent to it? Local studies should be undergone to study...
Table 5. Population of the regions in % nationally and in the data

| Regions            | Presence in the Gambia | Presence in the data |
|--------------------|------------------------|----------------------|
| 1 Kanyi           | 21.96%                 | 32.95%               |
| 2 Brikama          | 37.17%                 | 39.17%               |
| 3 Mansa Konko      | 4.38%                  | 10.54%               |
| 4 Kerewan          | 11.74%                 | 5.29%                |
| 5 Janjanburey-kuntaur | 12.01%            | 7.83%                |
| 6 Basse            | 12.74%                 | 4.14%                |

Table 6. Incomplete data by region

| Regions            | 1 | 2 | 3 | 4 | 5 | 6 |
|--------------------|---|---|---|---|---|---|
| Data              | 1550 | 2684 | 519 | 246 | 378 | 415 |
| Data %            | 3.8069 | 4.9980 | 4.0276 | 4.4316 | 4.1306 | 7.3764 |

such microscopic facts.

In Brikama, for 2012 and 2013, girls were well enrolled in the system according to their majority but the tendency came short to be persistent.

Conclusion 1. The GABECE data does not really represent the national statistics. Some regions are over represented (as the first region from East). The others are under represented with Basse being not very under present in the education data.

As well, the majority of the female population is not materialized in the education data. This, surely, is a sign of a poor gender equity policy to be solved.

Recommendation 1. In the statistical study, samplings should be used to extract from the data samples corresponding to national ratios for each region and gender to expect to have a fair comparison between of regions and of genders.

4. Incomplete data

The data presents a high number of incomplete data. The missing data concerns the grades G1, G2, G3, G4. For a grade, a missing data (-1 or -2) is assigned if the student did not attend the exam of that grade or he/she handed a blank copy. We only give here the percentages of missing data by region and year. We see that 2.5% to roughly 8% of the data are missing data.

Journal home page: http://www.jafristatap.net, www.projecteuclid.org/ euclid.ajas
5. Outliers

As explained in the introduction, the grades range from very excellent (1) to fail (9), with credits between 1 to 6. Grades (7) and (8) are for passing students with very average levels.

We will see later that the good students with grades (1), (2), (3) are very rare. They are so few that they seem to be outliers, that is data that can be taken as exceptional or accidental.

At the same times a bulk of the grades range over \{7, 8, 9\}.

The consequence is that outlier tests do detect students with credits as exceptional data, that should be studied apart. Here we use the simple Wilks tests for outliers which works as follows:

- Compute the first quartile $Q_1$, the second first quartile $Q_2 = Me$ which is the median and the third quartile $Q_3$
- compute the interquartile distance $D = Q_3 - Q_2$
- Compute the bounds $d_1 = Me - 1.5D$ and $d_2 = Me + 1.5D$.
- Declare outlier at left (OL) all elements below $d_1$ and outliers at right (OR) all observations beyond $d_2$

We applied that simple rule to each grade $G_1$, $G_2$, $G_3$ and $G_4$. The percentages of outliers at left (OL) and outliers at right (OR) are given by grade in Tables 8 - 11. In each table, The percentages of outliers at left (OL) and outliers at right (OR) are posted for every region. When we consider data by regions, the outliers are denoted by (OLr) [at left] and (ORr) [at left]. When data are restricted to regions and to a gender, outliers are denoted by (OLrf) [left] and (ORrf) [right] for females, (OLrm) [left] and (ORrm) [right] for males.

The percentage of outliers range roughly between 3% and 25%. The biggest values are related to male students.

The results confirm what has been said before.

Conclusion 2. Excellent grades are so rare that they have been declared outliers at left. Weak grades so very common that none of them is declared outlier at right.

Recommendation 2. There is no statistical interest to study the excellent grades. So it is highly recommended to regroup the grades into meaningful classes.
Table 8. Outliers in English grade by region and year

| Regions | 1     | 2     | 3     | 4     | 5     | 6     |
|---------|-------|-------|-------|-------|-------|-------|
| OLr     | 12.0388 | 14.2089 | 24.4279 | 23.9585 | 18.1123 | 20.6486 |
| ORr     | 0      | 0      | 0      | 0      | 0      | 0      |
| OLF     | 10.6113 | 10.0593 | 11.3523 | 11.8554 | 9.1617 | 24.2392 |
| ORF     | 0      | 0      | 0      | 0      | 0      | 0      |
| OLR     | 13.2076 | 13.3001 | 20.5589 | 16.9184 | 20.2336 | 22.8978 |
| ORR     | 0      | 0      | 0      | 0      | 0      | 0      |

Table 9. Outliers in Mathematics grades by region and year

| Regions | 1     | 2     | 3     | 4     | 5     | 6     |
|---------|-------|-------|-------|-------|-------|-------|
| OLr     | 23.4750 | 18.4683 | 16.9887 | 21.2629 | 14.9777 | 23.1817 |
| ORr     | 0      | 0      | 0      | 0      | 0      | 0      |
| OLF     | 24.5746 | 20.3071 | 18.5780 | 9.7478 | 18.0010 | 13.9349 |
| ORF     | 0      | 0      | 0      | 0      | 0      | 0      |
| OLR     | 22.5746 | 16.7094 | 15.4745 | 17.1072 | 24.5693 | 19.3186 |
| ORR     | 0      | 0      | 0      | 0      | 0      | 0      |

Table 10. Outliers in Sciences grades by region and year

| Regions | 1     | 2     | 3     | 4     | 5     | 6     |
|---------|-------|-------|-------|-------|-------|-------|
| OLr     | 14.3138 | 10.7826 | 23.8861 | 10.2356 | 18.4429 | 18.6336 |
| ORr     | 0      | 0      | 0      | 0      | 0      | 0      |
| OLF     | 15.0862 | 12.6854 | 10.8883 | 14.1512 | 9.0053 | 22.3374 |
| ORF     | 0      | 0      | 0      | 0      | 0      | 0      |
| OLR     | 13.6813 | 23.3288 | 19.7852 | 18.2779 | 23.2627 | 21.5179 |
| ORR     | 0      | 0      | 0      | 0      | 0      | 0      |

Table 11. Outliers in SES grades by region and year

example, we may regroup them into two class: Pass (including all grades not equal to 9) and Fail for grade 9; or into three class: credit (from 1 to 6), Pass (7 and 8) and Fail (9).

Journal home page: http://www.jafristatap.net, www.projecteuclid.org/euclid.ajas
6. General analysis of performances

We are giving some trends of the performances. More insight studies will be made in up coming papers.

Table 12 gives the percentages of grades (1 to 9) by discipline and year. We can see the following conclusions:

(C1) The fail grade is overwhelmingly dominant in Mathematics, Sciences and English disciplines. Even for SES which can be thought less tough than the three others, fail rates are significant and turn around 50%.

(C2) The excellency credit grade (1) is extremely rare and its rate does not exceed 2% at the exception of the disciplines SES for which it is still low and less than 5%.

(C3) The other credit grades are still low enough to be studied apart.
7. Conclusions

This paper is destined to be the basis of a series of more insightful papers using a variety of more or less deep statistical techniques. At least, the data is entirely described. The notations are posed. A general algorithm allowing to get all relevant variables and their local restrictions (in time (year), in area (region), in gender (female, male)).

First statistical trends are given, concerning incomplete data, outliers, fairness of representation of sub-groups in the educational data, some general evolutions of grades in the time. These statistical facts already direct to interesting leads and suggest hypotheses to be validated or not.

Acknowledgement The authors wish to thank the Ministry of Basic and secondary Education for making the GABECE data available.

References

Crawley M. J.(2009). The R book. Wiley.
Overview description of the Gambian GABECE Educational Data and Associated Algorithms and Unsupervised Learning Process.

James G., Witten D., Hastie T. and Tibshirani R.(2008) Introduction to statistical learning with Applications in R. (Corrected at 8th edition 2017). Springer

R Core Team (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Statistica.com: https://www.statistica.com/statistics/967843/total-population-of-gambia-by-gender/. Visited on 2020/11/13.

Wikipedia website. https://en.wikipedia.org/wiki/The_Gambia#Administrative_divisions. Visited on 2020/11/11.
Appendix I: More tables.

Tables of uncomplete data.

Table 14. Uncomplete data by year and region is absolute numbers

|       | 1   | 2   | 3   | 4   | 5   | 6   |
|-------|-----|-----|-----|-----|-----|-----|
| 2012  | 202 | 210 | 60  | 77  | 73  | 30  |
| 2013  | 190 | 188 | 68  | 36  | 70  | 37  |
| 2014  | 156 | 180 | 58  | 38  | 41  | 46  |
| 2015  | 248 | 572 | 97  | 26  | 46  | 149 |
| 2016  | 416 | 984 | 123 | 29  | 72  | 85  |
| 2017  | 338 | 550 | 113 | 40  | 76  | 68  |

Table 15. Uncomplete data by year and region in relative numbers

|       | 1    | 2    | 3    | 4    | 5    | 6    |
|-------|------|------|------|------|------|------|
| 2012  | 2.9640 | 2.5666 | 2.6548 | 7.3896 | 4.0782 | 3.9421 |
| 2013  | 2.8286 | 2.3497 | 3.1642 | 3.3364 | 4.3668 | 4.4100 |
| 2014  | 2.3532 | 2.1783 | 2.6471 | 4.4864 | 2.9202 | 5.0549 |
| 2015  | 3.6241 | 5.8654 | 4.6411 | 3.2019 | 3.1356 | 14.5650 |
| 2016  | 6.0473 | 10.3372 | 5.6318 | 3.3957 | 4.8032 | 8.4325 |
| 2017  | 4.9473 | 5.5088 | 5.6163 | 4.3663 | 5.4755 | 6.2872 |

Table 16. Uncomplete data by region and gender in absolute values

| Regions | 1   | 2   | 3   | 4   | 5   | 6   |
|---------|-----|-----|-----|-----|-----|-----|
| Females | 638 | 1130 | 250 | 130 | 224 | 175 |
| Males   | 912 | 1554 | 269 | 116 | 154 | 240 |

Table 17. Uncomplete data by region and gender in relative values

| Regions | 1    | 2    | 3    | 4    | 5    | 6    |
|---------|------|------|------|------|------|------|
| Females | 2.8776 | 4.1536 | 3.7976 | 4.6796 | 4.2464 | 7.0168 |
| Males   | 4.9180 | 5.8650 | 4.2678 | 4.1831 | 3.9731 | 7.6628 |
Overview description of the Gambian GABECE Educational Data and Associated Algorithms and Unsupervised Learning Process.

Table 18. Uncomplete data by year and gender in absolute values

| Years | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
|-------|------|------|------|------|------|------|
| Females | 327  | 320  | 251  | 476  | 694  | 479  |
| Males   | 325  | 269  | 268  | 662  | 1015 | 706  |

Table 19. Uncomplete data by year and gender in absolute values

| Years | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
|-------|------|------|------|------|------|------|
| Females | 3.0640 | 3.0834 | 2.3784 | 4.2296 | 5.9740 | 3.9810 |
| Males   | 3.1931 | 2.6873 | 2.7654 | 6.1678 | 9.8295 | 6.9310 |
Appendix: R codes

See Crawley (2009) for an overall review on R software.

The data are in the directory gabece, which is of a subdirectory where is this script.

The directory should contain the following files relation to the whole population of students, having the same length:

1. mesAppurements.txt: contains the variable okData with two values “ok” for complete data “ko” for incomplete data.
2. the original grades G1, G2, G2, and G1.
3. the covariates year.txt, gender.txt and region.txt

```r
appu <- read.table("gabece/mesAppurements.txt", h=TRUE)
attach(appu)
table(okData)

Gender <- read.table("gabece/gender.txt", h=TRUE)
attach(Gender)
gender
table(gender)

Year <- read.table("gabece/year.txt", h=TRUE)
attach(Year)
table(year)

Region <- read.table("gabece/region.txt", h=TRUE)
attach(Region)
table(region)

English <- read.table("gabece/G1.txt", h=TRUE)
attach(English)
table(G1)

Maths <- read.table("gabece/G2.txt", h=TRUE)
attach(Maths)
table(G2)

Sciences <- read.table("gabece/G3.txt", h=TRUE)
attach(Sciences)
table(G3)

SES <- read.table("gabece/G4.txt", h=TRUE)
attach(SES)
```

Journal home page: http://www.jafristatatap.net, www.projecteuclid.org/euclid.ajas
# Cleaning

# Variables including cases not presenting missing data

genderA = gender[okData == "ok"]
yearA = year[okData == "ok"]
regionA = region[okData == "ok"]
G1A = G1[okData == "ok"]
G2A = G2[okData == "ok"]
G3A = G3[okData == "ok"]
G4A = G4[okData == "ok"]

# yearly Analysis

# In the example below, you may change 2012 by any year in (2012, ..., 2017) and proceed to the analysis

yearOfStudy = 2012
regionOfStudy = "6"

genderAY = genderA[yearA == yearOfStudy]
regionAY = regionA[yearA == yearOfStudy]
G1AY = G1A[yearA == yearOfStudy]
G2AY = G2A[yearA == yearOfStudy]
G3AY = G3A[yearA == yearOfStudy]
G4AY = G4A[yearA == yearOfStudy]

# study of region

genderAR = genderA[regionA == regionOfStudy]
G1AR = G1A[regionA == regionOfStudy]
G2AR = G2A[regionA == regionOfStudy]
G3AR = G3A[regionA == regionOfStudy]
G4AR = G4A[regionA == regionOfStudy]

# study of region and year

genderAYR = genderA[(regionA == regionOfStudy) & (yearA == yearOfStudy)]
G1AYR = G1A[(regionA == regionOfStudy) & (yearA == yearOfStudy)]
G2AYR = G2A[(regionA == regionOfStudy) & (yearA == yearOfStudy)]
G3AYR = G3A[(regionA == regionOfStudy) & (yearA == yearOfStudy)]
G4AYR = G4A[(regionA == regionOfStudy) & (yearA == yearOfStudy)]

length(G4AYR)
O. Saine, S. Dembele, G. S. Lo, M. C. Haidara. Afrika Statistika, Vol. 7 (1), 2020. 829 - 850

Overview description of the Gambian GABECE Educational Data and Associated Algorithms and Unsupervised Learning Process.

#proceed to your analysis

Journal home page: http://www.jafristatap.net, www.projecteuclid.org/euclid.ajas