4open 2019, **2**, 19
© C. Cunha et al., Published by EDP Sciences, 2019
https://doi.org/10.1051/fopen/2019013

Special Issue: Statistical Inference in Copula Models and Markov Processes,
Case Studies and Insights
Guest Editor: Verónica Andrea González-López

Available online at:
www.4open-sciences.org

RESEARCH ARTICLE OPEN ACCESS

A copula-based consistency analysis of education indicators

César Cunha[1], Mariela Fernández[2], Jesús E. García[3], Verónica Andrea González-López[3,*], and Nícolas Romano[3]

[1] Department of Mathematics, University of Campinas, Sergio Buarque de Holanda, 651, Campinas, S.P., CEP 13083-859, Brazil
[2] B3, Praça Antônio Prado 48, São Paulo, S.P., CEP 01010-901, Brazil
[3] Department of Statistics, University of Campinas, Sergio Buarque de Holanda, 651, Campinas, S.P., CEP 13083-859, Brazil

Received 22 December 2018, Accepted 13 April 2019

Abstract – In this paper we investigate the consistency of quality indicators of the Brazilian public educational system. According to the newspaper Estado de São Paulo – Brazil, of January 18, 2017, only 7.3% of students in the third year of high school have an adequate level of mathematics, this shows the relevance of the evaluation and assessment of the Brazilian educational system. In this paper we explore the dependence between two indicators: (i) mean value between the proportions (in two subjects: Portuguese and Mathematics) of students *under the basic level* (SARESP classification) and (ii) rate of fails, during the years 2013, 2014 and 2015. (i) and (ii) are bases to define the educational quality of public schools for the population of young people, between 14 and 17 years old. This inspection is carried out through the Bayesian estimation of the parameters of the Asymmetric Cubic Sections (ACS) copula. We show that the dependence profile, year after year, behaves in a very unstable way, although during those years there were no substantial changes which justify such instability. Through the copula we compute conditional probabilities of tail events. We verify that an inversion occurred in the concordance/discordance between (i) and (ii). We compute the probability of (i) assuming high values, conditioned to a threshold in (ii). In 2013, as the threshold in (ii) increases the probability increases (concordance), in 2014 the threshold in (ii) is almost irrelevant to the probability and in 2015, as the threshold in (ii) increases the probability decreases (discordance). The inspection of the tail dependence allows to expose some kind of manipulation, in view of for instance, the maintenance of a global index índice de desenvolvimento da educação de São Paulo (IDESP) used to classify the educational institutions.

Keywords: Conditional probability, Asymmetric cubic sections copula, Bayesian estimation

MSC: 62H99, 62P99

1 Introduction

With information available almost constantly and coming from institutions, it is now possible to regularly review processes that impact on the life of those institutions, as is the case of institutions related to health, safety and education, among others, so that reviewing processes is a healthy task. For institutions to increase their performance, internal strategies are usually incorporated, such as measuring their processes along some period of time and using indices defined by consent. Some sectors have consolidated indices and can be used to identify performance changes. In general, indices are constructed with the intention of reproducing reality, summarizing it in just one value or few values that have simple interpretation and that are easy to calculate.

In the present study we investigate the relationship between two indicators of the Brazilian educational system. According to the newspaper Estado de São Paulo – Brazil, of January 18, 2017, only 7.3% of students in the third year of high school have an adequate level of mathematics, this shows the relevance of the constant inspection of the educational system performance. We restrict our study to the intermediate level (14–17 years old students) of public schools in the region of Guarulhos, years 2013, 2014 and 2015. Guarulhos is a city in the São Paulo state. The city of São Paulo, capital of the São Paulo state is the third most populated city in America, being behind New York and Mexico City. It is also the city with the largest Gross Domestic Product (GDP) in Latin America, which makes it a city of reference. The city of São Paulo has been approaching various municipalities of the state, because of its constant expansion. For instance, in the northeast with the municipality of Guarulhos. Guarulhos is the second most populous city in the state of São Paulo.

[*] Corresponding author: veronica@ime.unicamp.br

In this study we inspect two indicators, denoted by X and Y. These indicators compound a global index used in the state of São Paulo and called índice de desenvolvimento da educação de São Paulo (IDESP) created in 2007, http://idesp.edunet.sp.gov.br/. Thus, $X =$ the annual proportion of students classified below the baseline, per school and $Y =$ the annual failure rate, per school. Educational policies in São Paulo state, encourage the school monitoring in function of several indices, between them the IDESP. The proposal is to achieve a value of IDESP equal to or higher than five by 2030. Given that according to the Organisation for Economic Co-operation and Development (OECD) this value makes it possible to level public schools in Brazil with schools of excellence of member countries of the OECD, see more details in [1]. We see in Figures 1 and 2 that schools in Guarulhos expose a low IDESP value in comparison with the goal, despite the constant efforts made to improve their performance.

Since 2007 (year of creation of IDESP) the IDESP does not show a progressive evolution, which has led us to inspect some of its components, the most influential ones, which are X and Y. There are four levels at which students can be classified, those are: (i) under the basic level, (ii) basic level, (iii) adequate level, and (iv) advanced level, defined from an annual assessment called Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo (SARESP). Students under the basic level demonstrate insufficient mastery of the contents, the skills and the abilities desirable for the school serie in which they find themselves. For details see the next two sections of this paper. Under an ideal and simplistic perspective the variables X and Y should exhibit a linear/concordant relationship between them. In this case we do not perceive that (as we can conclude from Tab. 1), which leads us to study and model the dependence between X and Y assuming a more general approach. We use the Asymmetric Cubic Sections (ACS) copula to describe the dependence between X and Y. We perform the estimation of the parameters of the model, under a Bayesian perspective, year by year. This procedure allows us to construct annual estimates of $\text{Prob}(U > u|V > v)$ and annual estimates of the expected value $\mathbb{E}(U|V > v)$ where U are the ranks of X scaled to [0,1] and V are the ranks of Y scaled to [0,1]. In general terms, these quantities allow us to compare year by year the impact of high Y values on the values of X. More precisely, if we have observed high *failure rates*, we see how they affect the probability of high *rates of students below the baseline* and how those high *failure rates* impact in the mean value of *rates of students below the baseline*. The ACS family has already shown a good performance in applications in the area, see for example [2] and [3]. It is also compatible with our data which, as we shall see, shows very low correlation. Moreover, this family is analytically simple to treat, which facilitates its computational implementation.

In this paper, we will introduce the real problem as well as the description of the data in Section 2. Section 3 shows the model and the results. Finally we show our general conclusions in the Conclusion section, which is followed by the acknowledgments and the references.

2 Index of education development of São Paulo State

In this section we explain the construction of the IDESP and we show the reasons that lead us to study two quantities that contribute to its definition.

The SARESP system aims to evaluate the educational quality of the schools and not the performance of each student directly. This system provides different levels of classification: *under de basic*, *basic*, *adequate* and *advanced* and those levels are used to compose the IDESP, which serves as a measure of improvement in the quality of education in the state. The levels serve to diagnose the reality of the students of a given school, so it is possible through these results to develop projects in charge of the teachers of that school, in order to recover the skills not developed by that particular group of students. In the SARESP system, the classification of each student in one of the four levels is done separately in two subjects Portuguese and Mathematics. For each subject and for each school is computed the proportion of students inside each level, for *under the basic:* α_m, α_p; *basic:* β_m, β_p; *adequate:* γ_m, γ_p and *advanced:* δ_m, δ_p respectively. The quantities with subscript $m(p)$ are related to Mathematics (Portuguese) and $\alpha_m + \beta_m + \gamma_m + \delta_m = 1$, $\alpha_p + \beta_p + \gamma_p + \delta_p = 1$, respectively. Formally the IDESP index, denoted by η, is defined as follows:

$$\eta = \Delta\zeta,$$

where Δ is the mean value between Δ_m and Δ_p, $\Delta = \frac{\Delta_m + \Delta_p}{2}$, with

$$\Delta_m = \left(1 - \frac{(3\alpha_m + 2\beta_m + \gamma_m)}{3}\right)10 \text{ and } \Delta_p = \left(1 - \frac{(3\alpha_p + 2\beta_p + \gamma_p)}{3}\right)10.$$

And ζ is the proportion of approved students. For instance, when the proportion $\alpha_m = 1$, the other proportions are zero $\beta_m = \gamma_m = \delta_m = 0$, and we obtain $\Delta_m = 0$ (low quality in Mathematics). When $\delta_m = 1$, the other proportions are zero, $\alpha_m = \beta_m = \gamma_m = 0$ and $\Delta_m = 10$ (high quality in Mathematics). This means that high values of η indicate that the school shows a good overall performance. That is, as expected, high values of *under the basic* and high failure rates are indicators of poor performance, implying in low values of η. Each year, the schools receive individual goals to be achieved, and defined by the IDESP. These goals are generated by the Education Secretary (http://www.educacao.sp.gov.br/) and based on the result of the IDESP index of the previous year. When a school reaches the growth goal totally or partially, all the school

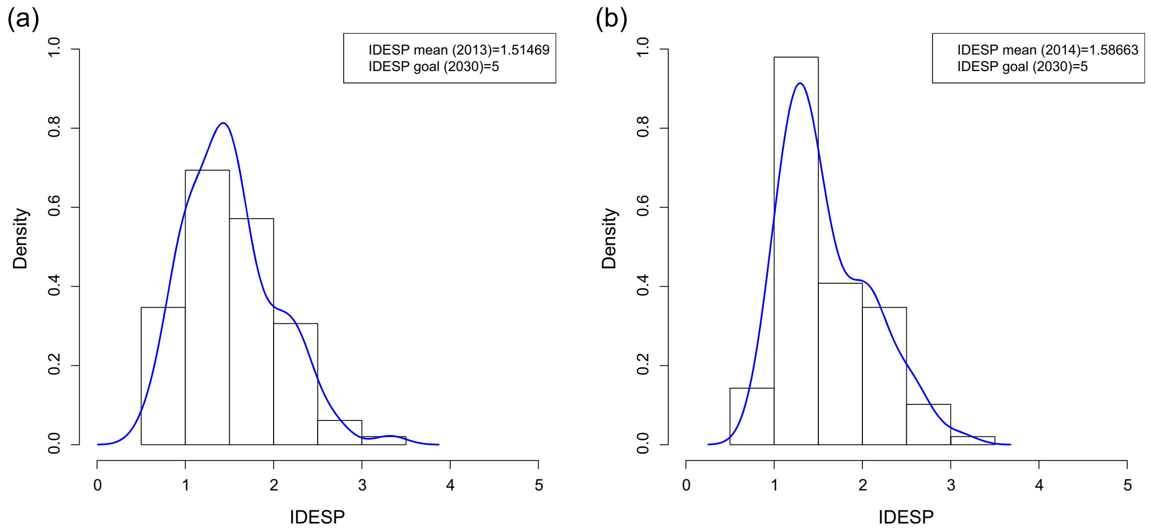

Figure 1. IDESP values distribution in Guarulhos, São Paulo State, Brazil (2013–2014). (a) Year 2013, (b) year 2014.

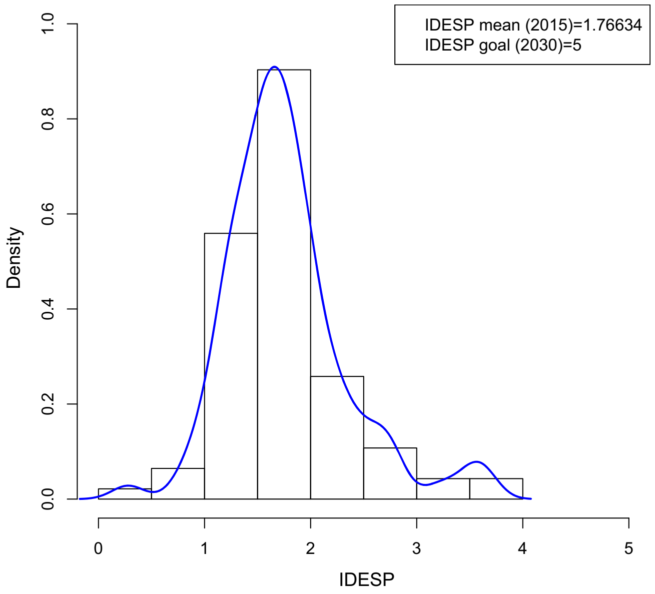

Figure 2. IDESP values distribution in Guarulhos, São Paulo State, Brazil in 2015.

Table 1. Spearman's correlation coefficient ρ between the ranks of the proportion of students classified under the basic level and ranks of the proportion of fails.

Year	2013	2014	2015
ρ	0.10493	0.04189	−0.08182

staff is awarded with a monetary complement, by merit, known as education bonus. If the school has high failure rates and a high number of students under the basic level, the school tends to have a low educational indicator, and consequently does not receive the bonus. If this continues, during three consecutive years, the school becomes a priority unit and as a consequence, the school can undergo by pedagogic interventions and detailed monitoring by the regional institution destined to do this, until the school changes its indicators. In the case of Guarulhos region this function is exercised by two sectors: *Diretoria de ensino Guarulhos Norte*, see http://deguarulhosnorte.educacao.sp.gov.br/ and *Diretoria de ensino Guarulhos Sul*, see http://deguarulhossul.educacao.sp.gov.br/. What usually occurs is that schools have high numbers of students

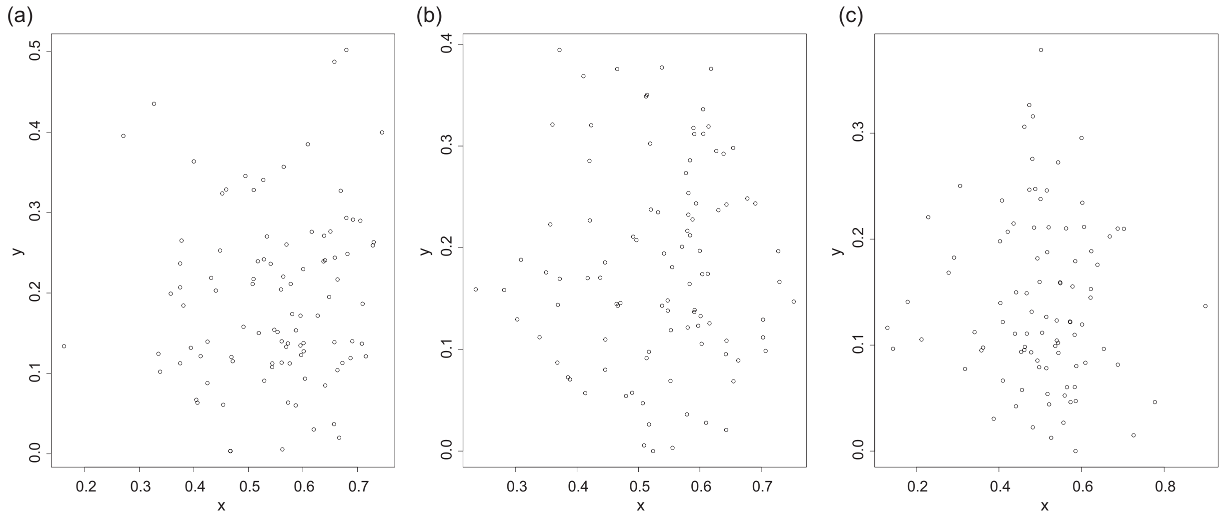

Figure 3. Plot between X (horizontal axis) and Y (vertical axis). (a) Year 2013, (b) year 2014, (c) year 2015.

classified under the basic level, which should also lead to high failure rates. But in order to mitigate the impact this would have on the indicator, the school regulates the failure rates, always keeping the same pattern in the indicator, regardless of the number of students under the basic level. That is, what regulates the promotions of the students is not how much they learn but a structural reality coming from methods of external control. This prospect is worrying, because every year some students receive the promotion to the next series without knowing the minimum required in the previous one. Consequently it becomes also more difficult to recover these students, in view of the accumulated great lag caused by this automatic promotion. It is uninteresting for a school to have high failure rates, consistent with the number of students below to the basic level, as besides impacting in the fall in the index and directly in obtaining the bonus, the school would also have more work, since it will be necessary to carry out a recovery plan, designed for these students.

2.1 Performance levels and fail rates

The data set consists of two scores X and Y recorded for each school and for the intermediate level (from 14 to 17 years old), $X =$ proportion of students classified under the basic level and $Y =$ proportion of fails for that school. We have annual data, from 2013 to 2015. Each school i receives a value $x_i = \frac{\alpha_m(i)+\alpha_p(i)}{2}$, which is the arithmetic mean between the proportion of students under the basic for each subject. For the second variable, each school i receives the value y_i, which is the proportion of fails, by year. State schools participating in this study are listed in http://www.ime.unicamp.br/~veronica/schools.htm. See also the behavior of IDESP for those years and those schools in Figures 1 and 2. Figure 3 shows the plots of the data X versus Y, for the three years.

Since our focus is to identify the dependence between X and Y, we appeal to the concepts derived from Sklar's theorem (see [4]). If X and Y have joint distribution H, with marginal distributions F and G respectively, that is, for values x and y, $F(x) = H(x, \infty)$ and $G(y) = H(\infty, y)$ there is a joint distribution C:$[0,1]^2 \to [0,1]$ (denoted by copula) such that

$$H(x, y) = C(F(x), G(y)).$$

As it is a matter of studying the dependence between X and Y, the marginal distributions F and G have nothing to report on the relationship between X and Y. Also, if we define $U = F(X)$ and $V = G(Y)$ the concordance/discordance between X and Y is preserved by U and V, since functions F and G are non-decreasing monotone functions. A natural representation of the values of U (and V respectively) are the empirical ranks of the observations scaled to [0,1] of X (and Y respectively). With this purpose, we compute the pseudo-observations $\hat{u}_i = \hat{F}(x_i)\frac{n}{n+1}$ and $\hat{v}_i = \hat{G}(y_i)\frac{n}{n+1}$ where $i = 1, \ldots, n$, $\hat{F}$ and $\hat{G}$ are the empirical distributions of X and Y, respectively and n denotes the number of observations (schools).

In Table 1 we expose the Spearman's correlation coefficients between X and Y, year by year.

We can note the low values of the Spearman's correlation coefficient ρ although both variables are related with an unsatisfactory performance and, by coherence need to be associated. We note the inability of the Spearman's correlation coefficient to capture dependence by showing a negative value in 2015. The results of Table 1 only means that is not identified a linear relation between the ranks of the observations, so the alternative is to use a non-linear model to represent the dependence between the rates. Thus, the focus of our study is the identification of C. To get to this identification we will delimit the possibilities of C into a sufficiently flexible family.

3 Model and results

Here we introduce the model explored in this paper. This model corresponds to a family of copulas that is a perturbation of the case of independence, that is $C(u,\ v) = uv$. With this proposal we seek to contemplate also situations of low correlation, as shown in Table 1.

Definition 3.1. *The biparametric ACS copula family is given by* $C(u,v|(a,b)) = uv + uv(1-u)(1-v)[(a-b)v(1-u)+b]$, *where* $(a,b)\in \Theta = \{|b|\le 1,\ \frac{b-3-\sqrt{9+6b-3b^2}}{2} \le a \le 1 \text{and}\, a \ne b\}$, $u, v \in [0,1]$.

Its density function is

$$c(u,v|(a,b)) = 1 + (a-b)\left(1-4u+3u^2\right)\left(2v-3v^2\right) + b(1-2u)(1-2v), \tag{1}$$

and the Spearman's correlation coefficient $\rho = \frac{a+3b}{12}$. It should be noted that if $a = b$, the copula corresponds to the Farlie-Gumbel-Morgenstern family $C(u,v|b) = uv + uv(1-u)(1-v)b$, which admits fragile positive as well as negative degrees of Spearman's correlation. Precisely, since the parameter $b \in [-1,1]$ then $\rho \in [-\frac{1}{3},\frac{1}{3}]$. Given the correlation spectrum allowed by the Farlie-Gumbel-Morgenstern family and according to the results of Table 1, our data could respond to this model. Thus, in relation to the estimation of parameters a and b, if they were similar, we could argue that the dependence between U and V is well represented by the Farlie-Gumbel-Morgenstern model.

Looking to explore stochastic-functional relationships between U and V, [5] shows a method of constructing copulas with the property of having cubic cross-sections, one of these models is given by the Definition 3.1. For instance, if we fix $v = v_0$ in Definition 3.1, we obtain:

$$C(u,v_0|(a,b)) = \theta(v_0)u + \vartheta(v_0)u^2 + \kappa(v_0)u^3,$$

with $\theta(v_0) = (v_0 + v_0(1-v_0)[(a-b)v_0 + b])$, $\vartheta(v_0) = (2v_0^2(1-v_0)(b-a) - bv_0(1-v_0))$ and $\kappa(v_0) = v_0^2(1-v_0)(a-b)$. Then, the copula is given by a cubic expression in u. Analogously, if we set $u = u_0$, the expression in Definition 3.1 corresponds to a cubic expression in v. In terms of the modeling process, these cubic forms aim to give greater flexibility to the dependence type between U and V, being this more general than a linear dependence type.

Given a specific year we compute the likelihood function of the sample of size n, that is

$$\prod_{i=1}^{n} c(u_i, v_i|(a,b)),$$

where the function c is given by the equation (1). Assuming a non-informative prior distribution on $(a,b)\in \boldsymbol{\Theta}$, $\pi(a,b) \propto 1$, the posterior distribution of $(a,\ b)$ is proportional to the likelihood function. We use a non-informative prior distribution on $(a,\ b)$ in order to contain the impact of the prior distribution in the posterior distribution of $(a,\ b)$. We also observe that the complexity of the parametric space Θ (see Definition 3.1) could hinder the use of an informative prior distribution without a very solid base. About literature linking copula's theory and Bayesian estimation, see [6] and [3]. The Bayesian estimates of a and b, under quadratic loss function, for each year are shown in Table 2. In 2015, five schools did not participate in the study, these are: *Profa. Alice Chuery*, *Conselheiro Crispiniano*, *Hugo de Aguiar*, *Profa. Ilia Zilda Innocenti Blanco* and *Vila Any*.

We see that in none of the three cases the model indicates the Farlie-Gumbel-Morgenstern copula, since the estimates of a and b look very different. A Bayesian approach is appropriate in those cases for several reasons, between them we note: a moderate sample size to implement a frequentist estimation of two parameters and the constrains over the parameters a and b. We estimate the probability $\text{Prob}(U > u|V > v)$ and the expected value $\mathbb{E}(U|V > v)$ by means of the values reported in Table 2, as follows. If X and Y are continuous with cumulative distributions F and G respectively, given $U = F(X)$ and $V = G(Y)$ with 2-copula C, $\text{Prob}(U > u|V > v) = \frac{1-u-v+C(u,v)}{1-v}$. Then, using the Definition 3.1 we can define the estimation of $\text{Prob}(U > u|V > v)$ as

$$\hat{P}(U > u|V > v) = \frac{1-u-v+C(u,v|(\hat{a},\hat{b}))}{1-v}, \quad u, v \in [0,1]\,. \tag{2}$$

Table 2. Bayesian estimators of a and b – see Definition 3.1.

Year	n = Sample size	$\hat{a}$	$\hat{b}$
2013	98	−0.65852	0.54581
2014	98	0.26329	−0.03267
2015	93	0.20679	−0.44183

Since $\text{Prob}(U \leq u|V > v) = \frac{u}{1-v} - \frac{C(u,v)}{1-v}$ and $C(u,v) = \int_0^u \frac{\partial}{\partial s} C(s,v)\mathrm{d}s$, then

$$\text{Prob}\,(U \leq u|V > v) = \frac{u}{1-v} - \frac{1}{1-v}\int_0^u \frac{\partial}{\partial s} C(s,v)\mathrm{d}s,$$

as a consequence

$$\mathbb{E}(U|V > v) = \frac{1}{2(1-v)} - \frac{1}{1-v}\int_0^1 u\frac{\partial}{\partial u} C(u,v)\mathrm{d}u. \tag{3}$$

Computing the partial derivative of the copula given by Definition 3.1 we obtain from the equation (3), $\mathbb{E}(U|V > v) = \frac{1}{2} + \frac{b}{6}v + \frac{(a-b)}{12}v^2$. Then, we propose the estimation:

$$\hat{\mathbb{E}}(U|V > v) = \tfrac{1}{2} + \tfrac{\hat{b}}{6}v + \tfrac{(\hat{a}-\hat{b})}{12}v^2, \quad v \in [0,1]\,. \tag{4}$$

Returning to the real problem, we expect the variables $X =$ *proportion of students classified under the basic level* and $Y =$ *proportion of fails* for that school, to show a performance compatible with what they are measuring. To investigate in detail the coherence in the dependence between X and Y, observed year after year, we first focus on the conditional dependence between tail events, estimated by the equation (2), then we show a more traditional study on the mean value of U (ranks of X) conditioned to thresholds in V (ranks of Y) estimated by equation (4).

3.1 Conditional tail dependence

The most reasonable behavior of (2) is to show an increasing tendency in the upper tail. This is, it is expected that high values of U to be concentrated with high values of V. We will show what we verify in the estimates, for certain values of U (ranks of X) and in relation to all possible values of V (ranks of Y). The behavior of (2), year by year is illustrated in Figures 4 and 5, for the cases $u = 0.5$, 0.7 and 0.9. See Table 3, for other values of u.

In 2013, (2) is given by a concave quadratic curve. We note that as u increases (2) changes by being formed only by the increasing part of the curve, also its concavity is less pronounced, revealing an almost linear and increasing aspect in the case of $u = 0.9$. The curves (2) of 2014 and 2015 are convex quadratic curves. For the year 2014, we see that as u grows, the curve goes taking a constant aspect. We can also verify this fact by inspecting Table 3 (case 2014). This statement can be better visualized in the Figure 6. For instance, given any threshold v, the probability of $U > 0.9$ is almost constant. In practical terms this means that large proportions of students below the basic level do not depend on any failure rate. Evidently, this exposes an extreme contradiction. In the case of 2015, we observe that as u grows the curve loses its convexity and exhibits an almost linear and decreasing behavior, for large threshold values in U (see also Tab. 3). That is, the higher the threshold in V, the smaller the chance of U exceeding values close to 1.

Since the dependence between X and Y is the same as the dependence between U and V, we see how there was a concrete deterioration from 2013 to 2015, of the relationship between X and Y. Arriving at the point of showing conditional discordance between X and Y (in 2015) and going through conditional independence (in 2014), which does not make sense from the meaning of the variables.

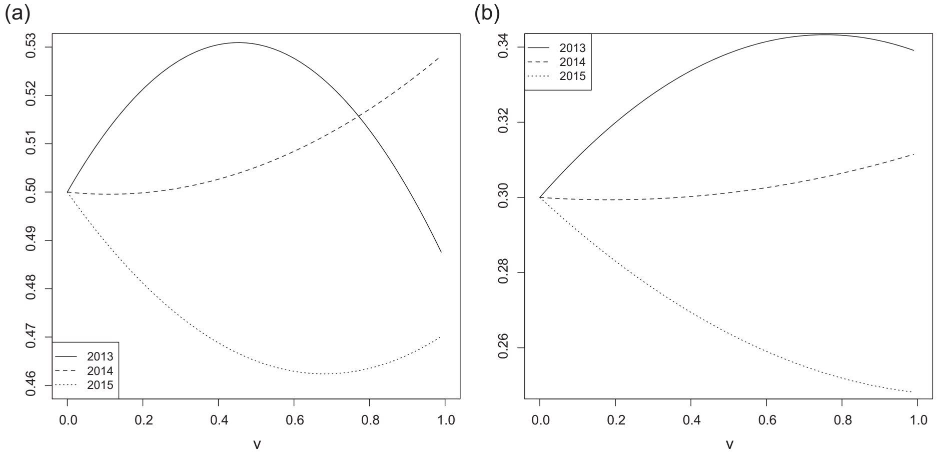

Figure 4. $\hat{P}(U > u_0|V > v)$ according to equation (2), for $v \in [0,1]$. (a) $u_0 = 0.5$, years: 2013, 2014 and 2015. (b) $u_0 = 0.7$, years: 2013, 2014 and 2015.

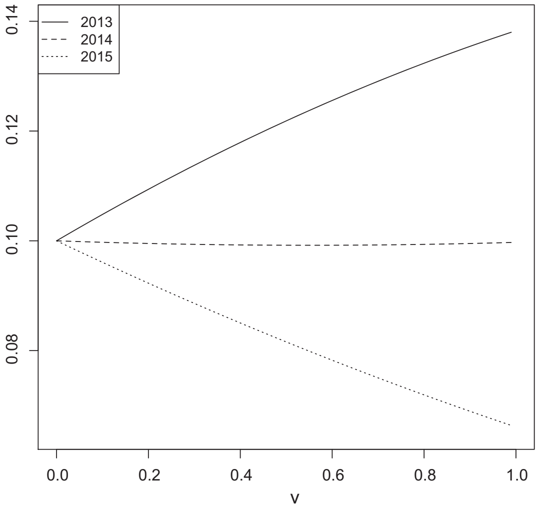

Figure 5. $\hat{P}(U > u = 0.9|V > v)$ according to equation (2), for $v \in [0, 1]$, years: 2013, 2014 and 2015.

Table 3. $\hat{P}(U > u|V > v)$ according to equation (2) with u, v = 0.5, 0.6, 0.7, 0.8 and 0.9.

u		0.5	0.6	0.7	0.8	0.9
$v = 0.5$	2013	0.53059	0.43659	0.33834	0.23403	0.12185
	2014	0.50516	0.40318	0.30123	0.19975	0.09920
	2015	0.46504	0.36255	0.26382	0.16984	0.08158
$v = 0.6$	2013	0.52768	0.43697	0.34146	0.23852	0.12557
	2014	0.50842	0.40552	0.30260	0.20027	0.09920
	2015	0.46291	0.35879	0.25904	0.16506	0.07824
$v = 0.7$	2013	0.52175	0.43504	0.34306	0.24225	0.12907
	2014	0.51241	0.40843	0.30433	0.20098	0.09925
	2015	0.46241	0.35628	0.25507	0.16069	0.07503
$v = 0.8$	2013	0.51282	0.43080	0.34314	0.24520	0.13236
	2014	0.51714	0.41191	0.30644	0.20188	0.09935
	2015	0.46352	0.35502	0.25193	0.15673	0.07192
$v = 0.9$	2013	0.50087	0.42425	0.34170	0.24738	0.13543
	2014	0.52261	0.41596	0.30893	0.20297	0.09951
	2015	0.46626	0.35500	0.24959	0.15319	0.06894

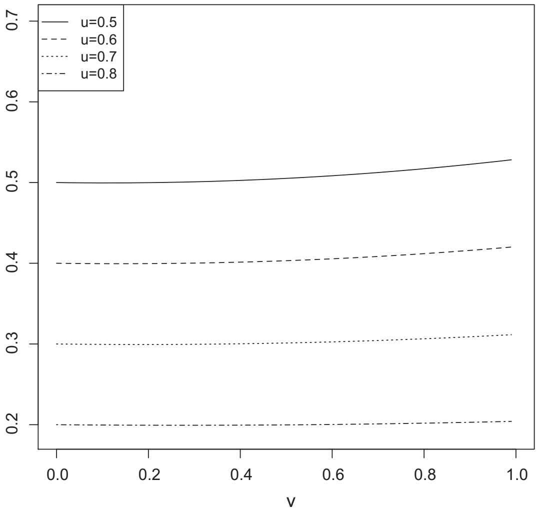

Figure 6. $\hat{P}(U > u|V > v)$ according to equation (2), for $v \in [0, 1]$ and year: 2014.

3.2 Central tendency

To build a global view of the behavior of U (ranks of students classified under the basic level) conditioned to values of V (ranks of fails) that exceed a threshold v, we will estimate $\mathbb{E}(U|V > v)$ by equation (4). When comparing the 3 years, a similar behavior of (4) is expected. Since we are inspecting consecutive years where non changes happened in the educational system. Figure 7 and Table 4 show the results.

We note that, the relationship between U and V exhibits different behaviors, when considered during these 3 years, one is a concave function and two are convex functions (see also Fig. 4). This fact shows the lack of robustness of the process of dependence between X and Y. We can compare the behavior of $\hat{\mathbb{E}}(U|V > v)$ with the conditional probability $\hat{P}(U > u_0|V > v)$ where u_0 is the value corresponding with the median of X, as listed by Table 5.

We verify that the functional performance of $\hat{\mathbb{E}}(U|V > v)$ (Fig. 7) and $\hat{P}(U > u_0|V > v)$ (Fig. 8) is similar as already anticipated when comparing Figures 4(a) and 7. In Table 4 we show the values given by equation (4) for $v = 0.2, 0.3, \ldots, 0.9$. Consider the year 2013 and $v = 0.8$, the expected value of X scaled into [0,1] under the condition $\{V > 0.8\}$ is approximately 0.50854 and $\hat{\mathbb{E}}(U|V > 0.8)$ belongs to the interval [0.47568, 0.51143] in the period: 2013–2015. In Table 6 we show the values given by (4) for $v = 0.2, 0.3, \ldots, 0.9$. So, the probability of U to exceed 0.49495 is approximately 0.51689 in 2013, under the condition $\{V > 0.8\}$, and $\hat{P}(U > u_0|V > 0.8)$ belongs to the interval [0.46352, 0.52243] in the period: 2013–2015.

These results lead us to observe Table 1, where the Spearman's correlation coefficient exposes its fragility. In the same way, it is to be expected that the mean values computed here do not clearly point out what is happening, in the tail region of $[0,1]^2$ where we are interested in tracking the concordance/discordance between U and V. This fact justifies the previously developed conditional study.

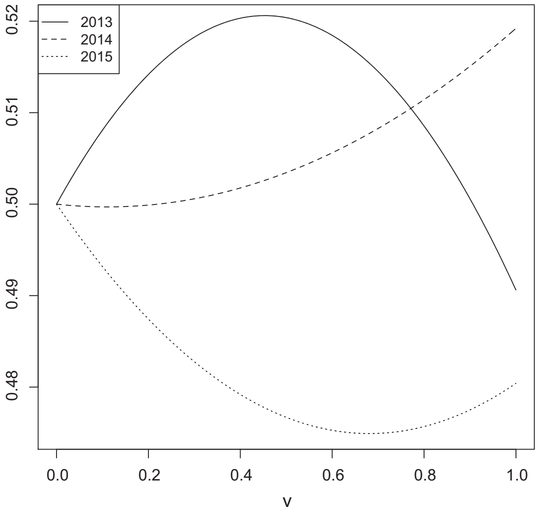

Figure 7. $\hat{\mathbb{E}}(U|V > v)$ according to equation (4) for $v \in [0, 1]$, years: 2013, 2014 and 2015.

Table 4. $\hat{\mathbb{E}}(U|V > v)$ according to equation (4), years 2013, 2014 and 2015.

v	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2013	0.51418	0.51826	0.52033	0.52039	0.51845	0.51450	0.50854	0.50058
2014	0.49990	0.50059	0.50177	0.50344	0.50561	0.50827	0.51143	0.51508
2015	0.48743	0.48277	0.47919	0.47669	0.47528	0.47494	0.47568	0.47751

Table 5. Median values of X and its corresponding u_0.

Year	Median of X	u_0
2013	0.56410	0.49495
2014	0.55220	0.49495
2015	0.51465	0.50000

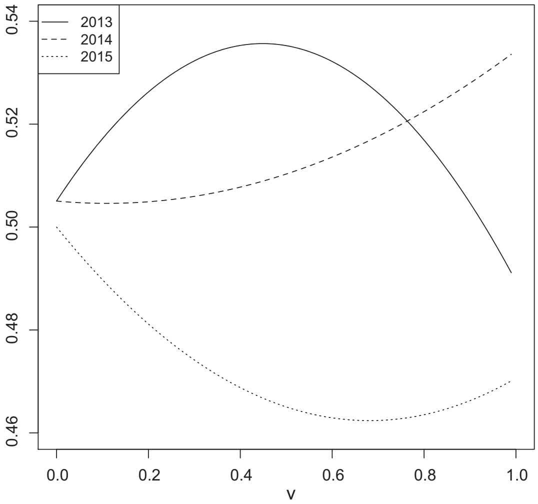

Figure 8. $\hat{P}(U > u_0|V > v)$ according equation (2), $v \in [0, 1]$, years: 2013, 2014 and 2015, with u_0 given by Table 5.

Table 6. $\hat{P}(U > u_0|V > v)$ according equation (2), years 2013, 2014 and 2015, with u_0 given by Table 5.

v=	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2013	0.52626	0.53230	0.53530	0.53526	0.53218	0.52605	0.51689	0.50469
2014	0.50491	0.50596	0.50776	0.51031	0.51360	0.51764	0.52243	0.52797
2015	0.48115	0.47416	0.46879	0.46504	0.46291	0.46241	0.46352	0.46626

4 Conclusion

In this paper we explore the dependence between two indicators: (i) mean between the proportions (in Portuguese and Mathematics) of students under the basic level (SARESP classification) and (ii) rate of fails, during the years 2013, 2014 and 2015. The data is coming from around 100 public schools of the Guarulhos city, the second largest city of the São Paulo state. The inspection of the dependence is carried out by means of a Bayesian copula estimation, through the Bayesian estimation of the parameters of the ACS copula, a model adopted for its flexibility. We show that the dependence profile, year after year, behaves in a very unstable way, although during those years there were no substantial changes which justify such variable behavior. The Bayesian point estimation of the parameters indicates this instability, see Table 2 and also confirmed by the influence of those estimations in the mean conditional curve given by equation (4). The mean value of the ranks of (i) conditioned to a threshold in (ii) shows a very different behavior when we compare the 3 years. According to the indications reported by Table 1, global measures, such as those computed via the conditional mean value (4) may not be appropriated to identify what is happening. Since, is suspected that some kind of handling may exists in (i) and/or (ii), due to the structural aspects of the educational system, which could explain the difference in dependence profiles, as is the case of Figure 7. To understand the relation between (i) and (ii) we inspect the conditional dependence in different upper tail regions of $[0,1]^2$ of the marginal ranks of (i) and (ii) scaled to [0,1]. We can see the representation of the behavior of tail events given by equation (2) in Figure 5. We see that in 2013 the behavior of the conditional probability is the expected, since, the higher threshold in rate of fail, the higher the probability of classification under the basic level be superior to 90%. In 2014, the thresholds of rate of fail do not influence the probability of classification under the basic level being greater than 90%. In 2015, to higher threshold in rate of fail is lesser the probability of classification under the basic level be superior to 90%. That is to say that the relation of concordance between (i) and (ii) verified in 2013 is inverted for discordance in 2015, precisely in the most critical values which are high failure rates and high proportions under the basic level.

Based on the study, we perceive the need to review the use of global indices such as the IDESP, for the development of policies to control the quality of education. As illustrated in Figure 1, the IDESP appears to exhibit some stability or very slight improvement and at the same time is able to mask relevant and decisive aspects for quality in education. More precisely, it allows mitigating the effects of relevant indicators, as the case of (i) and (ii).

Acknowledgments

The authors gratefully acknowledge the support for this research provided by the Graduate Program in Professional Masters National Network – PROFMAT to C. Cunha and by CAPES with a fellowship of the Master Program in Statistics - University of Campinas to N. Romano. Also, the authors wish to thank anonymous reviewers for their many helpful comments and suggestions on an earlier draft of this paper.

References

1. Cunha C (2017), Estudo sobre Componentes do IDESP na cidade de Guarulhos, Unpublished master's thesis. University of Campinas, Campinas, Brazil.
2. Fernández M, González-López VA (2013), A copula model to analyze minimum admission scores, in: AIP Conference Proceedings, 1558, 1479–1482.
3. Fernández M, González-López VA, Rifo LLR (2015), A note on conjugate distributions for copulas. Math Methods Appl Sci 38, 18, 4797–4803.
4. Sklar A (1959), Fonctions de répartition à n dimensions et leurs marges. Publ Inst Statist Univ Paris 8, 229–231.
5. Nelsen RB, Quesada Molina JJ, Rodríguez Lallena JA (1997), Bivariate copulas with cubic sections. J Nonparametr Statist 7, 205–220.
6. García Jesús E, González-López VA, Nelsen RB (2016), The structure of the class of maximum Tsallis-Havrda-Chavát entropy copulas. Entropy 18, 7, 264.

Cite this article as: Cunha C, Fernández M, García JE, González-López VA & Romano N, 2019. A copula-based consistency analysis of education indicators. 4open, **2**, 19.