Evidências do *Google Trends* de uma Crescente Exclusão Digital de Segundo Nível no Brasil

(Detailed local statistical analysis)

dos Santos, Renato P.[a]
Bülbül, M. Şahin[b]
Lemes, Isadora L.[a]

[a] Universidade Luterana do Brasil, Programa de Pós-Graduação em Ensino de Ciências e Matemática, Canoas, RS, Brasil
[b] Kafkas Üniversitesi, Dede Korkut Eğitim Fakültesi, Kars, Türkiye

The number of individuals who accessed the Internet in each state, as provided by CETIC.br data, and the rates of Internet searches for various relevant topics across Brazilian states in 2005, 2011, 2008, and 2013, as provided by downloaded GT data were analysed in detail, looking for spatial inequalities across the individual municipalities employing measures of inequality, such as the *Gini coefficient*, and measures of spatial autocorrelation, such as *Moran's I* statistic, first proposed by Moran (1948). As usual, near-zero values for Gini coefficient and Moran's I statistics indicate perfect equality and no spatial autocorrelation (random spatial distribution), respectively. In contrast, positive values indicate positive spatial autocorrelation, with spatial clusters of similarly low or high values between neighbour municipalities, and negative values indicate negative spatial autocorrelation, in which low values tend to have neighbours with high values and vice versa.

However, it must be noticed that classical Moran's I statistic and Gini coefficient are both whole-map, locationally invariant measures, in the sense that they can tell whether something is happening, but not where it is happening within the region of interest (Rey & Smith, 2013). Therefore, a spatial decomposition of the Gini coefficient was also done, according to a method introduced by (2013) that supports the detection of spatial autocorrelation and segregation, as provided by the *lctools* package (Kalogirou, 2019). Furthermore, the *lctools* package also provides a Monte Carlo simulation, in which the data are spatially reallocated in a random way to infer the share of overall inequality that is associated with non-neighbour pairs of locations and, therefore, to assess the significance of the evaluated Spatial Gini coefficient (Rey & Smith, 2013). Results are shown in Table 1 and Table 2.

Table 1

Measures of spatial autocorrelation and segregation for the indicators across the individual states.

Indicator	Year	N	Gini	gw.frac	ns.frac	*p*
Use of Internet	2005	26	0.21*	0.34	0.66	0.05
	2008	26	0.12*	0.36	0.64	0.05
	2011	26	0.12*	0.36	0.64	0.05
	2013	26	0.11*	0.37	0.63	0.05
Searches for ‘Biology’	2005	8	0.08	0.24	0.76	0.25
	2008	10	0.12*	0.26	0.74	0.05
	2011	22	0.18*	0.30	0.70	0.05
	2013	23	0.16	0.42	0.58	0.40
Searches for ‘Chemistry’	2005	8	0.07	0.26	0.74	0.50
	2008	11	0.08	0.31	0.69	0.10
	2011	23	0.13*	0.35	0.65	0.05
	2013	24	0.16	0.33	0.67	0.10
Searches for ‘Mathematics’	2005	9	0.10	0.23	0.77	0.10
	2008	14	0.10*	0.30	0.70	0.05
	2011	23	0.13*	0.34	0.66	0.05
	2013	25	0.17*	0.36	0.64	0.05
Searches for ‘Globo’	2005	15	0.11*	0.32	0.68	0.05
	2008	20	0.09	0.35	0.65	0.10
	2011	26	0.09	0.39	0.61	0.15
	2013	26	0.11*	0.36	0.64	0.05
Searches for ‘games’	2005	19	0.17	0.35	0.65	0.10
	2008	24	0.14*	0.37	0.63	0.05
	2011	26	0.08	0.38	0.62	0.20
	2013	26	0.08	0.38	0.62	0.10

Note: * indicates $p < .5$. ** indicates $p < .01$. **gw.frac** is the fraction of the first component of the spatial Gini. **ns.frac** is the fraction of the second component of the spatial Gini. ***p*** is the pseudo p-value of significance calculated from the Monte Carlo simulation. (Kalogirou, 2019).

Table 2

Measures of spatial autocorrelation for the indicators across the individual states.

Indicator	Year	N	Moran.I	EI	z.res	z.rand	*p.rsamp*	*p.rand*
Use of Internet	2005	26	0.23**	-0.04	8.51	8.44	1.7E-17	3.3E-17
	2008	26	0.22**	-0.04	8.14	8.09	4.0E-16	6.1E-16
	2011	26	0.22**	-0.04	8.10	8.05	5.6E-16	8.0E-16
	2013	26	0.20**	-0.04	7.50	7.44	6.3E-14	9.8E-14
Searches for 'Biology'	2005	8	0.05*	-0.14	1.93	2.26	0.05	0.02
	2008	10	0.35**	-0.11	5.29	5.12	1.2E-07	3.0E-07
	2011	22	0.23**	-0.05	6.73	6.66	1.7E-11	2.7E-11
	2013	23	-0.02	-0.05	0.79	0.78	0.43	0.44
Searches for 'Chemistry'	2005	8	-0.17	-0.14	-0.30	-0.30	0.77	0.77
	2008	11	0.25**	-0.10	4.04	3.84	5.4E-05	1.2E-04
	2011	23	0.12**	-0.05	4.46	4.42	8.2E-06	9.7E-06
	2013	24	0.03*	-0.04	2.10	2.64	0.04	8.2E-03
Searches for 'Mathematics'	2005	9	0.19**	-0.12	3.99	3.97	6.6E-05	7.3E-05
	2008	14	0.16**	-0.08	3.50	3.39	4.6E-04	6.9E-04
	2011	23	0.16**	-0.05	5.43	5.35	5.6E-08	9.0E-08
	2013	25	0.04**	-0.04	2.60	2.72	9.3E-03	6.6E-03
Searches for 'Globo'	2005	15	0.16**	-0.07	3.62	3.63	3.0E-04	2.9E-04
	2008	20	0.13**	-0.05	3.94	3.94	8.2E-05	8.2E-05
	2011	26	-0.05	-0.04	-0.37	-0.38	0.71	0.70
	2013	26	0.08**	-0.04	3.69	3.69	2.2E-04	2.2E-04
Searches for 'games'	2005	19	0.04*	-0.06	1.94	1.96	0.05	0.05
	2008	24	-0.04	-0.04	0.01	0.01	0.99	0.99
	2011	26	-0.05	-0.04	-0.29	-0.31	0.77	0.76
	2013	26	-0.06	-0.04	-0.73	-0.75	0.47	0.45

Note: * indicates $p < .5$. ** indicates $p < .01$. **EI** is the Expected Moran's I, to be compared with the Classic global Moran's I statistic **Moran.I** value. **z.res** is the z score calculated for the resampling null hypotheses test. **z.rand** is the score calculated for the randomization null hypotheses test. **p.rsamp** is the p-value (two-tailed) calculated for the resampling null hypotheses test. **p.rand** is the p-value (two-tailed) calculated for the randomization null hypotheses test. (Kalogirou, 2019).

From results in Table 1 and Table 2, there have been a noticeable spatial inequality in the rate of use of Internet across the country in the period 2005-2013. Rates of searches for Mathematics also show statistically significant spatial inequality in the entire period, differently from searches for Chemistry and Biology which were significant in certain years only. It should be noticed that Internet access was not still fully widespread in 2005 and, therefore, some Northern and North-eastern states had not enough search volumes for those terms as to be included in Google Trends results. Rates for searches for terms related to entertainment, and of "scarce social prestige" (Vogt & Castelfranchi, 2009, p. 23) such as Globo and games were significant only on 2005 as the inhabitants of those states also seem to have 'evolved,' in the sense of gradually learning to use the access also as a source for leisure, rather than learning only, as they become more used to the Internet.

Finally, scatterplots were drawn showing the rate of Internet access and the rate of Internet searches for a given term, values for each state and confidence ellipses, based on multivariate *Student-t* distributions, with a standard confidence level of 0.95, making use of the *ggplot2* package (Wickham & Chang, 2019, Chapter stat_ellipse) for the R statistical data analysis language (R Core Team, 2019). When the confidence ellipse is 'tilted,' the explanatory variables are correlated; in contrast, when its axes are parallel to the axes of the parameter space, the explanatory variables are uncorrelated (Fox, 2016, p. 221).

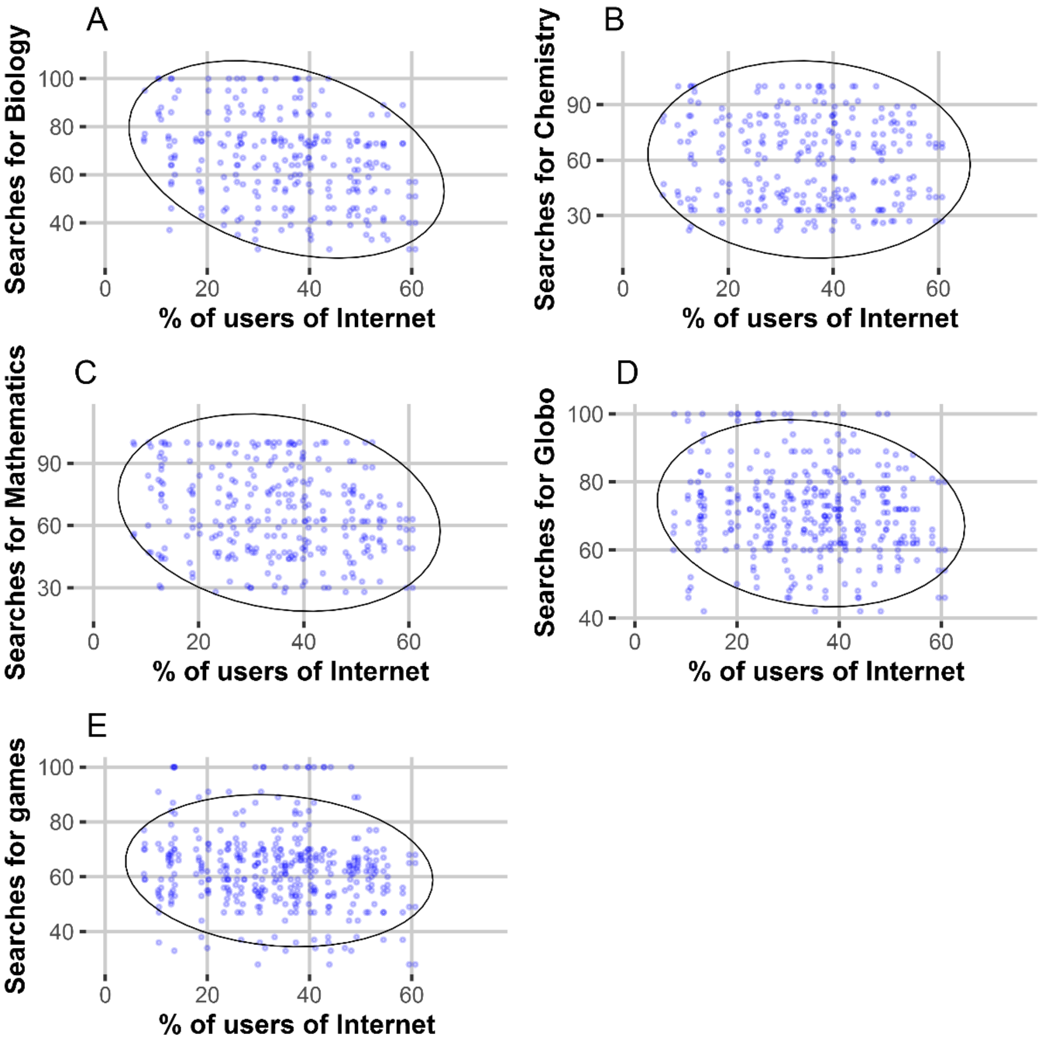

Figure 1. Scatterplots showing rates of access to the Internet and relative search volumes[a] for relevant terms for each state.

[a] These volume figures are relative and normalised, as discussed in Section 2.2.

As the major axis of the confidence ellipses in the scatterplot in Figure 2 for searches for 'Chemistry' and 'games' are almost parallel with the x-axis, very weak correlations with the rates of access to the Internet are expected for these two variables;

conversely, as the confidence ellipses for searches for ‘Biology’ and ‘Mathematics’ are the most ‘tilted’ ones, these two variables are expected to be the best-correlated ones with the rates of access to the Internet (Fox, 2016, p. 221).

Furthermore, *Pearson’s Correlation Coefficients* (PCC) could be calculated between the number of people who accessed the Internet in each state, as provided by CETIC.br data, and the rates of Internet searches for various relevant topics across Brazilian states in 2005, 2011, 2008, and 2013, as provided by downloaded GT data. A PCC value near +1 indicates that the highest the rate of Internet access in that state during that period, also the highest the rate of Internet searches for a given term; contrariwise, a PCC value near -1 indicates that either there happened many searches in a state with reduced access to the Internet or vice-versa.

However, the Shapiro-Wilk test, as provided by the *stats* package, which part of R-language itself, was done to verify the normality of the data distribution.

Table 3
Results of the Shapiro-Wilk test.

Indicator	***W***	***P***
Use of Internet	0.98	0.058
Searches for 'Biology'	0.98	0.255
Searches for 'Chemistry'	0.94**	0.005
Searches for 'Mathematics'	0.96*	0.017
Searches for 'Globo'	0.98	0.288
Searches for 'games'	0.96**	0.004

Note: *W* is the value of the Shapiro-Wilk statistic, and p is an approximate (for $n > 3$) p-value for the test. According to Royston (1995), the approximation produced by this algorithm “is acceptable for practical data analysis for $p < 0.1$, the critical region of most interest.” * indicates $p < .5$. ** indicates $p < .01$.

From Table 3, the p-values for ‘Chemistry,’ ‘Mathematics,’ and ‘games’ are less than the significance level 0.05 (with the p-value for ‘Use of Internet’ also not far from it) implying that the distributions of the data for these searches are not to be assumed as normal. Therefore, the non-parametric Spearman correlation test should be used instead of the usual Pearson’s one (Xu, Hou, Hung, & Zou, 2013) to investigate the possible correlations between the rates of access to the Internet and those relative search volumes.

In the following, as such “global” statistics are likely to hide significant spatial variation of the relationship between two variables (Kalogirou, 2015), we moved to use *Local Pearson’s Correlation Coefficients* (LPCC) based on a fixed number of nearest neighbours, as proposed by Kalogirou (2012, 2013) and provided by the *lctools* package

(Kalogirou, 2019). As argued by Kalogirou (2012), these LPCC would allow the identification of pairs of variables that not significantly correlated globally but higher correlated locally. The *lctools* package for R also provides a Monte Carlo simulation proposed by Hope (1968) and adapted by Fotheringham, Brunsdon, and Charlton (2002) to assess whether the spatial variation of the local correlation coefficients is statistically significant.

Table 4

Pearson's (r), Kendall's (τ), and Spearman's (ρ) correlation coefficients and respective p-values for the correlations between the rates of access to the Internet and relative search volumes[a] *for relevant terms.*

Correlation	Year	N	r	p	τ	p	ρ	p
Searches for 'Biology'	2005	8	-0.91**	1.96E-03	-0.84**	4.14E-03	-0.92**	1.11E-03
	2008	10	-0.90**	4.13E-04	-0.73**	2.21E-03	-0.85**	3.50E-03
	2011	22	-0.77**	2.48E-05	-0.59**	1.23E-04	-0.75**	5.97E-05
	2013	23	-0.46*	2.80E-02	-0.31*	3.67E-02	-0.41	5.17E-02
Searches for 'Chemistry'	2005	8	-0.54	0.17	-0.33	0.26	-0.46	0.26
	2008	11	-0.79**	3.73E-03	-0.61**	9.74E-03	-0.79**	3.48E-03
	2011	23	-0.67**	4.27E-04	-0.45**	2.82E-03	-0.61**	1.79E-03
	2013	24	-0.20	0.36	-0.22	0.14	-0.34	0.11
Searches for 'Mathematics'	2005	9	-0.87**	2.12E-03	-0.54*	4.64E-02	-0.73*	2.62E-02
	2008	14	-0.64*	1.37E-02	-0.54**	6.74E-03	-0.67*	1.01E-02
	2011	23	-0.67**	4.49E-04	-0.48**	1.51E-03	-0.62**	1.75E-03
	2013	25	-0.31	0.14	-0.28	4.95E-02	-0.39	5.64E-02
Searches for 'Globo'	2005	15	-0.75**	1.14E-03	-0.64**	1.04E-03	-0.81**	2.43E-04
	2008	20	-0.74**	2.11E-04	-0.56**	6.36E-04	-0.71**	4.31E-04
	2011	26	-0.07	0.74	-0.06	0.68	-0.08	0.70
	2013	26	+0.24	0.23	+0.15	0.29	+0.23	0.25
Searches for 'games'	2005	19	-0.39	9.76E-02	-0.31	6.34E-02	-0.43	6.78E-02
	2008	24	-0.35	8.93E-02	-0.31*	3.47E-02	-0.44*	3.24E-02
	2011	26	+0.03	0.89	+0.00	1.00	+0.00	0.98
	2013	26	-0.12	0.56	-0.14	0.32	-0.16	0.44

Note: [a] These volume figures are relative and normalised, as discussed in Section 2.2.
* indicates $p < .5$. ** indicates $p < .01$.

From Table 4, there are significant negative correlations ($p < 0.05$) between relative search volumes for 'Chemistry,' and 'Mathematics' and the rates of access to the Internet, be it measured by Pearson's (r), Kendall's (τ), or Spearman's (ρ) correlation coefficients; conversely, searches for 'Chemistry' and 'games' have no definitely significant correlations with the rates of access to the Internet. These results are in accordance with the analysis of the scatterplots of Figure 1.

However, when we move to the LPCC statistics across each state, Table 5 shows that most of those coefficients are not statistically significant, with the exception for searches for ‘Biology’ in 2011.

Table 5

Percentages and ranges of statistically significant ($p \leq 0.5$) correlations, as measured by LPCC values between rates of access to the Internet and relative search volumes[a] *for relevant terms*

Relation	**Year**	**N**	**% of $p \leq 0.05$**
Searches for ‘Biology’	2005	8	0.0%
	2008	10	10.0% [-0.97, -0.97]
	2011	22	40.9% [-0.94, -0.82]
	2013	23	8.7% [-0.83, -0.80]
Searches for ‘Chemistry’	2005	8	0.0%
	2008	11	0.0%
	2011	23	4.3% [-0.82, -0.82]
	2013	24	0.0%
Searches for ‘Mathematics’	2005	9	0.0%
	2008	14	0.0%
	2011	23	4.3% [-0.85, -0.85]
	2013	25	0.0%
Searches for ‘Globo’	2005	15	0.0%
	2008	20	0.0%
	2011	26	0.0%
	2013	26	0.0%
Searches for ‘games’	2005	19	0.0%
	2008	24	0.0%
	2011	26	0.0%
	2013	26	0.0%

[a] These volume figures are relative and normalised, as discussed in Section 2.2.

REFERENCES

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: Wiley.

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models* (3rd ed.). Los Angeles: Sage.

Hope, A. C. A. (1968). A Simplified Monte Carlo Significance Test Procedure. *Journal of the Royal Statistical Society: Series B (Methodological)*, *30*(3), 582–598. https://doi.org/10.1111/j.2517-6161.1968.tb00759.x

Kalogirou, S. (2012). Testing local versions of correlation coefficients. *Jahrbuch Für Regionalwissenschaft*, *32*(1), 45–61. https://doi.org/10.1007/s10037-011-0061-y

Kalogirou, S. (2013). Testing geographically weighted multicollinearity diagnostics. In *Proceedings of GISRUK 2013, Liverpool, UK, 3-5 April 2013*. Liverpool: Department of Geography and Planning, School of Environmental Sciences, University of Liverpool. Retrieved from http://gisc.gr/?mdocs-file=1140

Kalogirou, S. (2015). A spatially varying relationship between the proportion of foreign citizens and income at local authorities in Greece. In *Proceedings of the 10th International Congress of the Hellenic Geographical Society, Thessaloniki 22-24 October 2014*.

Thessaloniki: Aristotle University of Thessaloniki. Retrieved from gisc.gr/?mdocs-file=1048

Kalogirou, S. (2019, April 23). Package “lctools.” Retrieved from https://cran.r-project.org/web/packages/lctools/index.html

Moran, P. A. P. (1948). The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society: Series B (Methodological)*, *10*(2), 243–251. https://doi.org/10.1111/j.2517-6161.1948.tb00012.x

R Core Team. (2019). The R Project for Statistical Computing. Retrieved March 11, 2019, from https://www.r-project.org/

Rey, S. J., & Smith, R. J. (2013). A spatial decomposition of the Gini coefficient. *Letters in Spatial and Resource Sciences*, *6*(2), 55–70. https://doi.org/10.1007/s12076-012-0086-z

Vogt, C., & Castelfranchi, Y. (2009). Interesse, informação e comunicação. In M. Albornoz, Á. M. Ullastres, & L. A. Uli (Eds.), *Cultura científica en Iberoamérica. Encuesta en grandes núcleos urbanos* (pp. 21–36). Madrid: FECYT, OEI, RICYT. Retrieved from http://icono.fecyt.es/informesypublicaciones/Documents/CulturaCientificaEnIberoamerica.pdf

Wickham, H., & Chang, W. (2019, August 11). Package “ggplot2.” Retrieved from https://cran.r-project.org/web/packages/ggplot2/index.html