Prioritization of Government Auditing by Ranking: A Case Study Using Health, Education and Public Security Data from Distrito Federal's Administrative Regions

1st Augusto Nunes
student discipline MD/PPCA
UnB
Brasília, Brazil
augcrn@gmail.com

2nd Ennio Ferreira
student discipline MD/PPCA
UnB
Brasília, Brazil
ennioferreirab@gmail.com

3rd Mayra Campos
student discipline MD/PPCA
UnB
Brasília, Brazil
mayra.campos@gmail.com

***Abstract*—In this study we present a novel indicator that could be used in the annual planning phase of internal government auditing done by the executive branch of Distrito Federal's executive government branch responsible for internal auditing, the Comptroller-General of Distrito Federal *Controladoria-Geral do Distrito Federal*. Using open data in four dimensions (vulnerabilites): health, education, sociodemographic and public security, by clustering the 33 Administrative Regions into 06 clusters using unsupervised K-Means and ranking a subset of the variables that an internal auditor chose, we were able to suggest an indicator that achieved high concordance - 90% in general and 70% *intra* cluster - when evaluated against the 2022 internal auditing reports.**

***Index Terms*—data mining; internal auditing; risk-based auditing; cluster analysis; ranking statistics**

I. Introduction

A. Context Description

The Comptroller-General of Distrito Federal (CGDF) has, among its institutional attributions, "to secure good and regular application of public resources, and to coordinate the internal control system of the executive government branch"[1]. These institutional assignments unpack not only in the most direct actions on fraud detection and the enforcement on governance controls among all executive branches to reduce its risk, but also on asserting that ideally all of the public sector planning, decisions and activities are mindful of the stakeholders interest in the public and private sectors.

In order to assert that these attributions are carried out, one of the actions that are periodically executed are plannings of audits to be done by the internal auditors teams of CGDF at the end of each first semester of the year. These planning phases the longer term auditing processes that will be carried out for the next 12 months, directing the auditing efforts toward those that have the *biggest potential* for generating societal return. Furthermore, the audits are also carried when there are facts of general commotion reported in the media or when any irregular acts come to the knowledge of the auditors, usually by complaints made within the ombudsman service.

There are many approaches for evaluating which audits could have this biggest potential impact, and for ranking them having in mind that internal audits can't reach every single government action, either because there's scarcity of resources or because the expected return of an expensive auditing process would be minimal. One of those uses the discipline of *Risk-based auditing*, as established in the Internal Audit Capability Model (IA-CM) for the Public Sector framework.

Risk is anything that could lead to the objectives not being achieved. In this context, we can understand *risk* as both the risk that the government endeavour which is under scrutiny of the internal auditors did not meet its objectives, and as the risk of the relative success of the auditing process itself. This work mainly focuses on the first interpretation, which is to suggest by looking at some variables associated with the Administrative Regions of the Federal District we will be able to propose an index that will be capable to capture which of the Administrative Regions have citizens in general at a higher risk, and thus which of those Administrative Regions should receive precedence in the internal audits.

Risks vary in nature, some are very direct, such as the materiality of the endeavour - i.e. the amount of money the government spent or is planning to spend to achieve a certain objective - but others are not directly measurable by merely inspecting official documents. The Comptroller-General of the Union, in its Manual of Technical Guidelines for Government Internal Audit Activities[2], uses two categories for risk assess-

[1]Translated from https://www.cg.df.gov.br/competencias-2/

[2]Available in Portuguese at https://wiki.cgu.gov.br/index.php/Manual_de_Orienta%C3%A7%C3%B5es_T%C3%A9cnicas_da_Atividade_de_Auditoria_Interna_Governamental_do_Poder_Executivo_Federal

978-1-7281-4387-3/19/$31.00 © 2019 IEEE

ment: quantitative and qualitative. It states that qualitative risks have a certain degree of subjectivity, and exemplify them by referencing the social impact that is associated with an audit object.

B. Problem

The Federal District has both Brazil's highest GDP *per capita* in one of its Administrative Regions, *Lago Sul*; and the country's biggest *favela* in another, *Sol Nascente*. Such discrepancies in the geographic and social space are to blame for many side effects of economic inequality, as stated by (PATTUSSI *et al.*, 2001), (AZZONI; SERVO, 2002) and (SACHSIDA *et al.*, 2010). Even though these works are not focused on the problem of internal auditing, this suggests the inequality of access to basic needs poses a strong risk to the success of government endeavours in areas were it is most rampant.

In this work, we use a set of variables that lie in four dimensions - thus called *vulnerabilites* - of public policy evaluation, namely health, education, sociodemographic and public security, all of these freely available according to the discipline of open data, and we propose an index based on the qualitative risk attributed in these four vulnerabilities that could be used in the planning phase of the annual internal audits carried by the institution for the general composition of the 33 Administrative Regions that the Federal District currently has, and also on 6 clusters which were identified using the unsupervised K-Means algorithm. The clustering analysis also provides some insight into the compositional similarities between those administrative regions, identifying those which are at higher risk.

II. Literature Review

Generally speaking, the Institute of Internal Auditors (IIA) is the *de facto* authority regarding auditing activities both in private and public sectors. Internal auditing in public agencies and more generally government activities is most often treated as a separate practice from the private internal auditing, given its nuanced list of priorities and obligations which are not fundamental in private settings.

The most reputable reference published by the IIA is the Internal Audit Capability Model (IA-CM) for the Public Sector: Overview and Application Guide (INSTITUTE OF INTERNAL AUDITORS, 2009). This publication proposes an assessment framework for the phases of an internal audit in the governmental perspective, using the idea of Key Performance Areas (KPA) as a series of stepping stones for the internal auditors working in the public sector. None of these KPA have an approach that considers the vulnerabilites here studied.

We investigated reputable references that deal with the problem of prioritization of public government auditing, and concluded that many of these deal primarily with the auditing processes limited to the financial (i.e fraud, waste or collusion) perspective of materiality as the majoring factor in assessing risk.

Balaniuk *et al.* (2012) considered the use of a Naïve Bayes classifier trained on data pertaining to the purchases effected by the Federal Government, using variables related both to the executive branch of the government that ordered the purchase and the private entity that participated in the process. They proposed the classifier as a semi-automated risk assessment framework capable of mitigating the inherent financial menaces associated with these activities. So, even though this work utilises a classic approach in Data Mining, it falls short to address anything remotely close to what we propose here, instead focusing on the official documentation produced in the formal processes of government purchasing and later accounting reviews as the predictors of whether a similar process might also be fraudulent.

One interesting source we consulted was the work of Nascimento (2022): in his monograph he proposed an approach similar to the one we have chosen here, but applied it for the Brazilian state of Sergipe and its municipalities. It also uses the novel approach of considering dimensions and variables for risk assessing in the planning phase other than the typical object measures that are more often employed in financial internal auditing. Even though this is a similar work, we could not merely translate this work to the context of the Federal District, given that it is not exactly the same as municipalities in other states of Brazil in terms of constitutional law. We tried to circumvent this limitation treating the Administrative Regions as *municipalities*, but this also has the shortcoming that they are not, as stated in the organic law.

III. Methodology

A. Framework

For this work we followed what is proposed in the Cross Industry Standard Process for Data Mining, CRISP-DM, in particular the first five of its six phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation

The first topic is situated here in the Introduction section of this paper, the following three are contained in this Methodology section, and the end we present our findings as the Evaluation. As the nature of this work is mainly speculative, we were unable to apply the sixth phase, Deployment.

B. Data Understanding

These variables were chosen in discussions with one of the internal auditors of CGDF, and each of them have the same weight in the final indicator according to the opinion of this expert.

Ideally, we would have 33 observations of the 8 variables described in I, one for each of the Administrative Regions of the Federal District, but due to the varying nature of the sources used and the fact that some of the Administrative Regions have only been established very recently, we have

Source	Variable	Description
District Domicile Survey	SE1	Literacy level of the voters (used as a proxy of the proportion of all the population that finished at least high school)
	SE2	Proportion of families living in extreme poverty (R$ < 200) that receive social benefits (*Bolsa Família, BPC/Loas* or scholarships)
Datasus	S1	Rate of fetal mortality (per thousand born alive)
	S2	Rate of mortality by avoidable causes (per 10,000 inhabitants)
Public Security Secretary	SEG1	Rate of occurrence of crimes against property (per 1,000 inhabitants)
District School Census	E1	Student Average per class in elementary school
	E2	Age-grade lag rate in elementary school
	E3	Failure rate in elementary school

TABLE I
SOURCE, CODE AND DESCRIPTION OF THE VARIABLES

Administrative Region	Attributes with missing data
Candangolândia	SEG1
Sudoeste/Octogonal	SEG1
Park Way	SEG1
Sobradinho II	E3
Jardim Botânico	E3
Itapoã	E3
SIA	S2, SE2
Vicente Pires	E3
Fercal	E1, E2, E3
Sol Nascente/Pôr do Sol	S1, S2, E1, E2, E3

TABLE II
ADMINISTRATIVE REGIONS AND THE ATTRIBUTES FOR WHICH THERE IS MISSING DATA

a few instances of missing data, which will be imputed as described in the next subsection.

C. Data Preparation

- Standardization of variables: the scale chosen for each of the variables changes according to its interpretation and variability. In order to standardize the behaviour and maintain the most information about their location and scale properties we chose to use z-score normalization.
- Interpretation of variable SE1: in order to obtain an indicator that has the interpretation of risk, i.e. higher = most at risk, we chose to multiply SE1 by (-1), with the understanding
- Data imputation: some of the smaller Administrative Regions or those created more recently have missing data for some of the attributes. To impute this data, we used Ridge Regression as implemented in the `IterativeImputer` class of `scikit-learn`: each of the variables with missing values is treated as a feature in a Ridge Regression, and this process is repeated iteratively at most 10 times.
- Collinearity: it is not uncommon when studying demographic problems that strong correlations arise in the data, which generally poses itself as a problem for when adjusting models (HILL; ADKINS, 2001). In particular, the variables SE1 and SE2 presented a Fisher correlation score of 0.77, but it was a decision of the consulted internal auditor to keep both of these variables in the proposed index.

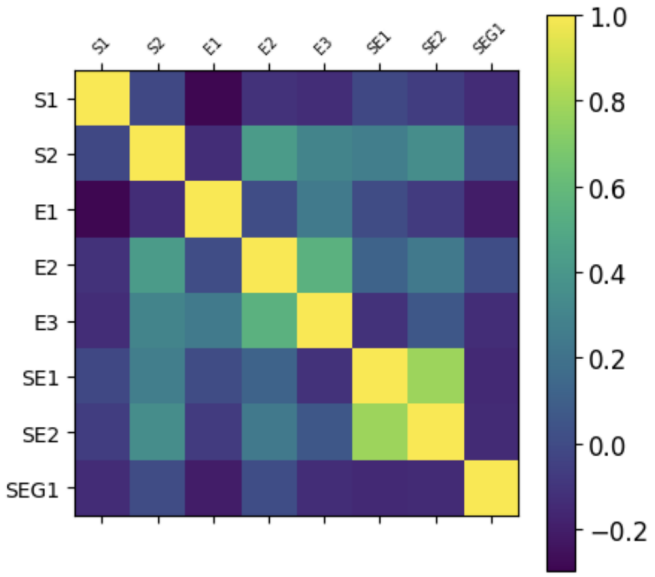

Fig. 1. Correlation matrix of the variables post normalization and imputation

- Definition of the Vulnerability Index: we chose the linear combination of the factors

$$\mathrm{X}_{\text{index.}} = S1+S2+E1+E2+E3+SE1+SE2+SEG1 \quad (1)$$

D. Modelling

- In order to segment the Administrative Regions by their vulnerability similarity, we used the K-Means clusterization technique, which tentatively separates the Administrative Regions which are most similar inside the same cluster, and put those least similar in different ones.
- One of the challenges we face when we use K-Means as an unsupervised clusterization technique is to determine the optimal number of clusters. An approach is to calculate the Elbow metric, which uses the concept of internal inertia of the clusters while it computes teh sum of the squared distances from each point to its assigned center, for a range of the tentative number of clusters we choose initially. The Elbow metric for our problem indicated that

6 is the optimal number of clusters, as illustrated in the Figure 2.

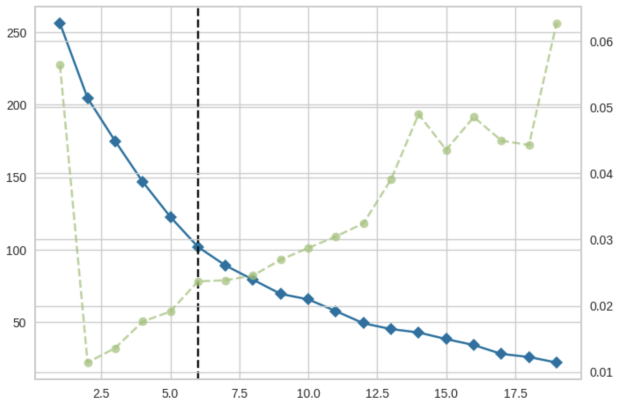

Fig. 2. Métrica Elbow. Quantidade de grupos formados

IV. Evaluation

Considering the general perspective of Administrative Regions according to the proposed vulnerability index V_index..

Administrative Region	Vulnerability Index
Fercal	8.712
SIA	6.228
Candangolândia	5.353
Sobradinho	5.353
SCIA/Estrutural	5.092
Planaltina	4.949
São Sebastião	4.393
Varjão	4.352
Ceilândia	3.055
Sol Nascente/Pôr do Sol	3.001
Brazlândia	2.663
Riacho Fundo II	2.607
Santa Maria	2.347
Itapoã	2.324
Park Way	1.681
Núcleo Bandeirante	1.542
Cruzeiro	1.281
Recanto das Emas	0.954
Paranoá	0.907
Taguatinga	0.422
Jardim Botânico	0.270
Lago Norte	-0.071
Samambaia	-0.165
Gama	-0.469
Águas Claras	-1.019
Vicente Pires	-1.173
Sobradinho II	-1.594
Lago Sul	-2.121
Plano Piloto	-2.524
Guará	-2.830
Riacho Fundo	-3.059
Sudoeste/Octogonal	-4.459

TABLE III
Administrative Regions and their respective Vulnerability Indices, sorted in decreasing order.

The proposed ranking makes sense when we look at the socio-economic conditions of some of those that ranked at the top: *Fercal* itself is an Administrative Region which has its roots in rural activites and owe its name to a cement industry, marred with inequality and pollution of air and water, is consistently facing problems regarding access to basic needs, such as infrastructure, job opportunities, and the scarcity of public endeavours in education and public health services (SANTOS, 2016).

SIA is an Administrative Region that originally catered exclusively to the industrial needs of the Federal District, such as the beverage industry and gasoline/diesel silos, but is now being gradually occupied by businesses and residencies. This occupation though is not being accompanied by the public services, particularly in regards to public security vulnerability, as evidenced by its value for the variable SEG1 (rate of occurrence of crimes against property) 5.42, much higher than for the other Administrative Regions.

At the bottom, we have some Administrative Regions such as *Sudoeste/Octogonal*, *Riacho Fundo* and *Guará*, which in general are better served with public infrastructure such as police departments, basic sanitation and schools.

Cluster	Administrative Region	Average Vulnerability Index (cluster)
6	SIA	6.228
5	Candangolândia Fercal Varjão	5.353
1	Brazlândia Ceilândia Itapoã Planaltina Riacho Fundo II SCIA/Estrutural Sobradinho Sol Nascente/Pôr do Sol São Sebastião	3.055
2	Cruzeiro Jardim Botânico Núcleo Bandeirante Paranoá Park Way Recanto das Emas Samambaia Santa Maria Taguatinga	0.954
3	Gama Guará Lago Norte Lago Sul Plano Piloto Sobradinho II Vicente Pires Águas Claras	-1.384
4	Riacho Fundo Sudoeste/Octogonal	-3.759

TABLE IV
Groups formed using K-Means and their respective average vulnerability indices

Using a mere socioeconomic view, the obtained clusters and their average indices provide a very coherent portrait of the disparities between Administrative Regions, and it was able to segment similar entities together according to their *de facto* similarities.

The clusters are plotted in a map of the Federal District in Figure 3, with each cluster being colored at the same tone.

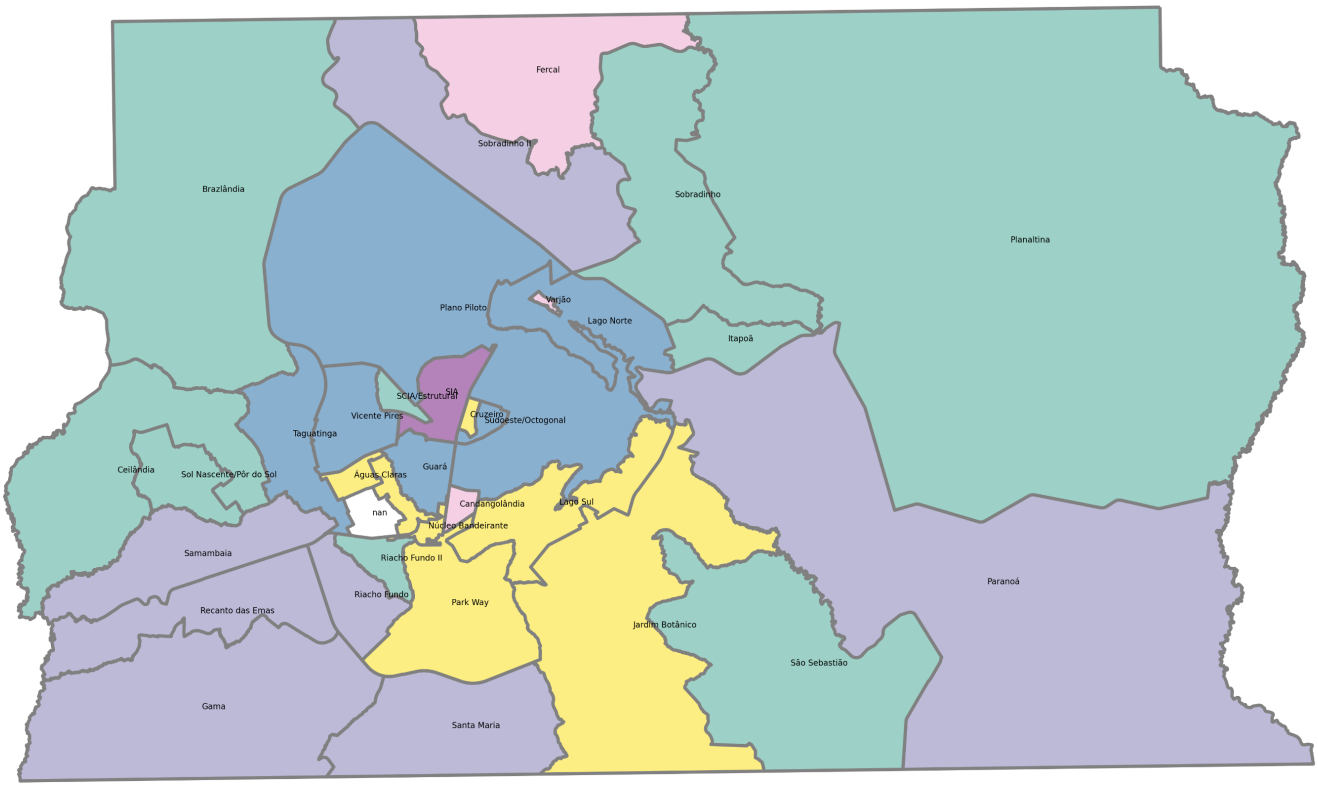

Fig. 3. Clusters Map

A. *Evaluation of the proposed index within the internal audits published in 2022*

As another tool to assess the proposed index, we compiled all the internal audits reports published in 2022. If the audit was related to one or more Administrative Regions, we noted its service order and respective region, as listed in the table IV-A.

When cross-evaluating the data presented in tables III and IV-A, we observed that the audit reports dealt with 5 of the top 10 Administrative Regions ranked according to our proposed index. When we evaluated this concordance regarding clusters, the first cluster - composed only of *SIA* - would be audited, the second biggest average risk (cluster 5) would have 1 out of 3 Administrative Regions audited, and the third cluster would have have 6 out of 9 Administrative Regions audited.

Service Order	Administrative Region
109	Fercal
108	Itapoã
104	Cruzeiro
101	SIA
100	Brazlândia
98	SCIA/Estrutural
96	Gama
82	Plano Piloto
82	Riacho Fundo II
79	Plano Piloto
74	Samambaia
74	Ceilândia
74	Lago Norte
74	Núcleo Bandeirante
69	Gama
69	Santa Maria
69	Plano Piloto
63	Jardim Botânico
56	Sobradinho
42	Paranoá
36	Guará
32	Plano Piloto
32	Guará
32	SIA
28	Plano Piloto
28	SIA
21	Santa Maria
19	Sobradinho II
17	Recanto das Emas
9	Lago Norte

TABLE V
INTERNAL AUDIT REPORTS THAT DEALT WITH AUDIT OBJECTS PERTAINING TO ONE OR MORE ADMINISTRATIVE REGIONS

V. Limitations and Conclusions

This study has its more direct limitation in the facts that Administrative Regions do not have the same statuses as municipalities in the Brazilian constitutional ordination: they do not have executive - in particular budgetary - independence to decide to what ends they will direct their efforts. An Administrative Region can't decide, for example, that it needs a new elementary school: this would have to be proposed by the Federal District's Secretary of Education. A similar conclusion applies to endeavours in all other vulnerabilities we considered here.

Still, we proposed an index that had a 50% concordance with what was, as a matter of fact, audited by the government internal auditors. Even though the reports used for our evaluation are not exclusively of internal audits that were planned during the planning phases, this suggests what an application of our vulnerability index would incur in a planning that was at least half the time in accordance to what is currently done.

At last, but not least, we would like to thank the internal auditors Renata Márcia Canuto Dumont and Leandro Shimabukuro for their collaboration with this work.

References

AZZONI, Carlos R; SERVO, Luciana MS. Education, cost of living and regional wage inequality in Brazil. **Papers in regional science**, Springer, v. 81, p. 157–175, 2002.

BALANIUK, Remis *et al.* Risk based government audit planning using naive bayes classifiers. *In:* ADVANCES in Knowledge-Based and Intelligent Information and Engineering Systems. [*S. l.: s. n.*], 2012.

CEVADA, Luana Zanini;
DAMY-BENEDETTI, Patricia de Carvalho. USO DA MATRIZ DE PRIORIZAÇÃO (MATRIZ GUT) COMO ALIADA EM AUDITORIAS. pt. **Revista Científica Unilago**, v. 1, n. 1, 2021. Number: 1. ISSN 2318-244X. Available from: https://revistas.unilago.edu.br/index.php/revista-cientifica/article/view/591. Visited on: 9 May 2023.

CONSTITUIÇÃO da República Federativa do Brasil. [*S. l.: s. n.*], 1988. Available from: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Visited on: 9 May 2023.

CONTROLADORIA-GERAL DO DISTRITO FEDERAL. **Competências**. pt-BR. [*S. l.: s. n.*]. Available from: https://www.cg.df.gov.br/competencias-2/. Visited on: 9 May 2023.

CONTROLADORIA-GERAL DO DISTRITO FEDERAL. **Portal da Transparência do Distrito Federal**. [*S. l.: s. n.*]. Available from: https://www.transparencia.df.gov.br/#/. Visited on: 9 May 2023.

CRISP-DM Help Overview - IBM Documentation. [*S. l.: s. n.*]. Available from: https://www.ibm.com/docs/en/spss-modeler/18.4.0?topic=dm-crisp-help-overview. Visited on: 9 May 2023.

DE ANDRADE MOREIRA, WELLINGTON. ESTRATÉGIAS DE PRIORIZAÇÃO DE AUDITORIAS NA FOLHA DE PAGAMENTO DO GOVERNO DO DISTRITO FEDERAL–GDF. *In:* available from: http: //consad.org.br/wp-content/uploads/2016/06/Painel-01-01.pdf.

HILL, R Carter; ADKINS, Lee C. Collinearity. **A companion to theoretical econometrics**, Blackwell Oxford, p. 257–78, 2001.

INSTITUTE OF INTERNAL AUDITORS, Research Foundation Staff. **Internal Audit Capability Model (IA-CM) for the Public Sector: Overview and Application Guide**. [*S. l.*]: Institute of Internal Auditors, Research Foundation, 2009. ISBN 978-0-89413-675-7. Available from: https://books.google.com.br/books?id=GuEISQAACAAJ.

LIMA, Dagomar Henriques. Seletividade do Controle Externo em Auditoria Operacional*. pt. **Revista do TCU**, n. 115, p. 26–35, May 2009. ISSN 2594-6501. Available from: https://revista.tcu.gov.br/ojs/index.php/RTCU/article/view/384. Visited on: 9 May 2023.

NASCIMENTO, José Leonardo Ribeiro. Índice de priorização de objetos de auditoria: um estudo de caso para municípios sergipanos. Idioma::Português:portuguese:pt, Sept. 2022. Accepted: 2022-12-01T16:45:29Z Publisher: Escola Nacional de Administração Pública (Enap). Available from: http://repositorio.enap.gov.br/jspui/handle/1/7396. Visited on: 9 May 2023.

PATTUSSI, Marcos Pascoal *et al.* Social deprivation, income inequality, social cohesion and dental caries in Brazilian school children. **Social science & medicine**, Elsevier, v. 53, n. 7, p. 915–925, 2001.

SACHSIDA, Adolfo *et al.* Inequality and criminality revisited: further evidence from Brazil. **Empirical Economics**, Springer, v. 39, p. 93–109, 2010.

SANTOS, Kelly Alves dos. Percepções socioambientais: um estudo da topofilia na comunidade Queima Lençol–Fercal/DF, 2016.

APPENDIX

Link para o vídeo no YouTube
Link para o GitHub

TABLE VI: Consolidated Data

RA	S1	S2	E1	E2	E3	SE1	SE2	SEG1	City
1.0	-0.841	-0.131	-1.511	0.176	-0.197	-1.357	-1.200	0.537	Plano Piloto
2.0	-0.670	-0.489	-0.042	-0.526	-0.669	0.661	-0.634	-0.101	Gama
3.0	-0.856	-0.004	-0.350	0.626	0.189	0.019	-0.414	0.211	Taguatinga
4.0	-0.408	2.069	0.176	-0.053	0.738	0.100	0.308	-0.267	Brazlândia
5.0	-0.543	2.424	-0.026	1.079	2.276	-0.357	0.753	-0.253	Sobradinho
6.0	-0.872	0.296	1.035	1.709	1.418	1.129	0.478	-0.244	Planaltina
7.0	-0.472	-0.171	0.053	-0.587	-0.831	0.996	0.794	0.126	Paranoá
8.0	1.009	-0.261	0.017	0.023	0.065	-0.016	-0.383	0.088	Núcleo Bandeirante
9.0	-0.964	0.571	0.649	0.855	1.039	0.048	0.714	0.142	Ceilândia
10.0	-0.704	-0.480	-1.467	-0.220	-0.112	-0.811	-0.813	-0.222	Guará
11.0	1.048	0.792	0.518	-0.309	-0.035	-0.726	-0.651	-0.354	Cruzeiro
12.0	-0.881	-0.331	0.270	-0.587	-0.522	0.284	0.570	0.032	Samambaia
13.0	-0.691	-0.087	1.724	0.751	-0.669	0.550	-0.182	-0.048	Santa Maria
14.0	-0.720	0.929	0.191	0.628	0.939	1.077	1.612	-0.262	São Sebastião
15.0	-0.683	-0.047	1.165	-1.623	-0.499	1.241	0.482	-0.082	Recanto das Emas
16.0	1.560	-1.214	1.037	-1.185	-1.117	-1.523	-1.287	-0.392	Lago Sul
17.0	-0.174	1.082	0.373	2.186	-0.252	-0.368	0.012	-0.253	Riacho Fundo II
18.0	0.456	0.142	-1.206	-0.142	1.349	-1.071	-1.251	-0.349	Lago Norte
19.0	1.889	0.971	-1.415	-0.068	-0.468	0.299	0.440	-0.295	Candangolândia
20.0	-0.772	-0.149	1.855	-0.897	-0.082	-1.511	-1.195	-0.268	Águas Claras
21.0	-0.293	-1.658	-0.171	-2.740	-1.542	0.193	0.459	-0.307	Riacho Fundo
22.0	0.071	-1.728	-0.732	-1.195	-0.855	-1.469	-1.180	-0.373	Sudoeste/Octogonal
23.0	2.520	-0.062	-2.037	-0.890	-1.527	1.270	1.418	-0.340	Varjão
24.0	1.858	-1.231	-0.348	1.450	1.727	-1.486	-0.858	-0.432	Park Way
25.0	-0.262	0.143	-0.950	1.748	-0.082	1.856	2.558	0.080	SCIA/Estrutural
26.0	-0.534	-1.270	0.054	-0.042	-1.628	0.463	-0.280	-0.358	Sobradinho II
27.0	-0.124	-1.914	2.285	-0.137	1.658	-1.145	-0.874	-0.480	Jardim Botânico
28.0	-0.489	-0.752	0.687	0.059	1.302	1.077	0.451	-0.012	Itapoã
29.0	-0.191	-0.052	-0.991	-0.145	-0.715	-1.084	-1.020	5.427	SIA
30.0	-0.385	0.965	-0.821	-0.151	-0.831	-0.590	-1.037	-0.322	Vicente Pires
31.0	2.309	1.141	0.016	-0.079	-0.124	1.785	0.054	-0.388	Fercal
32.0	-0.193	0.508	-0.037	0.287	0.058	0.465	2.154	-0.239	Sol Nascente/Pôr do Sol

TABLE VI: Consolidated Data

Note: The values of the SE1 variable were multiplied by -1, as they have a 'the higher, the better' relationship, correctly impacting the vulnerability index.