International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 41 No. 4 Feb. 2024, pp. 1088-1097

http://www.ijias.issr-journals.org/

Application of Correspondence Factor Analysis (CA) on data on the nutritional status of children under 5 years old, PRONANUT, from January to December 2022 in DR Congo

Don Folly FOFOLO Mulembu[1], Rostin MABELA Makengo Matendo[2], Fidèle MUAKU Mvunzi[3], Grace NKWESE Mazoni[3], and Camile LIKOTELO BINENE[3]

[1]Dpt. Computer Science & Mathematics, Technical Section, ISP Kikwit, RD Congo

[2]Dpt. Mathematics, Statistics and Computer Science, Faculty of Science and Technology, University of Kinshasa, RD Congo

[3]Dpt. Mathematics, Statistics and Computer Science, Faculty of Science and Technology, National Pedagogical University of Kinshasa Ngaliema, RD Congo

ABSTRACT: The objective pursued in this article is to study Correspondence Factor Analysis (CA), which is an extremely powerful tool for synthesizing information, widely used when dealing with a large mass of qualitative data. treat.
It also makes it possible to identify existing relationships between individuals by evaluating their similarities, as well as relationships between variables by evaluating their connections, and obtain a simple representation of the data cloud in a low-dimensional space closer to reality.
Factorial analysis of the data was applied using R software, version 4.3.1 (2023-06-16 ucrt).
The application of the said method was carried out on «the nutritional status of Congolese children under 5 years old, from January to December 2022, PRONANUT, DR Congo».
The data was summarized in a table of 26 row categories and 12 column categories. The 26 rows (individuals) represent the health provinces of DR Congo and the 12 columns, the variables relating to the activities of the PRONANUT Preschool Consultation.

KEYWORDS: Data analysis, Factorial correspondence analysis, Nutritional status of Congolese children under 5 years old.

1 INTRODUCTION

That today's world is experiencing a "data revolution": this is an undeniable fact. Because, with the dematerialization of an increasing number of processes and the appearance of completely new digital products and services, the quantity of digital data created increases exponentially (N. Shadboldt, H. Verdier, 2015-2016).

Also, in recent years, the explosion in the volume of data has been at the heart of the strategic priorities of governments, institutions, businesses and organizations. To the extent that this data is collected, analyzed and exploited, it offers unprecedented possibilities in terms of economic, medical, educational, managerial, financial, banking, advertising, commercial, electoral innovation, etc. This is why the development of new technologies such as Artificial Intelligence and Machine Learning only reinforce this phenomenon.

In addition, data analysis is a set of descriptive techniques, the major mathematical tool of which is matrix algebra, which is expressed without a priori assuming a probabilistic model (F. Z. Kasmi, 2019).

It includes, among other things :

- Principal component analysis (PCA), used for quantitative data, and its derived methods;
- Factorial correspondence analysis (CA) used on qualitative data (association table);
- Factorial analysis of multiple correspondences (AFCM or ACM) generalizing the previous one;

- Canonical analysis and generalized canonical analysis, which are more theoretical frameworks than easily applicable methods, extend several of these methods and go beyond description techniques (L. Lebart, M. Piron, A. Morineau, 2006).

Although these techniques described above are available in the standard extensions of R, it is often preferable to use two other more complete extensions, ade4 and FactoMineR, each having its advantages and different possibilities (J. Larmarange, et al. (2022).

Data analysis is also a set of techniques for structuring, possibly complicated, a multi-dimensional array of numbers and translating it into a structure at best. This structure can most often be represented graphically" (J-P. Fénelon, 1981).

That said, data analysis is a subfield of statistics that concerns itself with the description of joint data. We seek, through these methods, to provide the links that may exist between the different data and to derive statistical information which makes it possible to describe, in a more succinct way, the main information contained in these data. Hence our question :

1. What is the relevant information contained in the data table that must be extracted, summarizing it in order to make graphical representations that are both faithful to the initial data and easy to interpret?
2. Taking into account the similarities between individuals and the connections between variables, is it possible to summarize all the data by a restricted number of values without significant loss of information?
3. Is there a relationship between all the variables on the nutritional status of Congolese children under 5 years old?
4. Can we seek to reduce the number of variables describing the data, the quantity of information reduced, to the best maintained?

As we can see, all these questions ultimately aim to :

- Respond to the problems posed by statistical tables resulting from the study with large dimensions, that is to say the tables contain a high number of individuals, in rows (n) and several variables in columns (p) ;
- Summarize the information contained in a large table in the form of a matrix ;
- Organize and visualize information ;
- Identify useful information.

This is why this research studies the variability of factors in the nutritional status of Congolese children under 5 years old followed in preschool consultation structures (CPS). After the presentation of data relating to the nutritional status of children under 5 years old, coming from the National Nutrition Program (PRONANUT), we will carry out the Correspondence Factor Analysis (CA) or the model, before proceeding to analysis of the related results.

2 DATA PRESENTATION

In this study, the data come from the National Nutrition Program (PRONANUT) database. In fact, these are secondary data from the health provinces of the Democratic Republic of Congo detailing the nutritional status of Congolese children under 5 years old, i.e. the child's first 1000 days, at the national level.

After pre-processing the received data, our data frame includes 26 rows and 12 columns. And, the modalities rows (I) are the 26 health provinces of the DRC ; without forgetting that the column modalities (J) are the 12 variables on the nutritional status of Congolese children under 5 years old. All these data were obtained after pre-processing.

Furthermore, the target population is made up of Congolese children under 5 years old who participated in Pre-School Consultation (CPS) activities from January to December 2022.

Table 1. Data.frame of row (26) and column (12) modalities

PS	1	2	3	4	5	6	7	8	9	10	11	12
	EAEM6	**EAC611**	**EAC1223**	**AnjeM6**	**Anje611**	**EDEP1223**	**EDEP2459**	**CCPAM6**	**CCPA611**	**CCPA1223**	**CCPA2459**	**SVTA611**
1	5516	13344	7477	2246	2840	4628	6865	165	199	269	309	3849
2	10678	10438	5021	5591	4664	2350	2354	2323	1710	1159	802	6532
3	42375	30580	14418	30279	22052	17497	17412	2812	3100	2685	2299	23232
4	14366	20309	19846	15626	15660	16482	21917	4088	5549	6936	10846	14330
5	10167	11677	8032	4385	5052	6792	11247	388	548	535	557	5851
6	35403	34003	24511	25151	22445	16834	19846	933	1755	2020	2289	18569
7	29650	28096	18168	29238	24092	3049	3693	151	643	708	512	6766
8	31400	31177	32777	32365	28836	15253	25207	693	3729	6729	10334	16462
9	12257	33171	30131	13030	18326	16774	20431	374	1549	2617	2908	11052
10	20913	44358	60299	16815	22468	25646	29006	696	3645	6438	6342	18336
11	59630	39949	9908	42072	25511	7504	5903	1006	987	366	160	26072
12	28122	45841	63186	31981	40242	16500	19028	786	5715	8855	9524	21922
13	14927	21798	27454	14397	18127	14672	22353	1441	3843	5525	6849	13653
14	13135	13621	10263	12197	8642	8616	9941	588	584	783	699	9303
15	15852	28238	37216	17997	22753	16418	25458	219	1863	3829	4756	14489
16	7922	12745	12432	5946	6798	7752	9010	199	704	1134	1268	7593
17	9044	9202	5821	1274	1466	4235	4155	423	444	401	364	6318
18	12527	12820	8972	9384	8107	6726	7654	347	964	1265	1142	7098
19	87210	71287	70906	69511	54897	41684	55893	3194	8365	6058	7397	33681
20	12020	17048	18650	7649	22067	7173	8033	250	727	1060	752	7278
21	39567	67682	37362	45841	49677	28412	33615	1463	4586	5582	5554	28702
22	9294	14884	18466	8001	12133	6118	8731	871	2317	4466	4751	7414
23	23015	32403	34064	13196	20113	13106	14812	158	832	1666	1515	20009
24	14330	19086	18786	12154	15652	9955	11373	518	4960	6811	6792	10732
25	19345	20735	20453	13328	15011	13007	19063	327	574	797	930	10979
26	5884	10461	17723	4878	6838	13351	16894	1347	1686	2837	3751	7687

Source : Designed by us, following PRONANUT data, from January to December 2022.

3 MODEL

3.1 CHOICE OF MODEL

Correspondence factor analysis (CA) is used on qualitative data (association table). It is a statistical method of data analysis, which makes it possible to analyze and prioritize the information contained in a rectangular table of data and which, today, is particularly used to study the link between two qualitative (or categorical) variables).

The present study includes 26 individuals, i.e. the health provinces of DR Congo and 12 SPC activities as qualitative variables. The data table is of the *individuals*qualitative variables type,* based on factorial correspondence analysis (CA).

3.2 MODEL OVERVIEW

3.2.1 FACTOR CORRESPONDENCE ANALYSIS (CA)

It deals with the case of two qualitative or categorical variables.

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two qualitative or categorical variables with $\boldsymbol{p}$ and $\boldsymbol{q}$ categories (modalities), respectively, describing a set of $\boldsymbol{n}$ individuals (Z. Sayl, 2019-2020). Factor Analysis is based on the cloud of points, called a contingency table, denoted by **N^***. This is the matrix of observed numbers of $\boldsymbol{p}$ rows and $\boldsymbol{q}$ columns.

By crossing the two variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ we obtain :

$$\boldsymbol{N^* = \begin{pmatrix} x_{11} & \dots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{p1} & \dots & x_{pq} \end{pmatrix} \in \mathcal{M}(p * q)},$$

Or :

$\boldsymbol{x_{ij}}$ **:** this is the observed number, an element obtained by the intersection of **row** $\boldsymbol{i}$ and **column** $\boldsymbol{j}$**.**

DEFINITION 2.2.1.

We say that $\boldsymbol{N}$ is the contingency table crossing the qualitative variables $\boldsymbol{X}$ and $\boldsymbol{Y}$**,** if $\boldsymbol{N}$ is of the form **:**

$$N = \begin{bmatrix} x_{11} & \dots & \dots & \dots & x_{1q} \\ \vdots & \ddots & & & \vdots \\ \vdots & \dots & \boldsymbol{x_{ij}} & \dots & \vdots \\ \vdots & & & \ddots & \vdots \\ x_{p1} & \dots & \dots & \dots & x_{pq} \end{bmatrix}$$

Where $\boldsymbol{x_{ij}}$ is the number of observations simultaneously presenting the modalities $\boldsymbol{i}$ of $\boldsymbol{X}$ and $\boldsymbol{j}$ of $\boldsymbol{Y}$.

In our study, $\boldsymbol{x_{ij}}$ is the number of Congolese children under 5 years old belonging to health province $\boldsymbol{i}$ who participated in CPS activity $\boldsymbol{j}$.

DEFINITION 2.2.2.

The weight of row $\boldsymbol{i}$ is defined by :

$$p_i = \frac{x_{i\bullet}}{x\bullet\bullet}$$

The weight of column $\boldsymbol{j}$ is defined :

$$q_j = \frac{x_{\bullet j}}{x\bullet\bullet}$$

Or :

$$x_{\bullet j} = \sum_{i=1}^{n} x_{ij} \quad , \quad x_{i\bullet} = \sum_{j=1}^{p} x_{ij}$$

And $\boldsymbol{x} = \sum_{j=1}^{p} x_{\bullet j} = \sum_{i=1}^{n} x_{i\bullet} = \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}$

DEFINITION 2.2.3.

The profile of row $\boldsymbol{i}$ is defined by :

$$x_i = \left\{\frac{x_{ij}}{x_{i\bullet}}\right\} \; j = 1, \dots, p$$

The profile of column $\boldsymbol{j}$ is defined by :

$$y_i = \left\{\frac{x_{ij}}{x_{\bullet j}}\right\} \; i = 1, \dots, n$$

We pose :

$$D_I = diag(p_1, \dots, p_n)$$

$$D_J = diag(q_1, \dots, q_p)$$

$$X_{n*p} = \frac{1}{k} D_I^{-1} X$$

And

$$Y_{p*n} = \frac{1}{k} D_J^{-1t} X$$

DEFINITION 2.2.4.

We define two point clouds :

$$N(I) = \left\{(x_i, p_i), x_i \in \mathbb{R}^p, p_i > 0, \sum p_i = 1\right\}$$

$$N(J) = \left\{(y_i, q_i), y_j \in \mathbb{R}^n, q_j > 0, \sum q_i = 1\right\}$$

DEFINITION 2.2.5.

The distance between rows $\boldsymbol{x_i}$ and $\boldsymbol{x_i'}$ is given by :

$$\boldsymbol{d^2(x_i, x_i') = {}^t (x_i - x_i') A_I (x_i - x_i')}$$

Or $\boldsymbol{A_I = D_J^{-1}}$

The distance between the column profiles $\boldsymbol{y_i}$ and $\boldsymbol{y_i'}$ is given by :

$$\boldsymbol{d^2(y_i, y_i') = {}^t (y_i - y_i') A_I (y_i - y_i')}$$

Or $\boldsymbol{A_J = D_I^{-1}}$

NOTICED

This distance is called the metric of χ^2.

Indeed, by replacing $\boldsymbol{x_i}$ and $\boldsymbol{x_i'}$ respectively by $\left\{\frac{x_{ij}}{x_{i\bullet}}\right\}$ $j = 1, \dots, p$ and

$\left\{\frac{x_{i'j}}{x_{i'\bullet}}\right\}$ $j = 1, \dots, p$ in definition 2.2.5, we obtain :

$$\boldsymbol{d^2(x_i, x_{i'}) = {}^t (x_i - x_{i'}) A_I (x_i - x_{i'}) = \sum_{j=1}^{p} \frac{x}{x_{\bullet j}} \left(\frac{x_{ij}}{x_{i\bullet}} - \frac{x_{i'j}}{x_{i'\bullet}}\right)^2}$$

We do the same thing for the column profiles $\boldsymbol{y_i}$ and $\boldsymbol{y_i'}$ and we obtain :

$$\boldsymbol{d^2(y_j, y_{j'}) = {}^t (y_i - y_{j'}) A_J (y_j - y_{j'}) = \sum_{j=1}^{p} \frac{x}{x_{i\bullet}} \left(\frac{x_{ij}}{x_{\bullet j}} - \frac{x_{ij'}}{x_{\bullet j'}}\right)^2}$$

Simple correspondence analysis consists of carrying out a principal component analysis of the cloud of points $N(I)$, by setting : $V_I = {}^t X D_I X$ the inertia matrix relative to the origin, and AI the weight matrix.

4 RESULTS ANALYSIS

4.1 DATA PROCESSING

The computer processing was carried out using the R software, Version 4.3.1. from 06-16-2023. We used the packages :

- **FactoMineR** with the function: **CA** which performs factorial analysis of correspondences;
- **Factoextra** using the functions: **fvz_screeplot** to visualize the eigenvalues; **fviz_ca_row()** is used to visualize the columns and **fviz_ca_col()** is used to produce the column graph.

4.2 DATA ANALYSIS AND INTERPRETATION OF RESULTS

REPRESENTATION OF INDIVIDUALS ON THE MAIN FOREGROUND

Looking at graph 3.1., shown below, it is necessary to show the relationships between the rows points.

First, we see that rows with a similar profile are grouped together as follows :

- 1 (Bas Uele), 23 (Sud Ubangi), 20 (North Ubangi), 25 (Tshopo), 5 (Haut Uele), 16 (Maï-Ndombe), 15 (Lomami), 9 (Kwango) and 10 (Kwilu) ;
- 17 (Mangala), 14 (Lualaba), 6 (Ituri), 18 (Maniema), 19 (North Kivu) and 21 (South Kivu) ;
- 13 (Kasaï), 22 (Sankuru) and 24 (Tanganyika).

Then, the negatively correlated rows are positioned on opposite sides of the graph origin (opposite quadrants).

Finally, the distance between the rows points and the origin measures the quality of the rows points on the graph. Rows points that are far from the origin are well represented on the graph. We have :

- 4 (Haut Lomami) and 26 (Tshuapa) ;
- 2 (Ecuador), 3 (Haut Katanga) and 11 (Kinshasa) ;
- 1 (Bas Uele), 10 (Kwilu) and 23 (South Ubangi) ;
- 7 (Kongo Central).

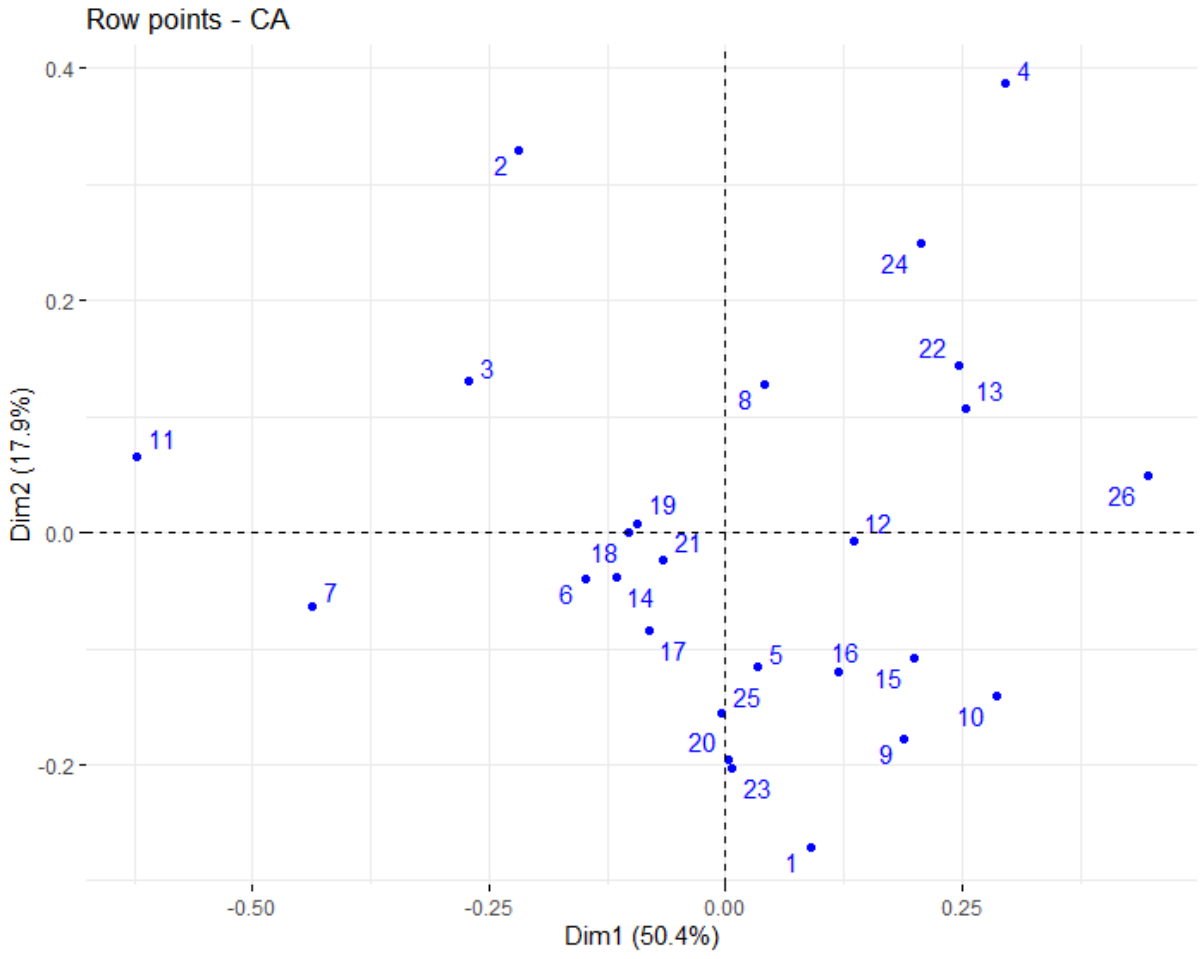

Fig. 1. Mapping of rows points

REPRESENTATION OF VARIABLES ON THE FIRST MAIN PLANE

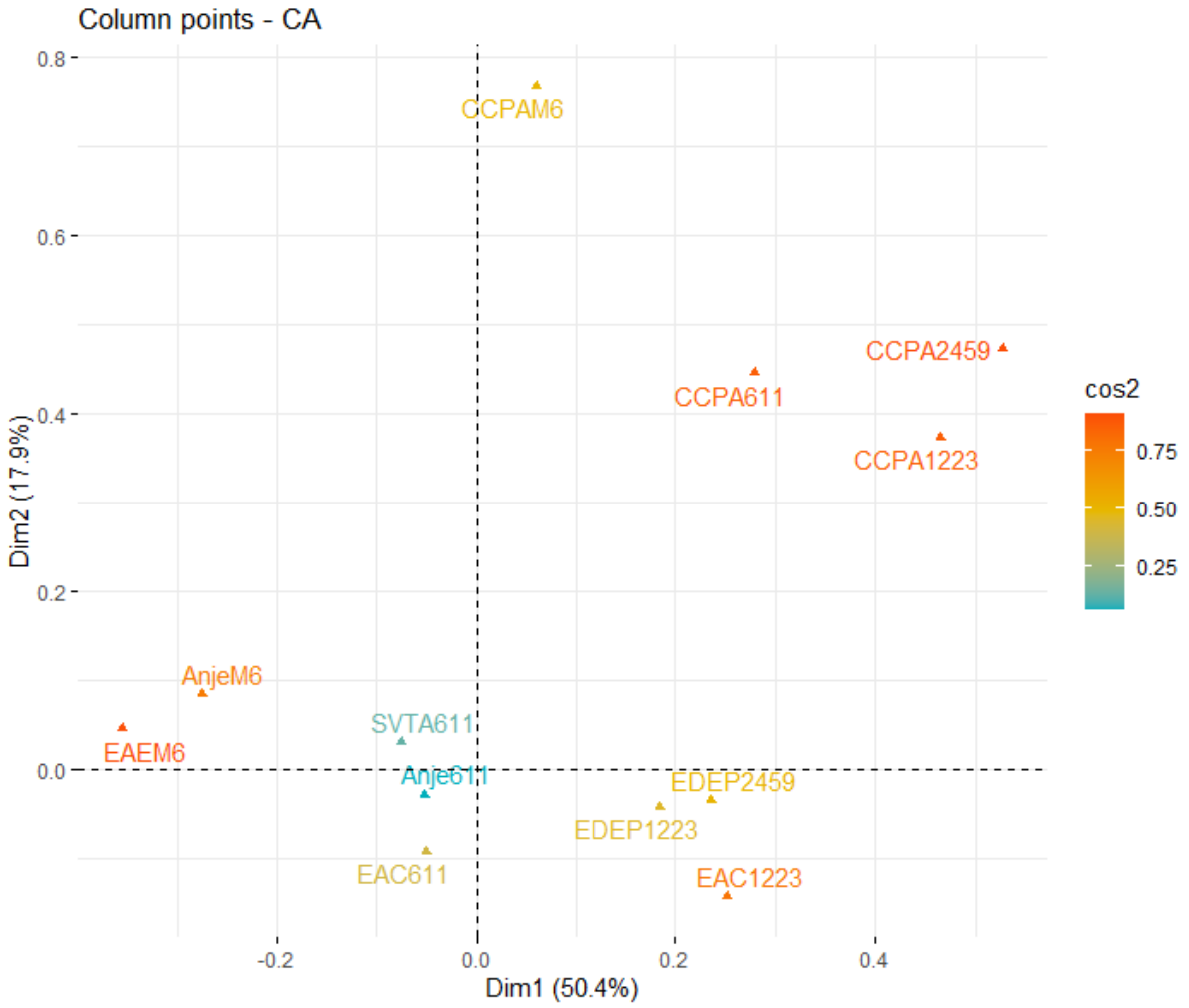

Fig. 2. Mapping of columns according to their cos2

REPRESENTATION OF INDIVIDUALS AND VARIABLES ON THE FIRST MAIN PLANE

According to Figure 3, below, rows are represented by blue dots and columns by red triangles.

The distance between rows points or between columns points gives a measure of their similarity (or dissimilarity). ***Rows points with a similar profile are close together on the graph. The same goes for column points.***

Through Figure 3, we see that :

- Row 4 (Haut Lomami) is most associated with columns CCPA611, CCPA1223 and CCPA2459 ;
- Rows 2 (Equateur) and 3 (Haut Katanga) are most associated with column AnjeM6 ;
- Rows 6 (Ituri), 14 (Lualaba) and 21 (South Kivu) are most associated with column Anje611 ;
- Rows 8 (Kasaï Oriental), 19 (North Kivu) and 18 (Maniema) are most associated with column SVTA611 ;
- Rows 1 (Bas Uele), 9 (Kwango), 10 (Kwilu) and 15 (Lomami) are most associated with column EAC1223 ;
- Rows 5 (Haut Uele), 17 (Mongala), 20 (North Ubangi), 23 (South Ubangi), 25 (Tshopo) are most associated with column EAC611 ;
- Rows 13 (Kasaï), 22 (Sankuru) and 26 (Tshuapa) are most associated with column EDEP2459 ;
- Row 24 (Tanganyika) is most associated with column CCPA1223 ;
- Rows 7 (Kongo Central) and 11 (Kinshasa) are most associated with column EAEM6.

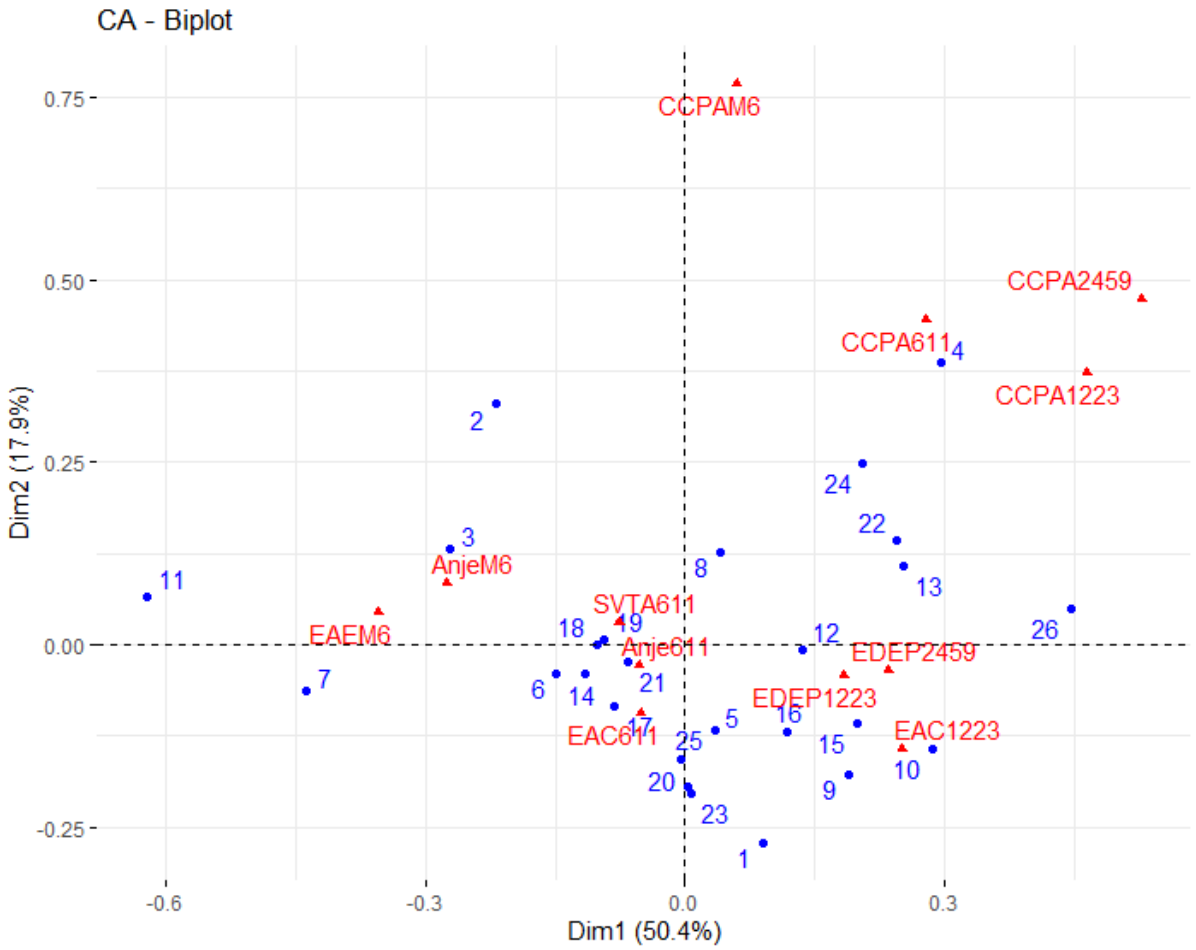

Fig. 3. Biplot of row and column modalities

SCREE-PLOT REPRESENTATION OF PERCENTAGE EIGENVALUES

Eigenvalues can be used to determine the number of axes to retain. An analysis is good when the first dimensions account for a large part of the variability. The first two axes explain **68.3%** of the total variance. This is an acceptable percentage. To determine the number of dimensions is to look at the graph of eigenvalues **(scree-plot)**, ordered from largest to smallest value. The number of axes is determined by the point, beyond which the remaining eigenvalues are all relatively small and of comparable sizes.

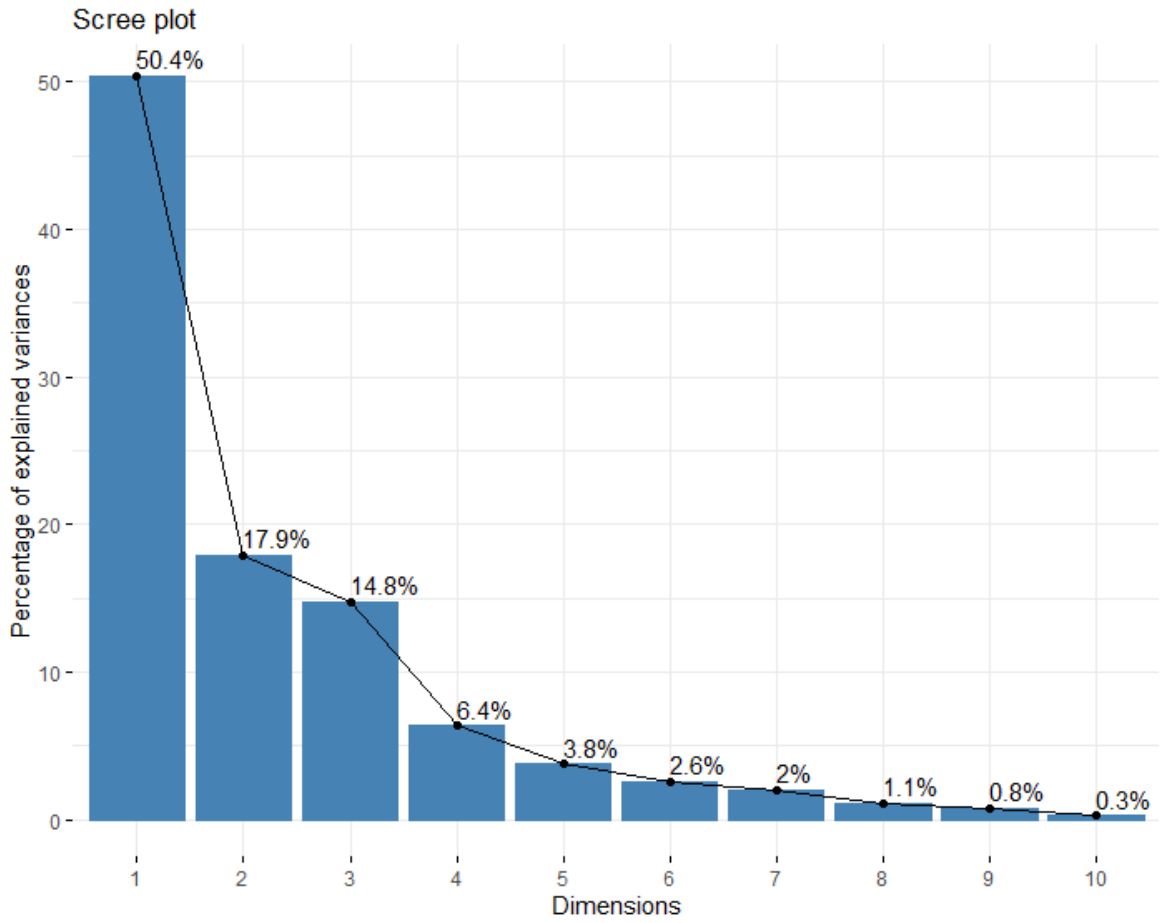

Fig. 4. Scree plot of eigenvalues in percentage

Dimensions 1 and 2 explain approximately **50.4%** and **17.9%** of the total inertia, respectively. This corresponds to a cumulative total of **68.3%** of the total inertia retained by the 2 dimensions. The higher the retention, the more subtlety contained in the original data is retained in the low-dimensional CA solution (M. Bendixen, 2003).

5 CONCLUSION

Our research focused on the Application of Correspondence Factor Analysis (CA) on data on the nutritional status of children under 5 years old, from January to December 2022, in the Democratic Republic of Congo.

The related data was reduced by obtaining a synthetic table or averaged table. The sample size ($\boldsymbol{n}$) is on average **4281453** Congolese children received in CPS during the months of January 2022 to December 2022, across the 26 health provinces of DR Congo, with 12 different CPS activities.

Also, the data were analyzed, visualized and interpreted, so that the CA results were obtained following the use of the different R. software packages, version 4.2.3. (2023-03-15 ucrt) in the following manner :

- Dimensions 1 and 2 explain respectively **50.4%** and **17.9%** of the total inertia. This corresponds to a cumulative total of **68.3%** of the total inertia retained by the 2 dimensions. The higher the retention, the more subtlety contained in the original data is retained in the low-dimensional CA solution (M. Bendixen, 2003).
- According to the chi-square test, the variables (modalities) of the rows and columns are statistically significantly associated.
- The p-value is less than **α=0.05**, which shows that there is a significant connection between the variables. Certainly, there is a dependency between the row and column modalities.
- The intensity of this association or this dependence is relatively weak in this specific case, because Cramer's **V** is **0.1.**
- The Contingency Coefficient (CC) is also low between the modalities. Or 0.316. Or approximately **31.6%.**
- The total inertia is **11.104.** All these indicators prove the weak connection between the row and column modalities.

REFERENCES

[1] Bendixen, M. (2003). "A Practical Guide to the Use of Correspondence Analysis in Marketing Research." Marketing Bulletin 14.
[2] Escofier, B., Pagès, J. (2008). Single and multiple factor analyses. Objectives, methods and interpretation, 4th Ed., Paris, Dunod, 328 p.
[3] Fénelon, J-P. (nineteen eighty one). What is data analysis ? Lefonen, Paris.
[4] Husson, F. (2018). R for statistics and data science, Rennes, PUR, 418 p.
[5] Kasmi, F.Z. (2019). Multivariate analysis and applications. Master's thesis, Abou-Bekr Belkaid University. Tlemcen.
[6] Larmarange, J. et Al. (2022). Analyze. Introduction to survey analysis with R and RStudio, Paris, 1403 p.
[7] Lebart, L., Piron, M., Morineau, A. (2006). Multidimensional Exploratory Statistics, Paris, Dunod, 464 p. (ISBN 978-2-10-049616-7).
[8] Sayl, Z. (2019-2020). Data analysis course. Level S6, Option : Economy, Ait Melloul, Morocco, 92 p.
[9] Shadboldt, N., Verdier, H. (2015-2016). Report of the Franco-British working group on the data economy. Franco-British TaskForce on innovation through data, Paris, 28 p.