

## P11

# Exploring the clinical characteristics of chronic obstructive pulmonary disease using multiple and joint correspondence analysis

Daniel Castro<sup>1</sup>, Vera Enes<sup>2,3,4</sup>, Gabriela Moura<sup>3,4</sup>, Alda Marques<sup>3,5</sup>

<sup>1</sup> Department of Mathematics, University of Aveiro, Aveiro, Portugal.

<sup>2</sup> Centre for Research & Development in Mathematics and Applications (CIDMA), University of Aveiro.

<sup>3</sup> iBiMED – Institute of Biomedicine (iBiMED), University of Aveiro, Aveiro, Portugal.

<sup>4</sup> Department of Medical Sciences, University of Aveiro, Aveiro, Portugal.

<sup>5</sup> Lab3R – Respiratory Research and Rehabilitation Laboratory, School of Health Sciences, University of Aveiro (ESSUA), Aveiro, Portugal.

## Introduction

Chronic obstructive pulmonary disease (COPD) is a major cause of morbidity and mortality worldwide [1] and it is characterised by chronic and irreversible airflow obstruction [2]. Little prospect exists for the development of COPD-specific biological treatments thus, research emphasis has been on prevention and establishment of causes, risk factors and covariates of severity and progression of COPD [1]. Multiple correspondence analysis (MCA), a variant of correspondence analysis (CA), is an unsupervised multivariate procedure which can detect, identify and represent underlying structures and non-linear interactions in a data set of categorical variables [3]. Joint correspondence analysis (JCA) is an extension of MCA that attempts to remedy discrepancies between CA and MCA and improve the interpretability of the graphics [4]. In a heterogeneous and complex disease as COPD, these procedures may be useful to enhance our knowledge on the disease. This study is aimed to illustrate the applicability of MCA and JCA in analysing a cohort of people with COPD.

## Methods

A database comprised of categorical nominal data of 498 individuals was used in MCA (with Burt matrix adjusted inertias) and JCA. Sociodemographic and clinical variables were characterized using descriptive statistics. Both methodologies were applied to all available variables with the exception of the GOLD-group and Group variables. Several numerical variables were converted into categories, including age, body mass index (BMI), smoking status (pack/years), comorbidities-Charlson comorbidity index (CCI), number of exacerbations in the previous year, lung function-forced expiratory volume in one second percentage predicted (FEV<sub>1</sub>.pp) and forced vital capacity percentage predicted (FVC.pp), impact of the disease-COPD assessment test (CAT), dyspnea modified British Medical Research Council questionnaire (mMRC), modified Borg scale (Dysp.Borg), symptoms of anxiety and depression-Hospital Anxiety and Depression Scale (HADS), Handgrip strength of dominant hand percentage predicted (Handgrip.pp), quadriceps muscle strength-Handheld Dynamometer (QMS.pp), functionality-1-minute sit-to-stand test (STS\_1min.pp), fatigue modified Borg scale (Fatig.Borg) and peripheral oxygen saturation (SpO<sub>2</sub>). A K-means algorithm (using Minkowski distance) was applied to JCA object scores (individuals) and the number of clusters was estimated by 30 indices, including silhouette and gap-statistic [5]. The optimal number of clusters was chosen by majority concordance of the indices used. For better interpretability, all graphic outputs were derived from JCA (except for Figure 1). To assess the concordance between the formed clusters and the COPD/Healthy groups the adjusted Rand Index was calculated. Furthermore, to assess if variable categories proportion between clusters were significantly different from each other, multiple proportion tests were performed. All statistical analysis was performed using R (version 4.0.1) [6], some graphic visualization obtained using R package ggplot2 [7].

### Keywords:

Chronic obstructive pulmonary disease, Joint correspondence analysis, Multiple correspondence analysis.

### Corresponding author:

Daniel Castro  
[d.castro@ua.pt](mailto:d.castro@ua.pt)

### Conflict of interest:

The authors declare no conflict of interests

First published: 23 OCT 2020



Open Access Publication

© 2020 Castro D, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Table 1** - Summary of sociodemographic and clinical characteristics of the sample (270 people with Chronic Obstructive Pulmonary Disease: chronic obstructive pulmonary disease and 228 age-, sex, and body mass index healthy individuals).

Variables		Category	Group	
			COPD N=270 54%	Healthy N=228 46%
Age years		<50	16(3.2%)	7(1.4%)
		[50, 60[	40(8%)	45(9%)
		[60,70[	104(20.9%)	92(18.5%)
		[70,80[	81(16.3%)	67(13.4%)
		>=80	29(5.8%)	17(3.4%)
Gender		Female	60(12%)	59(11.8%)
		Male	210(42.7%)	169(33.9%)
Education level		Primary or less	152(30.5%)	123(24.7%)
		Intermediate	93(18.7%)	82(16.4%)
		Higher	25(5%)	23(4.6%)
Body Mass Index (BMI) kg/m2		<18.5	9(1.8%)	0
		[18.5,25[	95(19.1%)	59(11.8%)
		[25,30[	103(20.7%)	115(23.1%)
		[>=30]	63(12.7%)	54(10.1%)
Smoking status (Pack/years)		0	58(11.6%)	164(32.9%)
		]0,25]	51(10.2%)	39(7.8%)
		]25,50]	83(16.7%)	14(2.8%)
		>50	78(15.7%)	11(2.2%)
Charlson Comorbidity Index (CCI) total score		[0,2[	12(2.4%)	45(9%)
		[2,5[	190(38.1%)	177(35.5%)
		>=5	68(12.6%)	6(1.2%)
Lung function	Forced Expiratory Volume in one second percentage predicted (FEV <sub>1</sub> .pp)	<30	28(5.6%)	0
		[30,50]	100(20%)	0
		]50,79]	107(21.5%)	20(4%)
		>=80	35(7%)	208(41.8%)
	Forced vital capacity percentage predicted (FVC.pp)	[30,50]	18(3.6%)	1(0.2%)
		]50,80[	122(24.5%)	42(8.4%)
		>=80	130(26.1%)	185(37.1%)
Number of Exacerbations in the previous year		[0,1]	199 (40%)	225(45.2%)
		>=2	71(14.3%)	3(0.6%)
COPD Assessment Test (CAT, points)		[0,9]	78(15.7%)	183(36.7%)
		]9,21[	128(26.7%)	43(8.6%)
		[21,30[	53(10.6%)	2(0.4%)
		>=30	11(2.2%)	0
Dyspnea	Modified British Medical Research Council dyspnea (mMRC, points)	grade0	40(8%)	161(32.3%)
		grade1	89(17.9%)	56(11.2%)
		grade2	69(13.9%)	7(1.4%)
		grade3	63(12.7%)	4(0.8%)
		grade4	9(1.8%)	0
	Modified Borg scale (Dysp.Borg)	0	171(34.4%)	209(42%)
		[0.5,2]	48(9.6%)	11(2.2%)
		3	31(6.2%)	6(1.2%)
		[4,6]	17(3.4%)	2(0.4%)
		[7,10]	3(0.6%)	0
GOLD (Global Initiative for Chronic Obstructive Lung Disease)	Groups <sup>a</sup>	A	102(5%)	——
		B	50(18.5%)	——
		C	86(31.8%)	——
		D	32(11.9%)	——
HADS (Hospital Anxiety and Depression Scale)	HADS Anxiety	Abnormal	95(19.1%)	54(10.8%)
		Normal	175(35.1%)	174(35%)
	HADS Depression	Abnormal	105(21.1%)	39(7.9%)
		Normal	165(33.1%)	189(38%)
Physical activity (BPAAT: brief physical activity assessment tool)		Insufficiently	187(37.6%)	135(27.1%)
		Sufficiently	83(16.7%)	93(18.7%)
Upper limb muscle strength (Handgrip strength of dominant hand percentage predicted: Handgrip.pp, kg)		[0,80[	20(4%)	14(2.8%)
		[80,100[	58(11.6%)	27(5.4%)
		>100	192(38.5%)	187(37.5%)
Quadriceps muscle strength (Hand held dynamometry percentage predicted: QMS.pp, kgf)		[0,80[	189(38%)	131(26.3%)
		[80,100]	54(10.8%)	64(12.9%)
		>100	27(5.4%)	33(6.6%)
Functionality (1-minute sit-to-stand test percentage predicted: STS_1min.pp)		[0,80[	143(28.7%)	46(9.2%)
		[80,100]	68(13.6%)	71(14.3%)
		>=100	59(11.8%)	111(22.3%)
Fatigue score (Modified Borg scale: Fatig.Borg)		[0,0.5[	168(33.7%)	195(39.2%)
		[0.5,2]	43(8.6%)	21(4.2%)
		]2,3]	33(6.6%)	7(1.4%)
		]3,6]	20(4.0%)	5(1%)
		]6,10]	6(1.2%)	0
Peripheral oxygen saturation (SpO <sub>2</sub> )		[85,90[	14(2.8%)	0
		[90,94]	84(17%)	24(4.8%)
		]94,100]	172(34.5%)	204(41%)

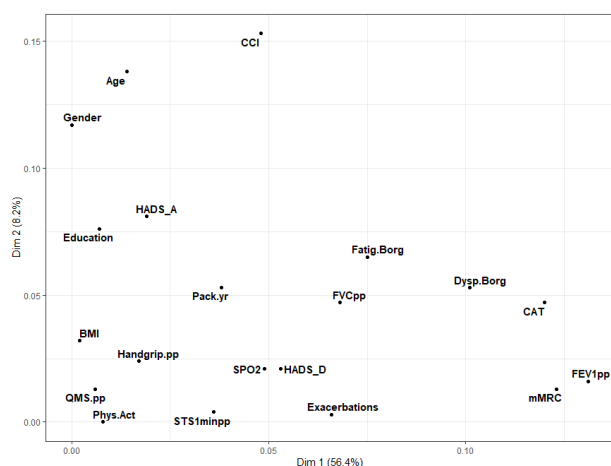
<sup>a</sup>variable not included in multiple or joint correspondence analysis..

## Results

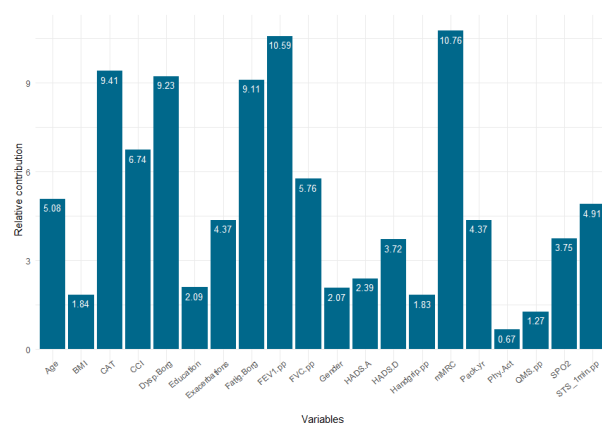
270 people with COPD (between 38 to 90 years, 210 male; GOLD stages: 1-100(37%), 2-86(31.9%), 3-65(24.1%), 4-19(7%), GOLD groups: A-102(5%), B-50(18.5%), C-86(31.8%), D-32(11.9%)) and 228 age-, sex- and BMI-matched healthy individuals (between 43 to 90 years, 169 male) were used in the analysis (Table 1). MCA variable contribution values are presented in Table 2 and Figure 1. Dimension 1 (Dim1) and 2 (Dim2) presented eigenvalues of 0.0376 and 0.0055, explained 56.43% and 8.24% of total variance, respectively (In total, all dimensions explained 80.5%). In JCA, the eigenvalues for Dim1 and Dim2 were 0.047 and 0.011 respectively, with a total explained variance of 98.5% from all dimensions (JCA biplot is shown in Figure 5). Variable contribution values for both techniques can be observed in Table 1, Figure 1 and 2. Clustering analysis on JCA outputs revealed 2 distinct clusters (C1 and C2, ARI:0.4): Healthy individuals were mostly grouped in C1 (73.3%, within cluster) and whereas subjects with COPD were mostly grouped in C2 (93.2%, within cluster). Significant differences in the proportion of variable categories between clusters were found (results of variables with relevant (>5%) contribution are presented in Table 3) except for age under 50 years, FEV<sub>1</sub>pp with percentage from 50 to 79, Dysp.Borg from 7 to 10 and Fatig.Borg score from 0.5 to 2.

**Table 2** - Multiple and Joint variable relative contribution values (%).

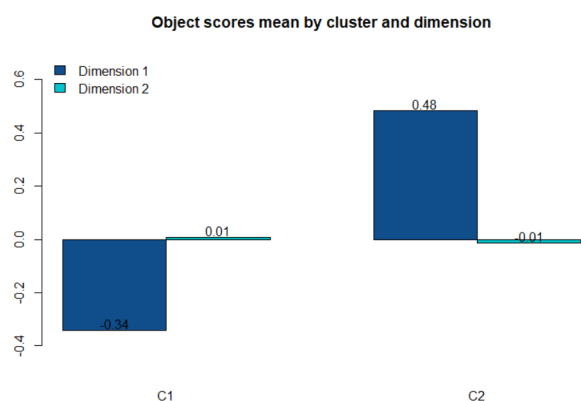
Variables	MCA		JCA
	Dim1	Dim2	
Age	01.4	13.8	5.08
Gender	0	11.7	2.07
Education level	0.7	7.6	2.09
Body composition - Body mass index (BMI <sup>1</sup> )	0.2	3.2	1.84
Smoking status (Pack/years <sup>1</sup> )	3.8	5.3	4.37
Comorbidities - Charlson Comorbidity Index (CCI <sup>1</sup> )	4.8	15.3	6.74
Lung function	Forced expiratory volume in one second % predicted (FEV <sub>1</sub> .pp <sup>1</sup> )	13.1	1.6
	Forced vital capacity % predicted: (FVC.pp <sup>1</sup> )	6.8	4.7
Number of exacerbations in the previous year	6.6	0.3	4.37
Impact of the disease - COPD assessment test total score (CAT <sup>1</sup> )	12	4.7	9.41
Dyspnea	Modified British Medical Research Council dyspnea (mMRC <sup>1</sup> )	12.3	1.3
	Modified Borg scale (Dysp.Borg <sup>1</sup> )	10.1	5.3
Hospital Anxiety and Depression Scale (HADS)	Anxiety (HADS_anxiety <sup>1</sup> )	1.9	8.1
	Depression (HADS_depression <sup>1</sup> )	5.3	2.1
Physical Activity - brief physical assessment tool (BPAAT)	0.8	0	0.67
Upper limb muscle strength - Handgrip strength of dominant hand % predicted (Handgrip.pp <sup>1</sup> ), kg	01.7	2.4	1.83
Quadriceps muscle strength - Hand held dynamometry % predicted (QMS.pp <sup>1</sup> ), kgf	0.6	1.3	1.27
Functionality - 1-minute sit-to-stand test % predicted (STS_1min.pp <sup>1</sup> )	3.6	0.4	4.91
Fatigue score - Modified Borg scale (Fatig.Borg <sup>1</sup> )	7.5	6.5	9.11
Peripheral oxygen saturation (SpO <sub>2</sub> <sup>1</sup> )	4.9	2.1	3.75
% of explained inertia	56.4%	8.2%	98%



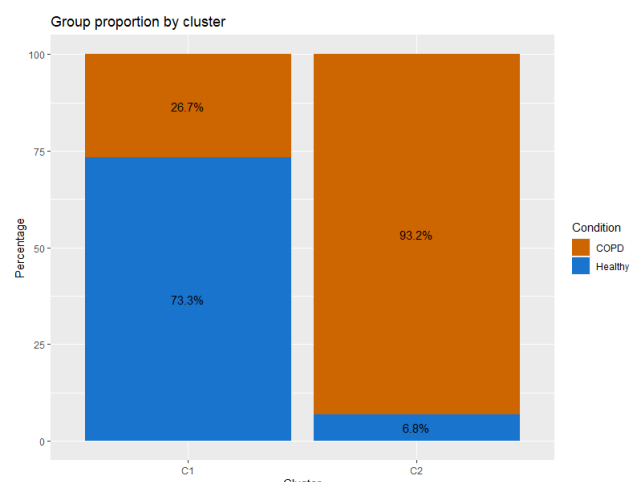
**Figure 1** - Multiple correspondence analysis variable contribution values by dimension.



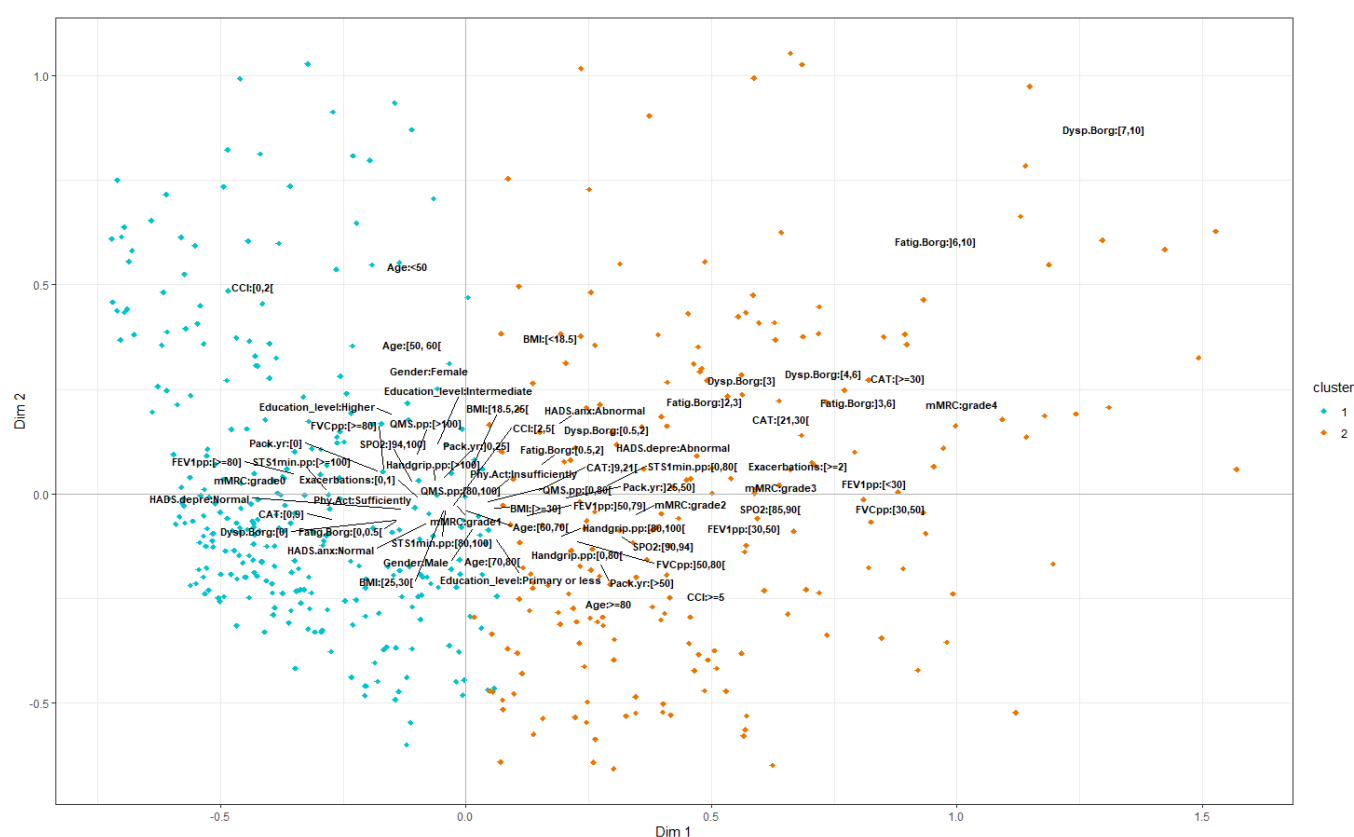
**Figure 2** - Joint correspondence analysis variable contribution



**Figure 3** - Object scores mean (of individuals) by cluster and dimension in Joint Correspondence Analysis..



**Figure 4** - Distribution of people with Chronic Obstruction Pulmonary Disease and healthy individuals by cluster in Joint Correspondence Analysis.



**Figure 5** - Joint Correspondence Analysis biplot of variable categories and individuals by cluster

## Discussion and conclusions

In MCA, the most contributing variables for the first dimension (D1) mainly comprise of ones connected to the disease, either physiological (lung function) or patient perceived (dyspnea and impact of the disease CAT) whereas in the second dimension (D2) is more related to risk factors. In JCA, individuals with COPD were grouped in both clusters with significant differences in category proportion. The attained results suggest that these methodologies might be useful for unravel the complexity and diversity associated with COPD.174(35%)

**Table 3** - Crosstabulations between relevant joint correspondence analysis variables and clusters.

Variables	Category	Cluster		Proportion test
		C1 n(%)	C2 n(%)	p-value
<b>Age</b> years	<50	15(3%)	8(1.6%)	0.2056
	[50, 60[	56(11.2%)	29(5.8%)	0.003191
	[60, 70[	119(23.9%)	77(15.5%)	0.001084
	[70, 80[	87(17.5%)	61(12.2%)	0.02594
	>=80	15(3%)	31(6.2%)	0.02354
<b>Charlson Comorbidity Index (CCI)</b> total score	[0, 2[	51(10.2%)	6(1.2%)	1.946e-09
	[2, 5[	230(46.2%)	137(27.5%)	1.512e-09
	>=5	11(2.2%)	63(12.7%)	7.185e-10
<b>Lung function</b>	Forced Expiratory Volume in one second percentage predicted (FEV <sub>1</sub> .pp)	<30	26(5.2%)	1.038e-05
		[30, 50[	6(1.2%)	< 2.2e-16
		[50, 79[	94(18.9%)	0.1285
		>=80	72(14.5%)	< 2.2e-16
	Forced vital capacity percentage predicted (FVC.pp)	[30, 50[	14(2.8%)	0.0002104
		[50, 80[	18(3.6%)	1.317e-05
		>=80	108(21.7%)	< 2.2e-16
<b>COPD Assessment Test</b> (CAT, points)	[0, 9[	227(45.6%)	34(6.8%)	< 2.2e-16
	[9, 21[	64(12.9%)	107(21.5%)	0.0004171
	[21, 30[	1(0.2%)	54(11%)	5.446e-13
	>=30	0	11(2.2%)	0.00243
<b>Dyspnea</b>	Modified British Medical Research Council dyspnea (mMRC, points)	grade0	13(2.6%)	< 2.2e-16
		grade1	92(18.5%)	0.0006401
		grade2	53(11%)	2.137e-09
		grade3	57(11.4%)	3.202e-14
		grade4	64(12.9%)	0.007389
		0	9(1.8%)	< 2.2e-16
	Modified Borg scale (Dysp.Borg)	[0, 5, 2[	107(21.5%)	0.0001711
		3	44(9%)	2.717e-06
		4(0.8%)	33(6.6%)	3.053e-05
		[4, 6[	19(3.8%)	0.2475
		[7, 10[	3(0.6%)	< 2.2e-16
<b>Fatigue score</b> (Modified Borg scale: Fatig.Borg)	[0, 0, 5[	259(52%)	104(21%)	0.2448
	[0, 5, 2[	27(5.4%)	37(7.4%)	2.865e-06
	[2, 3[	5(1%)	35(7%)	8.339e-06
	[3, 6[	1(0.2%)	24(4.8%)	0.04062
	[6, 10[	0	6(1.2%)	

## Funding

This work was supported by PRIME (PTDC/SAU-SER/28806/2017), iBiMED (UIDB/04501/2020) and GenomePT (POCI-01-0145-FEDER-022184), under the scope of Programa Operacional de Competitividade e Internacionalização – POCI, through Fundo Europeu de Desenvolvimento Regional – FEDER, and Fundação para a Ciência e Tecnologia – OE.

## References

1. Rabe Klaus, Watz Henrik. Chronic obstructive pulmonary disease. Lancet 2017;389(10082):1931–40. [https://doi.org/10.1016/S0140-6736\(17\)31222-9](https://doi.org/10.1016/S0140-6736(17)31222-9)
2. WHO. World Health Organization [Internet]. International: WHO. 2020 [cited 2020 Mar 16]. Available from: <https://www.who.int/respiratory/copd/definition/en/>
3. Salkind N. Encyclopedia of Measurement and Statistics [Internet]. Thousand Oaks (CA): Sage; 2007 [cited 2020 Mar 16]; [about 13 p.]. Available from: <https://personal.utdallas.edu/~herve/Abdi-MCA2007-pretty.pdf>
4. Camiz S, Gomes GC. Joint Correspondence Analysis Versus Multiple Correspondence Analysis: A Solution to an Undetected Problem [Internet]. In: Giusti A., Ritter G., Vichi M. (eds) Classification and Data Mining. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg; 2013 Sep [https://doi.org/10.1007/978-3-642-28894-4\\_2](https://doi.org/10.1007/978-3-642-28894-4_2)
5. Charrad M., Ghazzali N., Boiteau V., Niknafs A. “NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set”. Journal of Statistical Software. 2014;61(6):1-36. <https://doi.org/10.18637/jss.v061.i06>
6. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2020 [cited 2020 Jun 21]. Available from: <https://www.R-project.org/>
7. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. 2016 [cited 2020 July 6]. Available from: <https://ggplot2.tidyverse.org>