Corrigendum and addendum to the paper „Statistics on the use of AI technologies in the member states of the EU"

Imre Dobos
Budapest University of Technology and Economics, Department of Economics, Hungary
Email: dobos.imre@gtk.bme.hu

Gergely Lülök
Budapest University of Technology and Economics, Department of Management and Business Economics, Hungary
Email: lulok.gergely@edu.bme.hu

Zoltán Sebestyén
Budapest University of Technology and Economics, Department of Management and Business Economics, Hungary
Email: sebestyen.zoltan@gtk.bme.hu

Keywords:
artificial intelligence (AI), statistical analysis, European Union, Eurostat, multicollinearity

Online first publication date: 4 June 2026

Lülök et al. (2026) examined the areas of use of artificial intelligence (AI) in the European Union in their paper. One element of the statistical analysis was to examine which areas of use have a close statistical relationship, i.e., collinearity between variables. In this paper, the authors correct an inaccuracy in the collinearity analysis and analyse the problem using additional methods beyond the variance inflation factor (VIF). Such analyses can be variance proportion (VP) and cluster analysis.

Introduction

Lülök et al. (2026) examined the use of artificial intelligence in 11 applications. One of the questions that arose was how they relate to each area of use, what linear relationship, meaning what correlation relationship, they have. In the paper, the authors used the variance inflation factor method to filter out multicollinearity. However, a minor error was found during the analysis. The primary goal of the

current paper is to address a minor deficiency and to apply two additional methods to filter out collinearity, framing this step as a methodological refinement rather than a simple error correction.

In the rest of the corrigendum, we first correct the minor inaccuracy. While the variance inflation factor is suitable for signaling problematic variables, it does not clearly identify which variable pairs or groups cause the collinearity. Therefore, the substantive added value of this corrigendum is comparing three methods: the variance inflation factor, variance proportions, and variable-level cluster analysis.

Strong correlations among artificial intelligence technologies are not only expected in practice. Still, they are structurally inherent to technology co-diffusion processes, such as the joint deployment of machine learning with robotic process automation, text mining with speech recognition, and production with logistics. The professional focus of this corrigendum is to rigorously distinguish genuine technological co-evolution from statistical redundancy arising from multicollinearity. In addition to detecting collinearity, the aim of the methods used was to exclude the collinear statistical variable from further analyses. Importantly, the goal is not to question the main conclusions of the original article, but to improve the methodological foundation of variable selection and further analyses.

Testing collinearity with the three proposed methods

There are a number of methods available for examining collinearity. Ullah et al. (2019) describe 19 different methods for screening for linear relationships in their paper. This paper uses only three methods.

The first methods (VIF) only show whether a statistical variable is collinear or not. However, it does not give an answer to which other variables the filtered variable is collinear with (Vörösmarty–Dobos, 2020).

Cluster analysis gives an answer to which variables have a collinear relationship, but does not show which variable is worth filtering. In this case, in the cluster analysis, we use the Pearson correlation coefficient as the (half)-distance between two elements, instead of the usual Euclidean distance (Field 2010).

Finally, the variance proportion method – in addition to the linear relationship – also gives information about which variable is worth omitting (Belsley 1991).

Variance inflation factor

In the article, Lülök et al. (2026) only this method was used. This indicator is a function of the R^2 of a given variable in terms of the remaining variables, i.e., for the j^{th} variable

$$VIF_j = \frac{1}{1-R_j^2}.$$

Then the variable is filtered out sequentially by reducing the number of variables. As a stopping criterion, we choose that the VIF value should decrease to 5 or below 10. In most articles, the value 5 is chosen.

The study by Lülök et al. (2026) neglected this latter condition on page 15, as it states that the results obtained with VIF can be shown to be correlated with other variables, which is not fulfilled. However, the numbers highlighted in bold in the diagonal of Table 4 on page 18 are the VIF values. The table shows the VIF elements of variables remaining after filtering out the collinear elements, which are not needed in the analysis.

Table 1 shows here the excluded variables after the current VIF in the sequential method. We chose 5 as the target VIF value.

Table 1

Sequential elimination of multicollinear variables

Omitting variable	VIF value
Machine learning	15.222
Natural language generation	12.317
Production	9.234
Robotic process automation	6.552
Logistics	6.019

Source: own compilation with SPSS 31.

The correlation matrix shows which variable has collinearity, i.e., correlation. However, we cannot do this when applying the method, but only afterward, which does not guarantee that collinearity actually exists between these variables.

Collinearity analysis with cluster analysis

During cluster analysis, the collinearity test can be performed on the variables. In this case, the result is independent of the distance between the clusters, but when the distance between the variables is taken into account, the Pearson correlation is chosen. The dendrogram in Figure 1 shows which variables are closely correlated.

Figure 1 shows that there are four pairs of variables close to each other. These are with correlation:

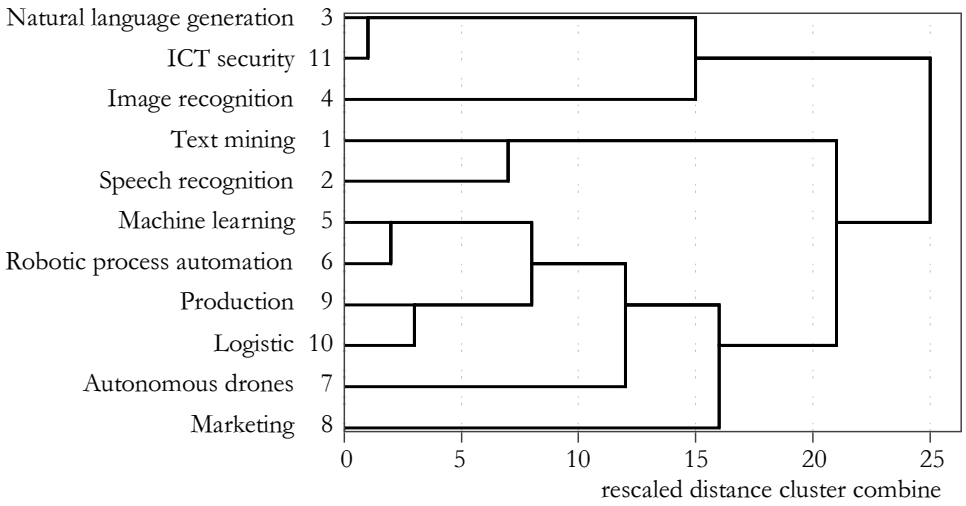
natural language generation	ICT security	0.900,
text mining	speech recognition	0.801,
machine learning	robotic process automation	0.878,
production	logistics	0.865.

This method, unlike VIF, can determine which variables have high collinearity and thus high correlation. However, it cannot determine which of the four pairs of variables can be neglected. It can also be seen that while in the case of VIF we can consider 5 variables as collinear with the others, in this case, only 4. Of the remaining

7 variables, this method does not provide information about which variable we can leave out.

Figure 1

Dendogram using average linkage (between groups)



Source: own compilation with SPSS 31.

Variance of proportions

The variance of proportions method was proposed by Belsley (1991) to filter out multicollinearity. The method starts from a matrix of variables normalized to unity and is based on a condition index formed from the eigenvalues of the transformed matrix. The calculations of the method are included in SPSS 31, which are presented in Table 2.

The variance of proportion values can be obtained from the linear regression equation, but the regression model will then be constant-free, through the origin.

In the variance proportions method, the condition index of the product of the normalized data matrix and the transposed matrix must be obtained, and then the eigenvalue of this matrix is examined. The definition of the condition index for the *j*-th eigenvalue is

$$CI_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}},$$

where λ_{max} is the largest eigenvalue, while λ_j is the jth largest eigenvalue of the product matrix. Then we get Table 2, where the columns of the variables contain the normalized variances.

Table 2

Variance proportions for variables

model		Collinearity diagnostics[a) b)]		variance proportions										
		eigen-value	condi-tion index	text mining	speech recog-nition	natural language generation	image recog-nition	machine learning	robotic process automation	autono-mous drones	marketing	produc-tion	logistics	ICT security
1	1	9.763	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2	0.445	4.682	0.02	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02
	3	0.249	6.257	0.02	0.11	0.00	0.02	0.02	0.00	0.02	0.03	0.00	0.00	0.00
	4	0.211	6.809	0.01	0.01	0.02	0.04	0.02	0.00	0.07	0.03	0.01	0.01	0.01
	5	0.114	9.240	0.00	0.03	0.00	0.03	0.04	0.00	0.04	0.27	0.01	0.00	0.03
	6	0.101	9.818	0.02	0.08	0.04	0.02	0.03	0.00	0.19	0.04	0.03	0.00	0.02
	7	0.042	15.177	0.05	0.11	0.00	0.02	0.03	0.29	0.06	0.02	0.03	0.24	0.01
	8	0.026	19.431	0.81	0.48	0.05	0.00	0.02	0.31	0.19	0.02	0.00	0.00	0.00
	9	0.019	22.617	0.04	0.02	0.01	0.05	0.05	0.03	0.11	0.02	0.73	0.46	0.07
	10	0.016	24.765	0.02	0.02	0.00	0.81	0.80	0.10	0.01	0.00	0.19	0.10	0.18
	11	0.012	28.177	0.00	0.14	0.87	0.01	0.00	0.25	0.31	0.58	0.01	0.17	0.66

a) Dependent variable: sum of variables.
b) Linear regression through the origin.
Source: own compilation with SPSS 31.

Three types of collinearity can be determined using the condition index: weak collinearity between 0–15, moderate collinearity between 15–30, and strong collinearity above 30.

Moderate collinearity exists for the European data on artificial intelligence, as our condition indices fall between 15 and 30 for 5 pairs of variables. However, for each pair of variables, we can filter out the variable with the highest variance proportion value in the given row.

The pairs will be as follows, with their corresponding correlations:

natural language generation	ICT security	0.900,
image recognition	machine learning	0.418,
production	logistics	0.865,
text mining	speech recognition	0.801,
robotic process automation	logistics	0.818.

The five variables to be omitted are marked in bold.

Comparison of the three methods

In the case of the three methods, only pairs can be determined in two of them, of which one of the collinear variables can be omitted from the analysis; these are *VIF* and *cluster analysis*, while the *variance proportions (VP) method* already shows the variable that is collinear with the variable to be omitted. Table 3 shows the results of the three methods.

Table 3

Variables remaining after applying the three methods

Method variables	Variance inflation factor	Cluster analysis	Variance of proportions
1.	Natural language generation	Natural language generation	Natural language generation
2.	Machine learning	Machine learning	Image recognition
3.	Production	Production	Production
4.	Logistics	Text mining	Text mining
5.	Robotic process automation	–	Robotic process automation

Source: own compilation with SPSS 31.

For pairs obtained using the cluster analysis method, we used the variables listed above, because this method does not provide information about which variable should be omitted.

Conclusion

This paper shows that the three collinearity screening methods act as complementary tools rather than distinct substitutes. While the variance inflation factor serves as an adequate initial filter, it cannot uncover the underlying structure of the collinear

relationships. Cluster analysis successfully visualizes tightly coupled pairs, including natural language generation with information and communication technology security, text mining with speech recognition, machine learning with robotic process automation, and production with logistics. Despite this visual clarity, the clustering approach offers no mathematical guidance on which variable should be omitted. The variance proportions technique fills this gap as a superior diagnostic instrument, simultaneously confirming collinearity and pinpointing the exact variables to drop.

The examined dataset exhibits moderate rather than extreme collinearity, evidenced by condition index values ranging from 15 to 30, a pattern that is characteristic of technology adoption datasets where multiple innovations evolve jointly. Recognizing this moderate level provides researchers with a nuanced view of the data structure. Implementing this improved methodology establishes a more dependable foundation for principal component analysis, future clustering, and complex cross-country comparisons, while reducing the risk of biased interpretation of inter-technology relationships. More importantly, these methodological corrections refine and solidify the original claim that the spread of artificial intelligence technologies creates structured, interconnected technological patterns across the member states of the European Union.

Finally, the act of dropping a variable cannot rely solely on mechanical statistical thresholds; it requires informed professional judgment grounded in both statistical diagnostics and domain-specific technological understanding. Highly correlated variables often capture genuine technological co-evolution. Purely statistical elimination risks removing valuable data about how these systems are synergistically deployed together in real business environments.

Acknowledgement

This research was supported by the University Research Scholarship Programme (EKÖP) within the framework of the Cooperative Doctoral Program (EKÖP-KDP-25), funded by the Ministry of Culture and Innovation and the National Research, Development and Innovation Fund.

REFERENCES

BELSLEY, D. A. (1991): A guide to using the collinearity diagnostics *Computer Science in Economics and Management* 4 (1): 33–50.

FIELD, D. A. (2010): *Discovering statistics using SPSS* Sage publications.

LÜLÖK, G.–DOBOS, I.–SEBESTYÉN, Z. (2026): Statistics on the use of AI technologies in the member states of the EU *Regional Statistics* 16 (1): 3–32. https://doi.org/10.15196/RS160101

ULLAH, M. I.–ASLAM, M.–ALTAF, S.–AHMED, M. (2019): Some new diagnostics of multicollinearity in linear regression model *Sains Malaysiana* 48 (9): 2051–2060.

VÖRÖSMARTY, G.–DOBOS, I. (2020): Green purchasing frameworks considering firm size: a multicollinearity analysis using variance inflation factor *Supply Chain Forum: An International Journal* 21 (4): 290–301.