Construction of an Immigrant Integration Composite Indicator through the Partial Least Squares Structural Equation Model K-Means

Venera Tomaselli, Mario Fordellone, and Maurizio Vichi

Abstract Integration is a multidimensional process, which can take place in different ways and at different times in relation to each of the single economic, social, cultural, and political dimensions. Hence, examining every single dimension is important as well as building composite indexes simultaneously inclusive of all dimensions in order to obtain a complete description of a complex phenomenon and to convey a coherent set of information. In this paper, we aim at building an immigrant integration composite indicator (IICI), able to measure the different aspects related to integration such as employment, education, social inclusion, active citizenship, and on the basis of which to simultaneously classify territorial areas such as European regions. For this application, the data collected in 274 European regions from the European Social Survey (ESS), Round 8, on immigration have been used.

1 Introduction

The immigrants' integration is a multidimensional process implying many economic, social, cultural, and political issues. This process is carried out according to several steps and in different conditions determining continuous redefinition of accomplishment outcomes. In fact, each single dimension, diachronically positioned over time, generates different integration levels. Hence, examining each single dimension is important as well as building composite indexes simultaneously comprehensive of all dimensions in order to obtain a complete description of a complex phenomenon and to convey a suitable set of information.

V. Tomaselli (✉)
Department of Political and Social Sciences, University of Catania, 8, Vitt. Emanuele II -95131, Catania, Italy
e-mail: venera.tomaselli@unict.it

M. Fordellone · M. Vichi
Department of Statistical Sciences, La Sapienza, University of Rome, Rome, Italy
e-mail: mario.fordellone@uniroma1.it
e-mail: maurizio.vichi@uniroma1.it

P. Mariani and M. Zenga (eds.), *Data Science and Social Research II*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-51222-4_28

According to the literature Entzinger (2000), Entzinger and Biezeveld (2003), the concept of integration can be broken down into different dimensions. Firstly, the socioeconomic dimension refers to housing conditions, work conditions, and income. The legal-political dimension takes into account the theme of citizenship and the rights of political participation, from the freedom of association to the voting right, which in some countries can be used at local government elections even without having achieved the citizenship status of the host country. Finally, the cultural and social dimension considers several elements, among which are knowledge of the language (Vermeulen 2004), free times activities, and access to information.

Due to the multidimensional nature of the integration concept, many studies underline the difficulty to identify core indicators (Ager and Eyber 2002; Strang et al. 2003) able to measure the integration level taking into account each dimension and subdimension of the integration concept (Cesareo and Blangiardo 2009). The factors more strictly connected to the host country approach toward migrants and also those related to country's socioeconomic conditions affect migrant integration (Di Bartolomeo et al. 2015) both at the local and regional levels (OECD 2018).

In this paper, we aim at providing a methodological proposal to build an immigrant integration composite indicator (IICI), able to measure the different aspects related to integration such as employment, education, social inclusion, and active citizenship and by which simultaneously to classify territorial areas (OECD 2008). With this in mind, we analyze the data collected in 274 European regions from European Social Survey (ESS), Round 8, by the structural equation modeling estimated via partial least squares (PLS-SEM) approach introduced by Lohmoller (1989) and developed by Tenenhaus et al. (2005).

In particular, we perform a simultaneous nonhierarchical clustering and partial least squares modeling, named partial least squares structural equation model k-means (PLS-SEM-KM), recently proposed by Fordellone and Vichi (2018), in order to obtain an immigrant integration composite indicator (IICI) and a clustering of the European regions.

Differently from the PLS-SEM methods, PLS-SEM-KM mainly focuses on the homogeneity between and within clusters of regions derived by a unique structural measurement model on immigrant integration. Thus, this study aims at both segmenting the immigrant population and simultaneously identifying the structural (i.e., the latent dimensions explaining the immigrants' integration) and measurement relations (i.e., the observed variables employed to build the latent dimensions) which have produced the segmentation among European regions grouped for immigrants' integration level.

The paper is structured as follows: in Sect. 2, a brief background on the PLS-SEM notation is provided. In Sect. 3, the PLS-SEM-KM model is presented; in Sect. 4, using the ESS data, the results obtained by IICI construction are shown.

2 Background Methods

2.1 Notation

Partial Least Squares (PLS) methodologies are algorithmic tools with analytic properties aiming at solving problems connected with stringent assumptions on data, e.g., distributional assumptions that are hard to meet in real life (Tenenhaus et al. 2005). Tenenhaus et al. try to better clarify the terminology used in the PLS field through a relevant review of the literature, focusing the attention on the Structural Equation Models standpoint.

Before showing the modeling details, the notation and terminology used in this paper are here presented to allow the reader to easily follow the subsequent formalizations and algebraic elaborations:

n, J	# of:	Observations, MVs
H, L, P	# of:	Exogenous LVs, endogenous LVs, LVs ($P = H + L$)
K	# of:	Clusters
Ξ	$n \times H$	Exogenous LVs matrix
$\mathbf{H}$	$n \times L$	Endogenous LVs matrix
$\mathbf{Y}$	$n \times P$	Scores matrix ($\mathbf{Y} = [\Xi, \mathbf{H}]$)
Γ	$L \times H$	Path coefficients matrix of the exogenous LVs
$\mathbf{B}$	$L \times L$	Path coefficients matrix of the endogenous LVs
$\mathbf{Z}$	$n \times L$	Errors matrix of the endogenous LVs
$\mathbf{X}$	$n \times J$	Data matrix
$\mathbf{E}$	$n \times J$	Errors matrix of the data
Λ_H	$J \times H$	Loadings matrix of the exogenous LVs
Λ_L	$J \times L$	Loadings matrix of the endogenous LVs
Λ	$J \times P$	Loadings matrix ($\Lambda = [\Lambda_H, \Lambda_L]$)
$\mathbf{T}$	$n \times H$	Errors matrix of the exogenous LVs
Δ	$n \times L$	Errors matrix of the endogenous LVs
$\mathbf{U}$	$n \times K$	Membership matrix (binary and row stochastic)

Usually, a PLS-SEM (called also PLS-PM, i.e., PLS path model) consists in a combination of two models:

- a structural model (or inner model) that specifies the relationships among latent variables (LVs). In this context, an LV is an unobservable variable (i.e., connected with a theoretical construct) indirectly described by a block of observable variables which are called manifest variables (MVs);
- a measurement model (or outer model) that relates the MVs to their LVs.

2.2 Structural Model

Let $\mathbf{X}$ be an $n \times J$ data matrix, with P endogenous and exogenous latent variables ($P \leq J$), let $\mathbf{H}$ be the $n \times L$ matrix of the endogenous LVs with generic element $\eta_{i,l}$, and let Ξ be the $n \times H$ matrix of the exogenous LVs with generic element $\xi_{i,h}$; the structural model is a causality model that relates the P LVs to each other through a set of linear equations (Vinzi et al. 2010). In matrix form:

$$\mathbf{H} = \mathbf{H}\mathbf{B}^T + \Xi\Gamma^T + \mathbf{Z} \tag{1}$$

where $\mathbf{B}$ is the $L \times L$ matrix of the path coefficients $\beta_{l,l}$ associated with the endogenous latent variables; Γ is the $L \times H$ matrix of the path coefficients $\gamma_{l,h}$ associated with the exogenous latent variables; $\mathbf{Z}$ is the $n \times L$ matrix of the residual terms $\zeta_{i,l}$.

Example 1 An example of structural model is shown in Fig. 1.

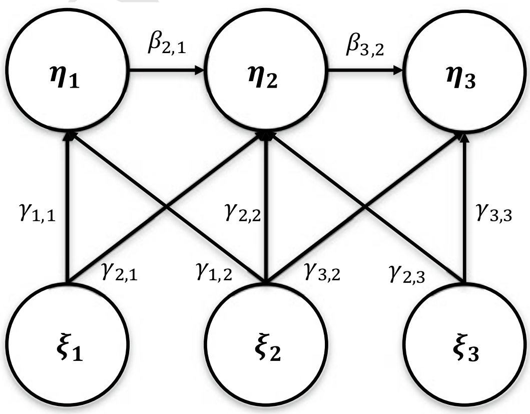

Fig. 1 Example of structural model with three endogenous LVs and three exogenous LVs

486489_1_En_28_Chapter ✓ TYPESET DISK LE ✓ CP Disp.:4/11/2020 Pages: xxx Layout: T1-Standard

2.3 Measurement Model

In PLS-SEM, unlike the traditional SEM approach, there are two ways to relate MVs to their LVs: reflective and formative ways (Diamantopoulos and Winklhofer 2001; Tenenhaus et al. 2005). In the reflective way, it is supposed that each MV reflects its LV, i.e., the observed variables are considered as the effect of the latent construct; a reflective measurement model can be written in matrix form as

$$
\begin{aligned}
\mathbf{X} &= \mathbf{Y}\Lambda^T + \mathbf{E} \\
&= \begin{bmatrix} \Xi & \mathbf{H} \end{bmatrix} \begin{bmatrix} \Lambda_H^T \\ \Lambda_L^T \end{bmatrix} + \mathbf{E} \\
&= \Xi\Lambda_H^T + \mathbf{H}\Lambda_L^T + \mathbf{E}
\end{aligned}
\tag{2}
$$

where Λ_H is the $J \times H$ loadings matrix of the exogenous latent constructs with generic element $\lambda_{j,h}$; Λ_L is the $J \times L$ loadings matrix of the endogenous latent constructs with generic element $\lambda_{j,l}$; $\mathbf{E}$ is the $n \times J$ residuals matrix with element $\epsilon_{i,j}$, which have zero mean and are uncorrelated with $\xi_{i,h}$ and $\eta_{i,l}$. Then, the reflective way implies that each MV is related to its LV by a set of simple regression models with coefficients $\lambda_{j,l}$.

Conversely, in the formative way each MV is supposed to be *forming* its LV, i.e., the observed variables are considered as the cause of the latent construct. Formally, for an exogenous latent construct, the model can be written as

$$
\Xi = \mathbf{X}\Lambda_H + \mathbf{T} \tag{3}
$$

whereas, for endogenous latent construct the model can be written as

$$
\mathbf{H} = \mathbf{X}\Lambda_L + \Delta \tag{4}
$$

where $\mathbf{T}$ and Δ are, respectively, the $n \times H$ and $n \times L$ errors matrices with elements $\tau_{i,h}$ and $\delta_{i,l}$, which have zero mean and are uncorrelated with $x_{i,j}$. Then, the formative way implies that each MV is related to its LV by a multiple regression model with coefficients λs.

Example 2 In Fig. 2, two examples of PLS-SEM with three latent constructs (η_1, ξ_1, and ξ_2) and six observed variables (x_1, x_2, x_3, x_4, x_5, and x_6) are shown. In particular, there are two exogenous LVs (ξ_1 and ξ_2) and one endogenous LV (η_1). The MVs are related to their LVs in reflective way (left plot) and formative way (right plot).

486489_1_En_28_Chapter ✓ TYPESET DISK LE ✓ CP Disp.:4/11/2020 Pages: xxx Layout: T1-Standard

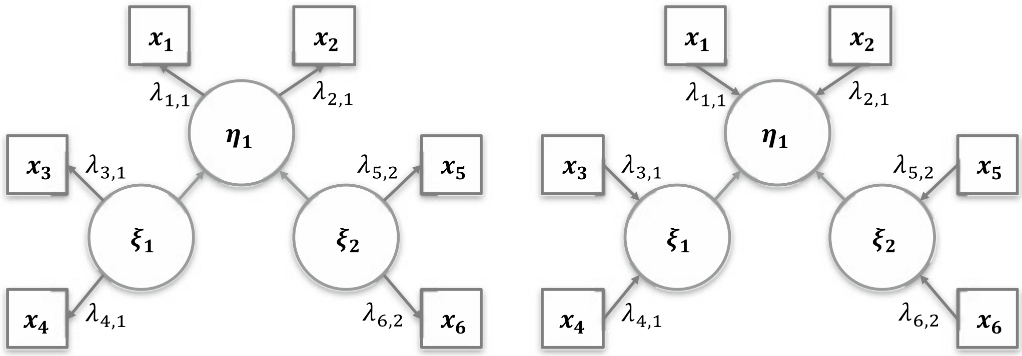

Fig. 2 Two examples of PLS path model with three LVs and six MVs: reflective measurement models (left) and formative measurement models (right)

3 Partial Least Squares *K*-Means

Given the $n \times J$ data matrix $\mathbf{X}$, the $n \times K$ membership matrix $\mathbf{U}$, the $K \times J$ centroids matrix $\mathbf{C}$, the $J \times P$ loadings matrix $\Lambda = [\Lambda_H, \Lambda_L]$, and the errors matrices $\mathbf{Z}$ $(n \times L)$ and $\mathbf{E}$ $(n \times J)$, the partial least squares structural equation model k-means (PLS-SEM-KM) model can be written as follows (Fordellone and Vichi 2018):

$$
\begin{aligned}
\mathbf{H} &= \mathbf{H}\mathbf{B}^T + \Xi\Gamma^T + \mathbf{Z} \\
\mathbf{X} &= \mathbf{Y}\Lambda^T + \mathbf{E} = \Xi\Lambda_H^T + \mathbf{H}\Lambda_L^T + \mathbf{E} \\
\mathbf{X} &= \mathbf{U}\mathbf{C}\Lambda\Lambda^T + \mathbf{E} = \mathbf{U}\mathbf{C}\Lambda_H\Lambda_H^T + \mathbf{U}\mathbf{C}\Lambda_L\Lambda_L^T + \mathbf{E},
\end{aligned}
\tag{5}
$$

subject to constraints: (i) $\Lambda^T\Lambda = \mathbf{I}$; and (ii) $\mathbf{U} \in \{0, 1\}$, $\mathbf{U}\mathbf{1}_K = \mathbf{1}_n$. Thus, the PLS-SEM-KM model includes the PLS and the clustering equations (i.e., $\mathbf{X} = \mathbf{U}\mathbf{C}$ and then, $\mathbf{Y} = \mathbf{X}\Lambda$ becomes $\mathbf{Y} = \mathbf{U}\mathbf{C}\Lambda$). The PLS-SEM-KM algorithm is composed by the following steps:

Algorithm 1 PLS-SEM-KM algorithm

1: Initialize $\Lambda = \mathbf{D}_\Lambda$;
Choose K through the *gap method* applied on scores matrix $\mathbf{Y} = \mathbf{X}\Lambda$;
$\omega = 10^{-12}$, iter=0, maxiter=300;
2: Random generate the memberships matrix $\mathbf{U}$;
Compute centers matrix $\mathbf{C} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{X}$;
Compute latent scores matrix $\mathbf{Y} = \mathbf{U}\mathbf{C}\Lambda$;
3: iter=iter+1;

Inner approximation

4: Estimate covariance matrix $\Sigma_Y = n^{-1}\mathbf{Y}^T\mathbf{J}\mathbf{Y}$ (with $\mathbf{J} = \mathbf{I}n^{-1}\mathbf{1}\mathbf{1}^T$);
5: Compute inner weights $\mathbf{W} = \mathbf{D}_B \otimes \Sigma_Y$;
6: Estimate new scores $\mathbf{Y}_W = \mathbf{Y}\mathbf{W}$;

Outer approximation

7: Update $\Lambda \rightarrow \Lambda_n = \mathbf{C}^T\mathbf{U}^T\mathbf{Y}_W(\mathbf{Y}_W^T\mathbf{Y}_W)^{-1}$; (Reflective way)
$\rightarrow \Lambda_n = (\mathbf{C}^T\mathbf{U}^T\mathbf{U}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{U}^T\mathbf{Y}_W$; (Formative way)
8: Update $\mathbf{U} \rightarrow \underset{\mathbf{U}}{\text{argmin}} \left\|\mathbf{X} - \mathbf{U}\mathbf{C}\Lambda_n\Lambda_n^T\right\|^2$,
subject to $\Lambda_n^T\Lambda_n = \mathbf{1}_P$, $\mathbf{U} = \{0, 1\}$, $\mathbf{U}\mathbf{1}_K = \mathbf{1}_n$;
9: Compute new centers $\mathbf{C}_n = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{X}$;

Stopping rule

10: Update $K \rightarrow Kn$ through the *gap method* applied on scores matrix $\mathbf{Y} = \mathbf{U}\mathbf{C_n}\Lambda_\mathbf{n}$
11: **if** $K_n \neq K$
go to step 2
12: **else**
13: **if** $\|\mathbf{C}\Lambda - \mathbf{C}_n\Lambda_n\|^2 > \omega$ & iter<maxiter, $\mathbf{C} = \mathbf{C}_n$, $\Lambda = \Lambda_n$;
repeat step 3-12;
14: **else**
exit loop 3-12;
15: **end if**
16: **end if**

Path coefficients estimation

17: **for** $l = 1$ to L **do**
18: **for** $h = 1$ to H **do**
19: Compute $\mathbf{Y}_h = \mathbf{X}\Lambda_h$
20: Compute $\mathbf{Y}_l = \mathbf{X}\Lambda_l$
21: Compute $\Gamma = (\mathbf{Y}_{h*}^T\mathbf{Y}_{h*})^{-1}\mathbf{Y}_{h*}^T\mathbf{Y}_l$
22: Compute $\mathbf{B} = (\mathbf{Y}_{l*}^T\mathbf{Y}_{l*})^{-1}\mathbf{Y}_{l*}^T\mathbf{Y}_l$
23: **end for**
24: **end for**

PLS-SEM-KM algorithm is based on the simultaneous optimization of PLS-SEM and reduced k-means (De Soete and Carroll 1994), where centroids of clusters are located in the reduced space of the LVs, thus, ensuring the optimal partition of the statistical units on the best latent hyperplane defined by the structural/measurement relations estimated by the prespecified model. The input parameters are the $n \times J$ standardized data matrix $\mathbf{X}$; the $J \times P$ design matrix of the measurement model $\mathbf{D}_\Lambda$, with binary elements equal to 1 if an MV is associated with an LV and 0 otherwise; the $P \times P$ path design matrix of the structural model $\mathbf{D}_B$, with binary elements equal to 1 if a latent exogenous or endogenous variable explains a latent endogenous variable and 0 otherwise. Matrix $\mathbf{D}_B$ is symmetrized.

Moreover, a different approach to select the optimal number of segments K is provided. In fact, PLS-SEM-KM algorithm includes the optimal K selection through the gap statistics proposed by Tibshirani et al. (2001). This statistics is embedded in the algorithm for estimating simultaneously the number of clusters together with PLS-SEM. In fact, the *gap method* may be applicable to any model-based clustering approach without restrictive assumptions on the scores distribution and therefore, is a valid method to be included in our methodology.

$\mathbf{Y}_h$ is the hth exogenous latent score and $\mathbf{Y}_l$ is the lth endogenous latent score; the symbol $\otimes$ indicates here the element-wise product of two matrices, while $*$ indicates the adjacent latent scores matrix, i.e., the set of latent scores that are related to the $\mathbf{Y}_h$ or $\mathbf{Y}_l$. The PLS-SEM-KM algorithm is a development of the Wold's original algorithm used to the PLS-SEM estimate in Lohmoller (1989). As you can see from the step 7 of the algorithm (i.e., in the loadings estimation), the method is performed for both reflective measurement models and formative measurement models. $\mathbf{U}$ matrix is optimized row by row solving an assignment problem through the objective function in the step 8 of the algorithm.

Therefore, the algorithm produces a matrix $\mathbf{U}$ of the segments assignment and a matrix $\mathbf{C}$ of centroids with a unique common measurement and structural model coefficients. However, researchers that wish determining segment specific measurement and structural model coefficients can apply group-specific PLS-SEM analysis. The unique measurement and structural model coefficients are interpreted as a consensus of the segment-specific coefficients.

The proposed methodology shows some important advantages with respect to the other proposed approaches for both cluster analysis and composite indicator construction: firstly, it is a simultaneous approach that identifies the best homogenous partition of the objects represented by the best causal relationships among latent and observed variables. Then, unlike a sequential approach, the identified partition is dependent on the prespecified composite-based (i.e., causal) relationships; moreover, distributional assumptions are not requested for the PLS-SEM-KM application (Fordellone and Vichi 2018), this because it uses a partial least squares (PLS) methodology that, unlike the covariance structure approach (CSA), is insensitive to the data distributional assumptions.

486489_1_En_28_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:4/11/2020 Pages: xxx Layout: T1-Standard

4 From Data to Results for IICI

The data used for the construction of the immigrant integration composite indicator (IICI) construction derive from the eighth iteration of the survey for ESS. Until now are available 18 of the 24 countries, which undertook fieldwork in 2016. Table 1 shows the principal topics included in ESS data.

Table 1 Topics and items of ESS survey

Items	Topic
Core A1–A6	Media use; internet use; social trust
Core B1–B43	Politics, including political interest, trust, electoral and other forms of participation, party allegiance, sociopolitical orientations immigration
Core C1–C44	Subjective well-being, social exclusion, crime, religion, perceived discrimination, national and ethnic identity, test questions (Sect. I), refugees
Core D1–D32	Climate change and energy, including attitudes, perceptions module and policy preferences
Core E1–E42	Welfare, including attitudes toward welfare provision, size of module claimant groups, attitudes toward service delivery and likely future dependence on welfare, vote intention in EU referendum
Core F1–F61	Sociodemographic profile, including household composition, sex, age, marital status, type of area, education and occupation, partner, parents, union membership, income and ancestry
Core Section H	Human values scale
Core Section l	Test questions

Table 2 Path coefficients estimated by PLS-SEM-KM

	Estimate	Std. error	t-value	Pr(> $\|t\|$)
(Intercept)	0.149	0.029	5.148	0.000
Politics	0.875	0.016	56.445	0.000
Economics	−0.215	0.029	−7.420	0.000
Social	0.211	0.022	9.385	0.000
Cultural	−0.383	0.022	−17.524	0.000
Crime	0.204	0.030	6.827	0.000
Religion	−0.185	0.019	−9.687	0.000
Structural	−0.046	0.012	−3.736	0.000
Household	−0.154	0.016	−9.743	0.000
Employment	0.216	0.013	16.211	0.000

F-statistic: 3756 on 9 and 264 DF (p-value = 0.000) $R^2 = 0.8823$, $R^2_{adj} = 0.882$

After data aggregation, our data set is composed of 274 regions of the 18 countries and 64 Likert scale variables, defining the following 9 dimensions, i.e., *politics* with 19 MVs, *economics* with 2 MVs, *social* with 2 MVs, *cultural* with 2 MVs, *crime* with 2 MVs, *religion* with 2 MVs, *structural* with 11 MVs, *household* with 9 MVs, and *employment* with 15 MVs.

The application of the PLS-SEM-KM model has detected a number of clusters $K = 5$ obtaining the estimates of the path coefficients shown in Table 2.

The estimates reported in Table 2 show an overall good performance of the model both in terms of path coefficients (i.e., all the estimated coefficients are statistically significant) and in terms of explained deviance (i.e., high R^2 values). Observing the single coefficients, we can see that more remarkable significant effect on IICI is given by the *politics* (0.875) and *cultural* (−0.383) constructs. In contrast, a very low impact on IICI is given by the *structural* dimension (−0.046), which includes important demographic features of the respondents, followed by *household* (−0.154) and *religion* (−0.185) constructs together with *economics* (−0.215), *social* (0.211), *crime perception* (0.204), and *employment* (0.216) dimensions.

Figure 3 shows the loading estimates obtained for each latent dimension.

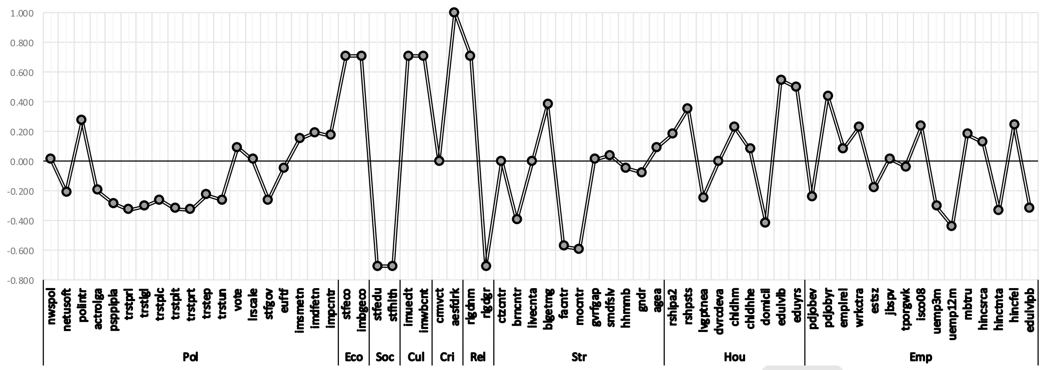

Fig. 3 Loadings estimates for each latent dimension

Figure 4 shows the cluster distributions on the 10 estimated latent scores (i.e., including also the composite indicator), while in Fig. 5 a geographical representation of the obtained clusters is shown. Note that the size of the 5 clusters comprising the identified partitions are 52, 46, 74, 39, and 63, respectively. In the representation of the loadings, we have used the official labels of the 64 MVs which we have selected for the definition of latent dimensions.[1]

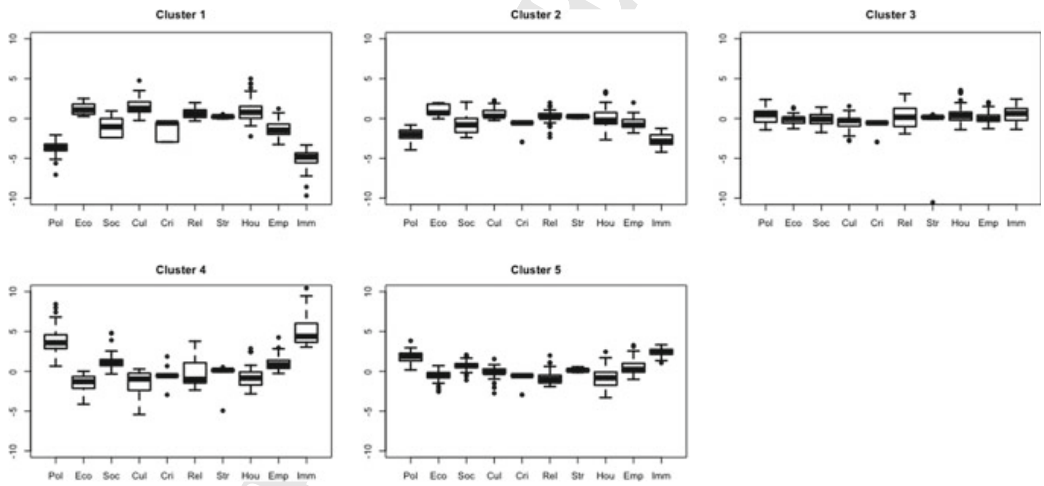

Fig. 4 Clusters distribution on the all latent constructs

[1]For more details on the selected MVs, you can see the official ESS website: http://www.europeansocialsurvey.org/about/news/essnews0038.html..

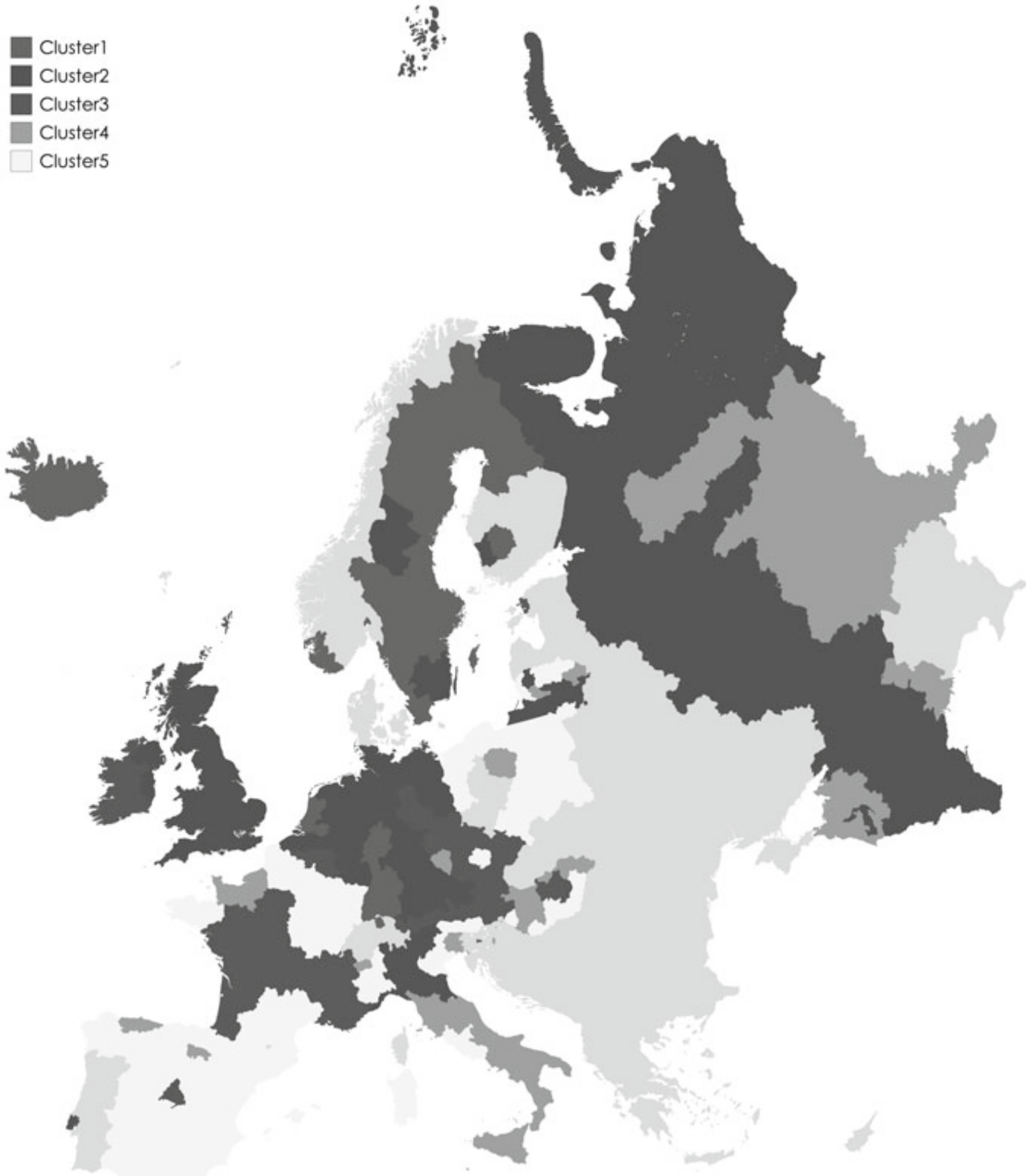

Fig. 5 Geographical representation of the clusters

From the cluster distributions in Fig. 4, we can note that the first and the fourth cluster are very discriminant of the immigrant integration level, because the IICI values are very low and very high, respectively. Moreover, we can also note that the political dimension has a very hard impact on the immigrant integration. So, the lower the IICI, the lower the political factor level is in the cluster 1. On the contrary, in the cluster 4 a high level of political dimension is related to a high level of the composite indicator.

The results obtained by employing the PLS-SEM-KM method show a reliable classification structure of 274 regions of 18 European countries where the level of immigrant integration is different for 4 clusters of regions.

The more discriminant ability of the 9 exogenous latent variables and also of the composite indicator (IICI) allows efficiently mapping overall the more northern regions in a cluster where the lowest values of the indicator represent a low level of immigrant integration while in the cluster 4, most of the southern and eastern regions are more discriminated on the basis of high values of the composite indicator for a higher level of immigrant integration. The effect of the political participation dimension is affecting the most both the classification of regions and the composite indicator building, assuming the same trend, thus, in each cluster: the higher/lower the level of political participation, the higher/lower the level of immigrant integration is in the European regions.

5 Conclusive Remarks

This work, employing the PLS-SEM methodology where SEM is estimated by PLS, is focused on the building of an integration composite indicator (IICI), in Europe. With this aim, we use a simultaneous PLS-SEM-KM approach introduced by Fordellone and Vichi (2018) (PLS-SEM-KM).

The results show a good performance of the global model, especially for the immigrant integration profile. Moreover, the conjoined clustering model defines partitions that add relevant information on the countries' features involved in the immigrant integration issue.

In our opinion, by employing composite indicators to measure a complex phenomenon like immigrant integration, an international comparative approach can help to focus and target nationally and locally immigrant integration policies.

References

Ager, A., & Eyber, C. (2002). Indicators of integration: A review of indicators of refugee integration. Report to the Home Office on Behalf of Michael Bell Associates.

Entzinger, H. (2000). The dynamics of integration policies: A multidimensional model. In R. Koopmans & P. Statham (Eds.), *Challenging immigration and ethnic relations politics* (pp. 97–118). Oxford: University Press.

Entzinger, H., & Biezeveld, R. (2003). *Benchmarking in immigrant integration*. Rotterdam: European Research Center on Migration and Ethnic Relations.

Cesareo, V., & Blangiardo, G. (2009). *Indici di integrazione*. Milano, IT: FrancoAngeli.

De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, & B. Burtschy (Eds.), *New approaches in classification and data analysis* (pp. 212–219). Berlin, Heidelberg: Springer.

Di Bartolomeo, A., Kalantaryan, S., & Bonfanti, S. (2015). Measuring integration of migrants: A multivariate approach. INTERACT RR 2015/01, Robert Schuman Centre for Advanced Studies, San Domenico di Fiesole, Firenze, IT: European University Institute.

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, *38*(2), 269–277.

486489_1_En_28_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:4/11/2020 Pages: xxx Layout: T1-Standard

Fordellone, M., & Vichi, M. (2018). Structural equation modeling and simultaneous clustering through the partial least squares algorithm. arXiv:1810.07677.

Lohmoller, J. B. (1989). *Latent variable path analysis with partial least squares*. Heidelberg: Physica.

OECD. (2008). *Handbook on constructing composite indicators*. Methodology and user guide. Paris, FR: OECD Publishing. http://www.oecd.org/std/42495745.pdf.

OECD. (2018). *The integration of migrants in OECD regions: A first assessment*. Regional Development Working Papers 2018/01. Paris, FR: OECD Publishing. https://dx.doi.org/10.1787/fb089d9a-en.

Strang, A., Ager, A., & Brien, O. O. (2003). Indicators of integration: The experience of integration. Report to the Home Office on behalf of Michael Bell Associates.

Tenenhaus, M., Vinzi, E. V., Chatelin, Y. M., & Lauro, N. C. (2005). PLS path modeling. *Computational Statistics and Data Analysis*, *48*(1), 159–205.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, *63*(2), 411–423.

Vermeulen, H. (2004). Models and modes of immigrant integration ...And where does Southern Europe Fit? In C. Inglessi, A. Lyberaki, H. Vermeulen & G. J. V. Wijngaarden (Eds.), *Immigration and integration in northern versus southern Europe* (pp. 11–26). Athens, EL: Netherlands Institute at Athens.

Vinzi, E. V., Trinchera, L., & Amato, S. (2010). PLS path modeling: From foundations to recent developments and open issues for model assessment and improvement. In V. Esposito Vinzi, W. Chin, J. Henseler & H. Wang (Eds.), *Handbook of partial least squares* (pp. 47–82). Berlin, Heidelberg: Springer.