VOLUME LXXII – N. 1 GENNAIO-MARZO 2018

RIVISTA ITALIANA DI ECONOMIA DEMOGRAFIA E STATISTICA

Sede Legale: C/O Studio Associato Cadoni, Via Ravenna n.34 – 00161 ROMA.
sieds.new@gmail.com, rivista.sieds@gmail.com

SIEDS
SOCIETÀ ITALIANA
DI ECONOMIA DEMOGRAFIA E STATISTICA

SEDE LEGALE:
C/O Studio Associato Cadoni, Via Ravenna n.34 – 00161 ROMA

sieds.new@gmail.com

rivista.sieds@gmail.com

VOLUME FUORI COMMERCIO – DISTRIBUITO GRATUITAMENTE AI SOCI

INDICE

Mario Fordellone, Venera Tomaselli, Maurizio Vichi
From tandem to simultaneous dimensionality reduction and clustering of tourism data 5

Leonardo Salvatore Alaimo
Demographic and socio-economic factors influencing the Brexit vote 17

Giovanna Da Molin, Maddalena Lenny Napoli, Elita Anna Sabella, Arjeta Veshi
Lifestyles of university students in Albania 29

Elisabetta Bilotta, Emanuela Trinca
Capitale umano e multinazionali estere 41

Valentina Ferri, Dario Guarascio, Andrea Ricci
Assetto proprietario e manageriale: evidenze empiriche su elementi di contesto ed organizzativi delle imprese italiane 53

Paola Naddeo, Stefania Cardinaleschi
Wage gaps by collective bargaining and firm size in Italy 65

Carmela Cuomo, Carlo Cusatelli, Massimiliano Giacalone
The possession of narcotics for personal use in the province of Salerno from 2010 to 2016 77

Vincenzo Marinello, Guglielmo L.M. Dinicolò
Drivers of brain drain phenomenon: possible association between macroeconomic variables in the international framework 87

Vincenzo Marinello, Guglielmo L.M. Dinicolò, Mariano Cavataio
Transparency and countering of corruption in public administration: strategies and policy concerning public contracting 99

FROM TANDEM TO SIMULTANEOUS DIMENSIONALITY REDUCTION AND CLUSTERING OF TOURISM DATA[1]

Mario Fordellone, Venera Tomaselli, Maurizio Vichi

1. Simultaneous *vs* sequential methods

In the market segmentation studies, but also in economic and social phenomena, many variables are often observed and a huge amount of objects are analysed in large data sets. The suitable statistical data analysis for modelling complex phenomena requires reducing both variables and units in order to extract the relevant information (Knowledge) from the data. From one end, the aim is to identify significant relationships, via dimensionality reduction either: of categorical variables by multiple correspondence analysis (MCA); or, of metric variables by principal component analysis (PCA) or factor analysis (FA). The reduction provide the measurement of hidden concepts, i.e., latent variables. From the other end, clustering produces a reduced sets of homogeneous objects described by the reduced set of latent variables.

The traditional sequential data reduction and clustering procedure is often not reliable due to some problems. Firstly, a reduced set of factors extracted from many variables could remove relevant information about the subsequent clustering data structure (De Sarbo *et al.*, 1994). Furthermore, noise could be introduced from those variables not so useful for clustering objects (Rocci, Gattone & Vichi, 2011). Thus, the sequential or *tandem analysis* (TA) could not clearly define factors and properly describe the clustering structure.

Focusing on factorial methods, the interpretation of the factors is often very hard. Factors, indeed, are defined as linear combinations of all variables, and generally have loadings usually different from zero; while, only few variables are effectively relevant for each factor. With the aim to simplify, alternative procedures have been proposed that combine the search for a reduced set of factors, such as multidimensional scaling or unfolding analysis, and clustering methods (Heiser, 1993; De Soete and Heiser, 1993, De Soete and Carroll, 1994, Bolton and Krzanowski, 2003).

[1] Invited paper to the 54th SIDES Scientific Meeting – Catania 2017.

As a good alternative to TA in the case of metric variables, a Factorial *K*-Means (FKM) model combining *K*-means cluster analysis with PCA was proposed by Vichi and Kiers (2001). The method selects the most relevant variables achieving factors that best identify the clustering structure in order to find in the data the best subspace that best represents this structure.

In the case of observed categorical variables, a similar methodology was proposed by Fordellone and Vichi (2016), named Multiple Correspondence *K*-Means (MCKM), for simultaneous dimension reduction and clustering. By means of an alternative least squares algorithm, in this innovative simultaneous definition of factors and clusters on the observed data, the minimization of a single objective function allows for identifying the best partition of *N* objects depicted by the best orthogonal linear combination of variables.

However, in the last years, Structural Equation Modeling (SEM) has become one of the reference statistical methodologies in the analysis of complex phenomena, where there are statistical relationships between variables directly observable (manifest variables) and non-directly observable (latent variable). Then, SEM are often used to assess unobservable hidden constructs (latent variables) by means of observed variables and to evaluate the relations between latent constructs. Covariance Structure Approach (CSA) (Jöreskog, 1978) and Partial Least Squares (PLS) (Lohmöller, 1989) are the two alternative statistical techniques for estimating such models. However, PLS is considered preferable to CSA in three specific cases: (*i*) when the sample size is small, (*ii*) when the data to be analyzed is not multi-normal as required by CSA, and (*iii*) when the complexity of the model to be estimated may lead to improper or non-convergent results (Bagozzi and Yi, 1994; Squillacciotti, 2010).

To detect homogenous tourism profiles, the sequential procedure is often employed in the tourism market segmentation studies to find clustering structure with a reduced set of factors (Pina & Delfa, 2005; Dolnicar, 2005; Asero *et al.*, 2013). In the following sections of the present paper, tourism survey data are processed with the aim to compare the findings by sequential and simultaneous methods. A new methodology for simultaneous non-hierarchical clustering and PLS-modeling was recently proposed and named Partial Least Squares K-Means (PLS-KM) (Fordellone and Vichi 2017). The model is based on the simultaneous optimization of PLS-SEM and Reduced K-Means (De Soete and Carroll, 1994), where centroids are laying 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 SEM pre-specified model.

In our hypothesis, TA shows some fallacies to correctly classify units and synthesize the relationships among observed categorical or metric variables. Specifically, the loss function of the TA is only imprecisely estimated by the

sequential procedure. Employing the simultaneous procedure, instead, the loss function is optimized. Hence, well-characterized dimensions are detectable and more homogenous and well-separate clusters identifiable, making even easier the interpretation of the achieved results.

2. Simultaneous procedure

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 $\mathbf{\Lambda} = [\mathbf{\Lambda}_{\mathrm{H}}, \mathbf{\Lambda}_{\mathrm{L}}]$[2], and the errors matrices $\mathbf{E}$, $\mathbf{Z}$, $\mathbf{D}$, the Partial Least Squares *K*-Means model can be written as follows (Fordellone and Vichi, 2017):

$$\begin{aligned} \mathbf{H} &= \mathbf{H}\mathbf{B}^{\mathrm{T}} + \mathbf{\Xi}\mathbf{\Gamma}^{\mathrm{T}} + \mathbf{Z}, \\ \mathbf{X} &= \mathbf{\Xi}\mathbf{\Lambda}_{\mathrm{H}}^{\mathrm{T}} + \mathbf{H}\mathbf{\Lambda}_{\mathrm{L}}^{\mathrm{T}} + \mathbf{E}, \\ \mathbf{X} &= \mathbf{U}\mathbf{C}\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}} = \mathbf{U}\mathbf{C}\mathbf{\Lambda}_{\mathrm{H}}\mathbf{\Lambda}_{\mathrm{H}}^{\mathrm{T}} + \mathbf{U}\mathbf{C}\mathbf{\Lambda}_{\mathrm{L}}\mathbf{\Lambda}_{\mathrm{L}}^{\mathrm{T}} + \mathbf{D}, \end{aligned} \tag{1}$$

under constraints: *(i)* $\mathbf{\Lambda}^{\mathrm{T}}\mathbf{\Lambda} = \mathbf{I}$; and *(ii)* $\mathbf{U} \in \{0,1\}$, $\mathbf{U}\mathbf{1}_K = \mathbf{1}_n$. Where, $\mathbf{H}$ is the $n \times L$ matrix of the endogenous LVs with generic element $\eta_{i,l}$, $\mathbf{\Xi}$ be the $n \times H$ matrix of the exogenous LVs with generic element $\xi_{i,h}$, $\mathbf{B}$ is the $L \times L$ matrix of the path coefficients $\beta_{l,l}$ associated to the endogenous latent variables, $\mathbf{\Gamma}$ is the $L \times H$ matrix of the path coefficients $\gamma_{l,h}$ associated to the exogenous latent variables, $\mathbf{\Lambda}_{\mathrm{H}}$ is the $J \times H$ loadings matrix of the exogenous latent constructs with generic element $\lambda_{j,h}$, and $\mathbf{\Lambda}_{\mathrm{L}}$ is the $J \times L$ loadings matrix of the endogenous latent constructs with generic element $\lambda_{j,l}$.

Thus, the PLS-KM model includes the PLS and the clustering equations. In fact, the third set of equations is the model of Reduced K-means (De Soete and Carroll, 1994). The simultaneous estimation of the three sets of equations will produce the estimation of the pre-specified SEM describing relations among variables and the corresponding best partitioning of units.

When applying PLS-KM, the number of groups is unknown and the identification of an appropriate number of *K* clusters is not straightforward. Then, often you need to rely on some statistical criterion. In this paper we use the *gap method* proposed by Tibshirani at al. (2001) for estimating the number of clusters, i.e., a *pseudo*-F designed to be applicable to virtually any clustering method.

In the preliminary step of the PLS-KM algorithm, the estimation of the PLS-SEM over the entire dataset is carried out; subsequently, the number of the *K* classes is obtained according to the maximum level of the *pseudo*-F function

[2] Note that $H+L=P$

computed on the estimated latent scores. Then, once chosen the number of clusters, the PLS-KM algorithm optimize the following overall objective function:

$$\text{argmin}_{\mathbf{UC\Lambda}} \left\| \mathbf{X} - \mathbf{UC\Lambda\Lambda}^{\mathrm{T}} \right\|^{\mathbf{2}} \qquad (2)$$

Note that because the constraints on **U**, the method can be expected to be rather sensitive to local optima. For this reasons, it is recommended the use of some randomly started runs to find the best solution.

3. Data and findings in sequential reduction and clustering

The methods for dimensionality reduction and clustering are employed to analyse incoming tourism data collected by a survey based on a sampling design known as *Time-Location Sampling* (Kalton, 2003; 2009). This methodology suits the mobile aspect of tourists and the absence of a complete list of sampling units for an unknown population. Tourists are often not immediately identifiable at destination access points because they mingle with residents or other types of travellers.

On the basis of official data on tourist flows, the survey was carried out at the main Sicilian ports and airports in spring, summer and early autumn when more than 80% of tourists visit the Island (Asero *et al.*, 2013). Data were collected by a questionnaire divided into different sections based on main tourism related dimensions, specifically: tourist nationality, first-time *vs.* repeated visitors, age, reason/motivation for vacation, family/friends or alone on holiday, information tool, travel arrangements, type of accommodation, type of holiday, total expenditure and specific items, expectations, and level of satisfaction. In addition, a special section was dedicated to mobility among different destinations within Sicily. Tourists were asked to indicate all the locations visited where they had spent at least one night; specifying the number of nights spent and the type of accommodation.

All of these variables were derived from a thorough review of tourism segmentation literature (Kozak, 2002; Pina & Delfa, 2005; Dolnicar, 2005; Martínez-Garcia & Raya, 2008). In the segmentation studies, the variables can assume the role of bases if they directly generate the process of classification into groups of units under observation, or behave as descriptors when they come to the interpretation of the segments (Brasini *et al*., 2002). The variables of market segmentation, as a consequence, should allow for significant differentiation of the members of the various segments in terms of expectations, attitudes and consumer behaviour. Furthermore, they must support the construction of explanatory

hypotheses about the factors affecting the observed phenomena (Idili and Siliprandi, 2005).

The analysis of the above survey data focused on tourists autonomously visiting the destinations in order to identify tourist profiles through motivational and behavioural variables leading to choose a destination. Thus, a number of 3233 self-organized tourists were selected from the original sample (3935).

On this sample of tourism data, the sequential approach for dimensionality reduction and clustering was performed. Specifically, the selection of useful features was performed step-by-step by the integrated multidimensional analysis strategy aimed to synthesize the variables into analytical dimensions for the subsequent clustering procedure (CA). For categorical variables, Multiple Correspondence Analysis (MCA) was employed and Principal Component Analysis (PCA) for the metric variables. Object scores by MCA and factor scores by PCA were used to define market segments by CA. The clustering was a non-hierarchical grouping procedure involving 'centres on the move'. It performed directly a partition in stable and homogeneous groups without progressive aggregation of pairs or groups of objects. The aggregation method, related to single links or to the closest units, was based on the smallest distances among contiguous units into a single group. The identification of stable groups was performed using the criterion of Ward known as 'minimum variance' which, after decomposition of the variance in *between* and *within*, aggregates the elements that together make up the cluster with the least internal variance (Fabbris, 1997). The agglomerative hierarchical method was used, proceeding *via* the progressive aggregation of groups into 'nodes', in order to build a tree whose terminal elements were precisely the stable groups previously identified.

Following this procedure, two clusters were obtained (Asero *et al.*, 2013): in the first, younger Italians were recovered, who spent less time on holiday, less interested in the 'beach' product, travelling alone or with friends, preferring free accommodation or B&B; in the second cluster, mainly older foreigners, for longer periods, preferring the beach and almost exclusively with family, with a *per capita* expenditure of around double in comparison with the first cluster, overall for internal trips and restaurants as well as for hotels and residence as accommodation.

4. Main results by simultaneous method

Dataset consists in 9 variables (8 categorical and 1 continuous) that represent the answers of 3233 self-organized tourists. The set of categorical variables have been re-scaled in an indicator matrix (called also *complete disjunctive table*) centred and normalized both by rows and by columns. In particular, $J^{1/2}\mathbf{JDL}^{1/2}$ is

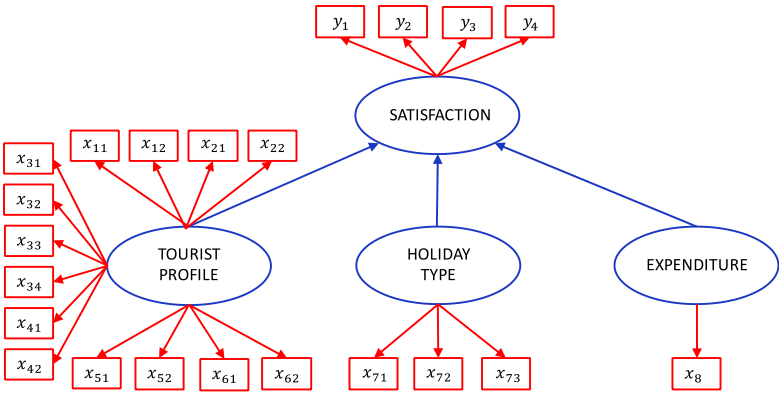
the centred categorical data matrix corresponding to the J qualitative variables, with the binary block matrix $\mathbf{D} = [\mathbf{D}_1, \dots, \mathbf{D}_\mathrm{J}]$ formed by J indicator binary matrices $\mathbf{D}_\mathrm{J}$ with elements $d_{ijm} = 1$, if the i^{th} individuals has assumed category m for variable j; $d_{ijm} = 0$, otherwise; $\mathbf{L} = diag(\mathbf{D}^\mathrm{T}\mathbf{1}_\mathrm{n})$; $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\mathrm{T}$ is the idempotent centering matrix with $\mathbf{1}_n$ the n-dimensional vector of unitary elements. The details on the original variables are shown in Table 1.

Table 1 – *Description on the variables included in the model*

Variable	Description
Nationality (x_1)	x_{11}: *Italy* x_{12}: *Foreign Country*
Gender (x_2)	x_{21}: *Male* x_{22}: *Female*
Age classes (x_3)	x_{31}: *<25 years* x_{32}: *25-44 years* x_{33}: *45-64 years* x_{34}: *>64 years*
Travel-friends 1 (x_4)	x_{41}: *Alone* x_{42}: *With someone*
Travel-friends 2 (x_5)	x_{51}: *Alone* x_{52}: *With family*
Travel-friends 3 (x_6)	x_{61}: *Alone* x_{62}: *With friends*
Type of holiday (x_7)	x_{71}: *Beach only* x_{72}: *Partial beach* x_{73}: *No beach at all*
Expenditure (x_8)	x_8: *Cost of the holiday* (continuous)
Satisfaction (y)	y_1: *High satisfaction* y_2: *Medium-high satisfaction* y_3: *Medium-low satisfaction* y_4: *Low satisfaction*

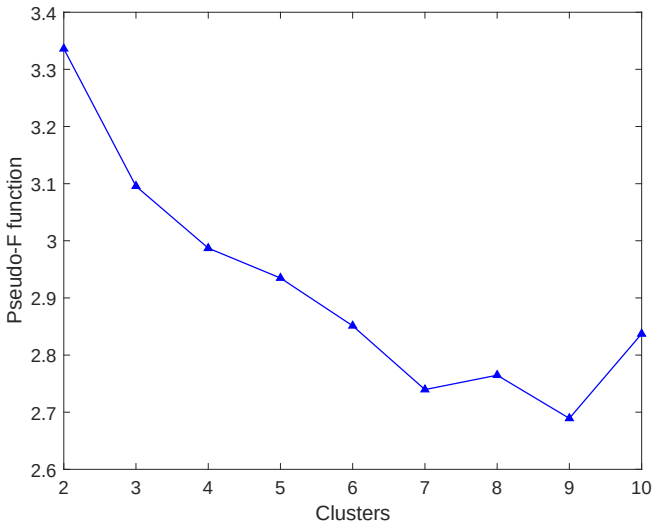
The causal structure used for modelling the relationships among latent constructs and observed variables, is represented by the path-diagram shown in Figure 1.

Figure 1 – *Path-diagram of the tourism model*



In order to apply the Partial Least Squares *K*-Means, the first step to consider is the choice of the number of clusters *K*. The optimal number of classes is obtained according to the maximum level of the *pseudo*-F function computed on the latent scores firstly estimated by PLS approach (Tenenhaus et al., 2004). The gap method proposed by Tibshirani et al. (2001) suggests $K = 2$, with the corresponding value of *pseudo*-F equal to 3.336 as shown in Figure 2.

Figure 2 – *Pseudo-F function obtained via gap method on the PLS scores from 2 to 10 clusters*



Once the number of clusters has been chosen, PLS-KM has been applied fixing a number of random starts equal to 15. The latent hyperplane obtained by the model has an Average Variance Explained (AVE) for factors equal to 54%.
The results obtained on the structural model and measurement models are shown in Table 2 and Table 3, respectively.

Table 2 – *Structural model estimated by PLS-KM*

	Tourist profile	Holiday type	Expenditure	Satisfaction
Tourist profile	0	0	0	**0.482**
Holiday type	0	0	0	**0.342**
Expenditure	0	0	0	**0.175**
Satisfaction	0	0	0	0

Table 3 – *Measurement models estimated by PLS-KM*

	Tourist profile	Holiday type	Expenditure	Satisfaction
x_{11}	**0.404**	0.000	0.000	0.000
x_{12}	**-0.404**	0.000	0.000	0.000
x_{21}	**-0.047**	0.000	0.000	0.000
x_{22}	**0.047**	0.000	0.000	0.000
x_{31}	**0.109**	0.000	0.000	0.000
x_{32}	**-0.403**	0.000	0.000	0.000
x_{33}	**0.326**	0.000	0.000	0.000
x_{34}	**0.114**	0.000	0.000	0.000
x_{41}	**-0.179**	0.000	0.000	0.000
x_{42}	**0.179**	0.000	0.000	0.000
x_{51}	**0.323**	0.000	0.000	0.000
x_{52}	**-0.323**	0.000	0.000	0.000
x_{61}	**-0.227**	0.000	0.000	0.000
x_{62}	**0.227**	0.000	0.000	0.000
x_{71}	0.000	**-0.646**	0.000	0.000
x_{72}	0.000	**-0.107**	0.000	0.000
x_{73}	0.000	**0.756**	0.000	0.000
x_{8}	0.000	0.000	**1.000**	0.000
y_{1}	0.000	0.000	0.000	**-0.715**
y_{2}	0.000	0.000	0.000	**0.667**
y_{3}	0.000	0.000	0.000	**0.173**
y_{4}	0.000	0.000	0.000	**0.122**

From the results obtained by structural and measurement models, is interesting to see that the *satisfaction* of the tourists is most influenced by *tourist profile* dimension (path coefficient equal to 0.48), with respect to the *holiday type* dimension (path coefficient equal to 0.34) and the *expenditure* dimension (path coefficient equal to 0.17). In Table 3 is possible see the intensity and the direction of the relationships among the four latent constructs and the single observed variables.

In Table 4 are shown the summary statistic on the normalized latent scores observed on the two clusters.

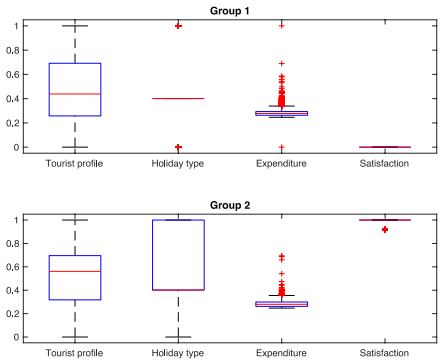
Table 4 – *Summary statistics of the latent scores observed on two groups*

	Group 1				Group 2			
	Tourist profile	Holyday Type	Expenditure	Satisfaction	Tourist profile	Holyday Type	Expenditure	Satisfaction
Min	0.000	0.000	0.000	0.000	0.000	0.000	0.247	0.918
Q1	0.258	0.401	0.263	0.000	0.318	0.401	0.262	1.000
Median	0.438	0.401	0.278	0.000	0.562	0.401	0.280	1.000
Mean	0.458	0.443	0.284	0.000	0.530	0.507	0.288	0.992
Q3	0.692	0.401	0.294	0.000	0.696	1.000	0.299	1.000
Max	1.000	1.000	1.000	0.000	1.000	1.000	0.694	1.000
	N = 2614				***N = 619***			

From the summary statistics, it is easy to note that in two observed clusters (formed by 2614 and 619 observations, respectively) represent the unsatisfied and the very satisfied tourists, respectively.
In fact, in the *tourist profile* dimension, no relevant differences are observed among clusters. In particular, in the first group (satisfied tourist) the 44% of tourists are Italian tourists and the 56% are foreign tourists, half of tourists are men (women), and most of them (about 88%) has age included between 25 and 64 years old. In the second group (no satisfied tourists) the 36% of tourists are Italian tourists and the 64% are foreign tourists, also in this group half of tourists are men (women), while about the 90% of tourists have age included between 25 and 64 years old.
As regards the cost of the holiday, it is possible to see that the second group has a *per capita* expenditure bigger than first group (mean of expenditure equal to 70.24 and 63.54 respectively).

Finally, in Figure 3 the box-plots of the normalized latent scores observed on the two groups are shown

Figure 3 – *Box-plots of the latent scores observed on two groups*



5 Conclusion

Tandem Analysis (TA) is a well-known sequential procedure for clustering and dimension reduction. This methodology is frequently used in applications for both quantitative data and qualitative/categorical data.

However, in some case this approach has several limitations. In particular, TA can fail to find the correct clustering structure of data because the noise variables could mask it. As alternative to TA there are the simultaneous approaches, e.g. Factorial *K*-Means (FKM) (Vichi and Kiers, 2001) in the case of metric variables, and Multiple Correspondence *K*-Means (MCKM) (Fordellone and Vichi, 2016) in the case of categorical variables. Both these approaches apply simultaneously a dimension reduction model (PCA and MCA, respectively) and a clustering model (*K*-Means).

In this work a new model recently proposed in (Fordellone and Vichi 2017) and named Partial Least Squares *K*-Means (PLS-KM) has been applied on mixed categorical and continuous variables. This methodology combines the Structural Equation Model (SEM) estimated via PLS algorithm and the *K*-Means model.

A comparison between TA and PLS-KM has been carried in order to analyse the incoming tourism phenomenon. Both approaches provide a number of clusters $K = 2$, but the results obtained show many differences in terms of tourist profile. The simultaneous procedure shows more homogenous and well-separate clusters than the sequential approach. Moreover, PLS-KM in addition of clustering model provides a model to study relationships among variables, where it is possible to analyse the tourist satisfaction as influenced by other aspects of the holiday.

References

ASERO V., D'AGATA R., TOMASELLI V. 2013. Analysing Tourism Demand by Motivations. In: Oliveri A. M., De Cantis S. (a cura di), *Analysing Local Tourism. A Statistical Perspective*. Maidenhead (UK): McGraw-Hill Education, pp. 297-307.

BAGOZZI R. P., YI Y. 1994. Advanced topics in structural equation models. *Advanced methods of marketing research*, 151.

BOLTON R. J., KRZANOWSKI W. J. 2003. Projection pursuit clustering for exploratory data analysis. *Journal of Computational and Graphical Statistics*, Vol. 12, No. 1, pp. 121-142.

BRASINI S., FREO M., TASSINARI F., TASSINARI G. 2002. *Statistica aziendale e analisi di mercato,* Bologna: Il Mulino.

DE SARBO W., JEDIDI K., COOL K., SCHENDEL D. 1991. Simultaneous multidimensional unfolding and cluster analysis: An investigation of strategic groups. *Marketing Letters*, Vol. 2, No. 2, pp. 129-146.

DE SOETE G., HEISER W. J. 1993. A latent class unfolding model for analyzing single stimulus preference ratings. *Psychometrika*, Vol. 58, No. 4, pp. 545-565.

DE SOETE G., CARROLL J. D.. 1994. K-means clustering in a low-dimensional Euclidean space. In: *New approaches in Classification and Data Analysis*. Berlin Heidelberg: Springer, pp. 212-219.

DOLNICAR S. 2005. Empirical market segmentation: What you see is what you get. In THEOBOLD W. (Ed.), *Global Tourism the Next Decade*, Oxford: Butterworth-Heinemann, 3rd, pp. 309–325.

FORDELLONE M., VICHI M. 2016. Multiple Correspondence K-Means: a new approach for dimension reduction and clustering for categorical data. In: Mola CONVERSANO F. C., Vichi M. (Eds), *Classification, (Big) Data Analysis and Statistical Learning* (series: Studies in Classification, Data Analysis, and Knowledge Organization), Berlin Heidelberg: Springer.

FORDELLONE M., VICHI M. 2017. Partial Least Squares Path Modeling and simultaneous clustering [Submitted].

HEISER W. J.. 1993. Clustering in low-dimensional space. In: *Information and Classification*. Berlin Heidelberg: Springer, pp. 162-173.

IDILl L., SILIPRANDI L. 2005. *Il marketing degli operatori turistici.* Milano: FrancoAngeli.

JÖRESKOG K. 1978. Structural analysis of covariance and correlation matrices. *Psychometrika*, Vol. 43, No. 4, pp. 443-477.

KALTON G. 2003. Practical methods for sampling rare and mobile populations. *Statistics in Transition*, Vol. 6, No. 4, pp. 491–501.

KALTON G. 2009. Designs for surveys over time. In PFEFFERMANN D., RAO C. R. (Ed.) *Handbook of Statistics.* Vol. 29A *Sample Surveys: Design, Methods and Applications.* North-Holland: Elsevier, pp. 89-108.

KOZAK M. 2002. Comparative analysis of tourist motivations by nationality and destinations. *Tourism Management*, Vol. 23 No. 3, pp. 221-232.

LÖHMOELLER J. B. 1989. *Latent Variable Path Analysis with Partial Least Squares*. Heidelberg: Physica.

MARTINEZ-GARCIA E., RAYA J. M. 2008. Length of stay for low-cost tourism. *Tourism Management*, Vol. 29, No. 6, pp. 1064-1075.

PINA I. P. A., DELFA M. T. D.. 2005. Rural tourism demand by type of accommodation. *Tourism Management*, Vol. 26, No. 6, pp. 951-959.

ROCCI R., GATTONE S. A., VICHI M.. 2011. A New dimension reduction method: Factor discriminant K-means. *Journal of Classification*, Vol. 28, No. 2, pp. 210-226.

SQUILLACCIOTTI S.. 2010. Prediction oriented classification in PLS path modelling. *Handbook of Partial Least Squares*, Berlin Heidelberg: Springer, pp. 219-233.

TENENHAUS M., VINZI V. E., CHATELIN Y. M., LAURO C. 2005. PLS path modelling. *Computational Statistics & Data Analysis*, Vol. 48, No. 1, pp. 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 (Statistical Methodology)*, Vol. 63, No. 2, pp. 411-423.

VICHI M., KIERS H. A. 2001. Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, Vol. 37, No. 1, pp. 49-64.

SUMMARY

From tandem to simultaneous dimensionality reduction and clustering of tourism data

The study of tourist demand is a critical component of a successful destination management strategy. In order to define tourist segments, many factors play an important role in the decision-making process. Tourism motivations are often used as segmentation bases of tourism market since they can affect the choices about travel destination, type of holiday and consumer behaviour. A tourist destination offers many experiences and products, which appeal different market segments. This paper aims to identify *a posteriori* segments of tourism demand by means of multidimensional approach employing a simultaneous factorial dimensionality reduction and clustering method. On the basis of results, tourists are classified in two clusters in order to understand the relationship between motivations and consumer behaviour. In particular, the two observed clusters represent the very satisfied tourists and the tourists unsatisfied at different level, respectively. Moreover, in terms of cost of the holiday, the first group has a per capita expenditure bigger than second group.

Mario FORDELLONE, Department of Statistical Sciences, University "La Sapienza", Roma, mario.fordellone@uniroma1.it

Venera TOMASELLI, Department of Political and Social Sciences, University of Catania, tomavene@unict.it.

Maurizio VICHI, Department of Statistical Sciences, University "La Sapienza", Roma, Maurizio.Vichi@uniroma1.it.

DICHIARAZIONE SOSTITUTIVA DI ATTO DI NOTORIETA'
(sull'attribuzione della responsabilità dei singoli autori di lavori congiunti)
(Artt. 19 e 47 del D.P.R. 28.12.2000, n. 445)

La sottoscritta TOMASELLI Venera nata a Catania l'1/9/1961, residente a Pedara (provincia di CT) Corso Ara di Giove n. 12, C.A.P 95030, consapevole che, ai sensi dell'art. 76 del D.P.R. 445/2000, dichiarazioni mendaci, formazione o uso di atti falsi sono puniti ai sensi del codice penale e delle leggi speciali in materia,

DICHIARA

che nel lavoro a firma congiunta:

Fordellone M., TOMASELLI V., Vichi M. (2018). From Tandem to Simultaneous Dimensionality Reduction and Clustering of Tourism Data. RIVISTA ITALIANA DI ECONOMIA, DEMOGRAFIA E STATISTICA, Vol. LXXII n. 1, p. 5-16, ISSN: 0035-6832.

Lista A.N.V.U.R. area 13 SSD 13/D3, Peer-review, IF=0,27 (researchgate.net). Fascia A, lista A.N.V.U.R., area 13, SSD 13/D3, Peer-review, IF=1.78,

il contributo degli autori è da considerarsi paritetico sotto ogni aspetto e l'ordine degli autori è esclusivamente alfabetico.

L'attribuzione della redazione dei paragrafi, tuttavia, è da intendersi nel seguente modo:

Fordellone M.: paragrafo 2
TOMASELLI V.: paragrafo 1, 3 e 4
Vichi M.: paragrafo 5.

La sottoscritta dichiara di essere informata, ai sensi dell'art. 10 della legge 675/96, che i dati sopra riportati saranno utilizzati nell'ambito del procedimento per il quale la presente dichiarazione viene resa.

Catania, 25/3/2018

La sottoscritta
Venera Tomaselli
Venera Tomaselli