CARMA 2025

7th Int. Conf. on Advanced Research Methods and Analytics
2-4 July 2025, Rome, Italy
DOI: https://doi.org/10.4995/CARMA2025.2025.20568

Composite indicators and Machine Learning techniques. An Application to the tourism industry

Giulio Pedrini, Giovanni Bonaccolto, Fabio Aiello, Marina Bonaccolto-Toepfer, Vincenzo Conti, Vincenzo Fasone, Raffaele Scuderi, Vincenzo Marinello, Vladislava Stankova, Emily Alaimo
Kore University of Enna, Italy.

How to cite: Pedrini, G.; Bonaccolto, G.; Aiello F.; Bonaccolto-Toepfer, M.; Conti, V.; Fasone, V.; Scuderi, R.; Marinello, V.; Stankova, V.; Alaimo, E. 2025. Composite indicators and Machine Learning techniques. An Application to the tourism industry. In: 7th International Conference on Advanced Research Methods and Analytics (CARMA 2025). Rome, 2-4 July 2025. https://doi.org/10.4995/CARMA2025.2025.20568

Abstract
Composite indicators are essential tools for summarizing complex and multidimensional phenomena into a single measure, aiding decision-making in various fields, including tourism. This paper preliminary reviews the main composite indicators used in the literature to assess the competitiveness of tourism destinations, along with their sustainability. Then the paper proposes the construction of composite indicators for the tourism sector, leveraging on Principal Component Analysis and Factor Analysis as key statistical methodologies enhanced with machine learning methods to improve accuracy, robustness, and interpretability. Through an empirical analysis on Italian tourism data, we compare standard and regularized methodologies, demonstrating how machine learning-enhanced approaches can improve the reliability and interpretability of composite indicators. Our findings provide valuable insights for researchers and policymakers seeking to develop robust and data-driven tourism performance measures.

Keywords: *Composite indicators; Tourism sector; Sustainability, Competitiveness, Principal Component Analysis, Machine learning techniques.*

1. Introduction

The first section starts on page two. Paper length must be between 4 and 8 pages, including the abstract page and all text, references, figures, and tables. Manuscripts that deviate from the guidelines, such as exceeding eight pages, adjusting margins incorrectly, or not conforming to the provided template, will be promptly rejected without evaluation of their content.

Composite indicators (CIs) have emerged as fundamental tools for summarizing complex and multidimensional information into a single measure, facilitating comparative analyses and policy decisions. These indicators are widely applied in fields such as economics, environmental studies, public health, and socio-economic development. Within the tourism sector, CIs play a crucial role in capturing the multifaceted nature of tourism performance by integrating diverse variables related to economic activity, infrastructure, demographics, and environmental factors. Indeed, a wide and heterogeneous set of variables and indicators has been proposed in the literature to assess the competitiveness and sustainability of tourist destinations. In particular, while for tourism competitiveness the set of proxies is quite restricted, referring to the multidimensional notion of tourism sustainability implies a significant increases of the number of relevant indicators (Manente, 2024).

Despite their advantages, constructing reliable and interpretable CIs presents several methodological challenges. The main difficulty lies in aggregating multiple heterogeneous indicators while preserving statistical rigor, reducing redundancy, and ensuring comparability. Traditional methods, such as Principal Component Analysis (PCA) and Factor Analysis (FA), are commonly employed for this purpose due to their ability to reduce dimensionality and extract latent structures from data. PCA transforms correlated variables into uncorrelated components, enabling the identification of key dimensions underlying the dataset, while FA estimates latent factors that explain observed relationships between indicators.

However, standard PCA and FA methods are not without limitations. The presence of highly correlated variables may distort the weights assigned to different indicators, leading to biased or misleading results. Additionally, traditional techniques do not perform feature selection, often incorporating all variables in the analysis, even those with minimal contribution to the final index. Furthermore, interpretability issues arise when principal component or factor loadings exhibit mixed signs or complex interactions, making it difficult to derive meaningful conclusions from the estimated CI.

To address these challenges, this paper explores the integration of machine learning techniques into the construction of CIs for the tourism sector. By incorporating methods such as Sparse PCA, Elastic Net regularization, and non-negative PCA, we aim to enhance feature selection, improve robustness against multicollinearity, and provide more interpretable results. Additionally, we take into account the use of advanced machine learning models, including tree-based algorithms and autoencoders, to refine the selection and aggregation of indicators.

The empirical analysis applies these methodologies to a dataset of Italian municipalities, focusing on tourism-related variables such as accommodation capacity, economic conditions, and demographic factors. We compare standard approaches with their machine learning-enhanced counterparts, evaluating their effectiveness in capturing tourism sector dynamics. Our findings contribute to the literature offering practical insights for researchers and policymakers seeking to develop data-driven and interpretable tourism performance measures.

2. Composite indicators: from concept to practical implementation in the tourism industry

2.1. Composite indicators

The term CI is widely used across various research and application domains to refer to tools suitable for measuring and representing a latent, often complex concept, commonly referred to as a construct. CIs are obtained from the combination of multiple individual indicators into a single numerical index. According to the OECD's (2008) handbook on CIs, they are constructed based on a conceptual framework that defines the phenomenon being measured. Since the phenomenon is observable but not directly measurable, it is decomposed into measurable quantities, known as components, dimensions, or preferably simple indicators (Aiello and Attanasio, 2008). Accordingly, the construction of a CI can be conceptualized as a forward-backward process, moving from the construct to its empirical representation and vice versa. Then, their construction involves several methodological challenges, particularly regarding the:

1. identification of elements (dimensions or components);
2. selection of empirical variables (i.e., simple indicators) and statistics to represent the defined dimensions;
3. normalization of simple indicators to ensure comparability and homogeneity;
4. identification of an appropriate weighting system to aggregate the transformed indicators;
5. selection of an aggregation rule to combine all measured and transformed information, to achieve the final measurement of the CI (Aiello and Attanasio, 2008);
6. robustness the final measurement of the CI (Greco et al., 2018).

The construction of CI thus follows a structured process, starting with the construct decomposition and then the identification of indicators to be selected according to theoretical and/or empirical criteria. This process is based on assumptions and operational choices, often characterized by a high degree of subjectivity and influenced to varying extents by the researcher's perspective. It relies on both substantive and methodological considerations (Aiello and Attanasio, 2008). The substantive ones pertains to the selection and orientation of variables: the choice of simple indicators is primarily guided by non-statistical evaluations, where they are often selected based on political-ideological considerations, reflecting specific viewpoints and biases.

CIs are applied, for instance, to construct the Human Development Index (HDI), the Environmental Performance Index (EPI), and the Global Competitiveness Index (GCI). A motivation for the construction of CIs is the need to summarize huge amounts of information into a more conceivable structure (Greco et al., 2018). Consider economic well-being that is not captured by GDP alone but requires broader indicators such as income distribution, education, and health. In addition, CIs are useful for tracking progress over time and assessing the impact of policy interventions.

2.2. Composite indicators in the tourism industry

As far as the tourism industry is concerned, illustrative CIs have been proposed by Asmelash and Kumar (2019), Ko (2005), and Kozic and Mikulic (2014). Asmelash and Kumar [2019] used a DPSIR framework and selected four dimensions: economic sustainability, environmental sustainability, socio-cultural sustainability, and institutional sustainability. Each dimension was then expanded into further sub-dimensions and items, as follows:

- Economic sustainability is divided into 11 items and three sub-dimensions;
- Environmental sustainability includes 12 items and four sub-dimensions;
- Socio-cultural sustainability consists of 18 items and 5 sub-dimensions;
- Institutional sustainability includes 12 items and 4 sub-dimensions.

Ko (2005) enlarged the number of relevant dimensions hierarchically grouping them into two ”systems”, the human system and the ecosystem:

- The human system includes the political dimension, the economic dimension, the socio-cultural dimension, and the production dimension, i.e. the quality of tourism services;
- The ecosystem includes the general environmental impacts, the quality of water, land and air, the biodiversity of flora and fauna, and the environmental policy and management.

Each of the eight dimensions then encompasses 4 indicators for a total of 32 indicators evenly distributed between the two systems. For each system, the relevant subset of indicators should be identified and measured to comprehensively assess the sustainability of that system with regard to the tourism industry.

Finally, Kozi´c and Mikulic (2014) proposed a more traditional set of indicators divided across the traditional sustainability dimensions (Table 1):

Table 1: Sustainable tourism indicators by dimension. Source: Kozic and Mikulic (2014).

Dimension	Indicator
Social	Tourists to locals Average net wage in hospitality to county net wage Low quartile to peak quartile employees Value for money (Accommodation, Gastronomy)
Economic	Seasonality of tourist arrivals Tourist expenditures per capita Daily tourist expenditures Hospitality subjects per 1,000 locals
Environmental	Water quality on beaches Tourist density per square km Wastewater management Biodiversity and landscape protection

This porposal also deserves attention for the use of three different weighting procedures of the selected indicators:

1. survey-based expert panel weighting
2. factor analysis weighting
3. equal weighting

The authors highlight that by adopting alternative weighting procedures, such as the ones listed above, the researcher or the decision-maker is able to check their different outcomes in terms of: (a) different weights (absolute and relative),(b) different rankings of tourism destinations; (c) different implications for the policymaker.

2.3. Weighting

The weighting step is crucial, as different techniques can lead to significant differences in terms of estimates and interpretation (Greco et al., 2018; OECD, 2008). Well-known weighting methods are: Equal Weighting (EW), Analytic Hierarchy Process (AHP), and Benefit of the Doubt (BoD). EW builds on the assumption that all indicators have the same contribution to the overall indicator. EW may not reflect the true significance of each variable. However, it is easy to implement and intuitive, and often produces good results. For instance, in the financial literature, EW is typically used to build financial portfolios that, although naive (as they do not rely on any estimation or optimization process), are often recognized for their outstanding out-of-sample performance (Bonaccolto et al., 2019). AHP is a multi-criteria decision-making method that has been extensively applied across various decision-making domains. Yet, its primary limitation lies in its sensitivity to biases and inconsistencies in expert assessments. BoD derives weights from observed data and revealed preferences, assigning higher weights to indicators on which performance is better and lower weights to indicators on which performance is poorer (Shwartz et al., 2009). This technique allows for unit-specific weighting schemes but may lead to difficulties in cross-comparisons (OECD, 2008).

Alternative methods are, for instance, PCA and FA. We focus on these two approaches in this study. The main advantage of PCA and FA is their ability to uncover latent structures in the data while reducing dimensionality. The methods help to reduce subjectivity by allowing the data itself to determine the weights, ensuring that the resulting factors or components are derived objectively rather than being influenced by predefined assumptions.

3. Machine learning-enhanced principal component analysis

PCA is one of the most common methodologies applied to the construction of CIs. By transforming correlated variables into uncorrelated components, PCA ensures that the most significant dimensions of variation are retained. However, its application to CIs is not without challenges. These include difficulties in interpretability when principal components (PCs) exhibit mixed signs in loadings, and considerable estimation errors when dealing with large datasets. Addressing these challenges is crucial to constructing reliable and meaningful indexes. In particular, in formative models, where indicators define the latent variable, PCA can produce misleading results due to its reliance on the correlation structure of the data. Specifically, problems can arise when PCA assigns incorrect signs to the eigenvectors, and does not preserve the intended polarities (positive or negative contributions to the latent concept). Additionally, PCA inherently assigns larger weights to highly correlated indicators and smaller weights to poorly correlated ones. This approach then leads to unstable weights in the presence of multicollinearity, especially when increasing the number of variables (Mazziotta and Pareto, 2024).

We can address the dimensionality issue by means of a particular machine learning technique: the Least Absolute Shrinkage and Selection Operator (LASSO) introduced by Tibshirani (1996). More directly related to our research, LASSO has been applied to PCA as well (Zou et al., 2006). The relevance of LASSO in PCA primarily stems from its ability to induce sparsity in the loadings of the PCs.

This approach, known as Sparse PCA (SPCA), simplifies the interpretation of the PCs by associating them with a (smaller) subset of variables. In our study, we take a step forward and also adopt a constrained PCA where both non-negativity and sparsity constraints are enforced on the principal axes (Sigg and Buhmann, 2008). Integrating Machine Learning techniques into the construction of CIs for the tourism sector can offer several methodological advantages. First, using feature selection algorithms, such as recursive elimination of less relevant variables through tree-based models like Random Forests or XGBoost, can improve the identification of the most significant indicators to include in the analysis. Similarly, rather than relying solely on the linear dimensionality reduction provided by PCA and FA, employing autoencoders or unsupervised learning techniques such as Self-Organizing Maps (SOM) could capture more complex relationships within the data. Furthermore, a hybrid approach that combines traditional statistical methods with advanced regression models, such as gradient boosting machines or neural networks, would allow for a more dynamic and effective estimation of the weight of different dimensions compared to a simple FA. Moreover, predictive models based on recurrent neural networks, such as LSTMs, or more recent transformer-based architectures could capture temporal dynamics more effectively than traditional statistical approaches. The integration of these advanced techniques could significantly enhance the accuracy and reliability of CIs, making them even more valuable tools for analysis, benchmarking, and decision-making in the tourism sector.

4. Standard and regularized PCA in constructing CIs: an empirical application to tourism data

As a practical application of the methodology presented in Section 3, we developed an empirical study focused on the tourism sector. This case study provides insight into the role of various socio-economic and geographical factors in shaping tourism dynamics. For this purpose, we collected data for 2,247 Italian municipalities, with observations referring to the year 2019. Data on tourism supply (hotels, Bed & Breakfast, other accommodation facilities) were extracted from the official figures on accommodation facilities provided by the Italian National Office of Statistics (ISTAT). Then, for each municipality, we considered a wide set of social, economic, and institutional characteristics, such as demographic trends, income levels, and political representatives of the municipality of interest. All these additional variables were drawn from the municipal database (the so-called ’Atlante Statistico dei Comuni’) also made available by the Italian National Office of Statistics (ISTAT). Finally, we included policy-related variables drawn from the database “opencoesione.it”, which collects all the information about the projects funded by cohesion policies in Italy in each municipality, notably their amount and objective. Grounding on this dataset, we selected a set of variables that are representative of the main dimensions encompassing the territorial characteristics of tourism destinations, to which we apply a logarithmic transformation to our data to have more symmetric distributions and reduce large differences in the scale of the original variables. Before applying the regularized PCA, we also standardize our data such that each variable has zero mean and variance of one.

The scores of the first PC result from the loadings displayed in Figure 1. Interestingly, the loadings associated with Population and ForeignP, as well as those corresponding to Hotels, OtherAcc, HotelsFund, and OthersFund, take similar negative values. Furthermore, their magnitude appears to be relatively large compared to the other variables, leading to the risk of obtaining highly distorted scores, in the light of their significant correlation. We can address this issue by estimating the SPCA model, which is subject to the ELNET penalty function (Zou et al., 2006).

We show the pattern of the regularized loadings of the first PC in Figure 2. Here, we presentthe estimates obtained for 100 values of the tuning parameter $\lambda 1$, which controls the intensity of the $\ell 1$-norm penalty. The parameter values range from 0.004 to 3, following an exponential progression. For simplicity, we set the $\lambda 2$ parameter to 1e-6. The x-axis of Figure 2 reports the logged $\lambda 1$ values. As expected, the solutions become sparser as $\lambda 1$ increases. Many loadings quickly converge to zero, while others reach zero for sufficiently large values of $\lambda 1$. Notably, the presence of negative loadings almost disappears for moderate values of $\lambda 1$, with the exception of Altitude, which takes strongly negative values and only reaches zero in the final part of Figure 2. In contrast, the ForeignP loading always takes positive values, and converges to one as $\lambda 1$ approaches its maximum value.

The results depicted in Figure 2 highlight the ability of ELNET to address the issues arising from the presence of highly correlated variables. However, it does not resolve the problem of mixed signs in the loadings, which poses challenges in terms of interpretation. This creates the need to impose a dual constraint: one on sparsity and the other on the sign of the loadings. This new setup leads to the SNNPCA method introduced by Sigg and Buhmann (2008), that can be estimated using the nsprcomp function available within the nsprcomp in R. In contrast to spca, nsprcomp requires the number of non-zero loadings rather than a sequence of $\lambda 1$ values. We show the sparse and constrained loadings associated to the first PC in Figure 3. Notably, all loadings take non-negative values. Similar to Figure 2, the greater contribution of ForeignP to the first PC also emerges in Figure 3: it increases from left to right along the x-axis and equals one when all the others are zero.

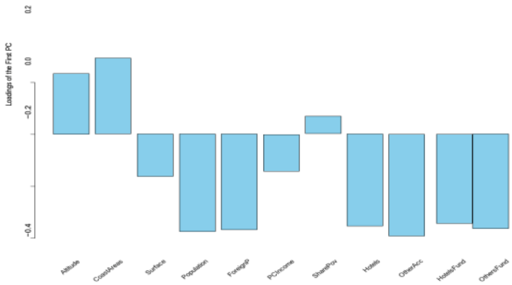

Figure 1: Loadings of the first PC estimated for the 11 variables in our dataset.

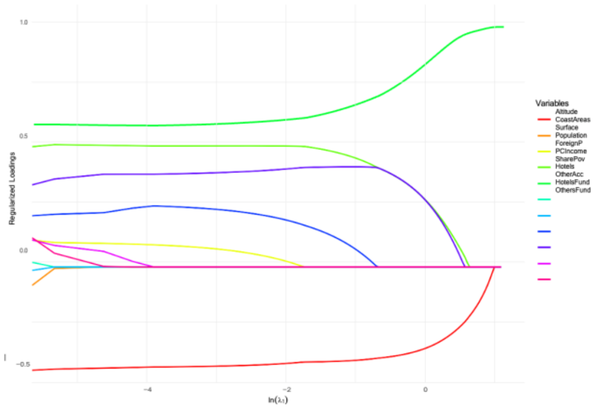

Figure 2: Regularized loadings as a function of ln λ1.

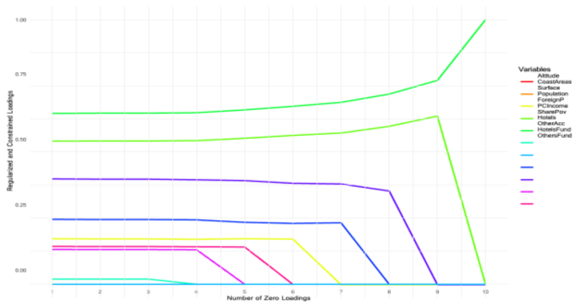

Figure 3: Regularized and constrained loadings.

5. Concluding remarks

The construction of CIs is a critical task in analyzing complex and multidimensional phenomena, particularly in the tourism sector, where various economic, social, and geographical factors interact. This paper has explored the methodological advancements in CI estimation by leveraging PCA, further enhanced through machine learning techniques. Our study highlights the strengths and limitations of standard approaches and demonstrates how advanced statistical and computational methods can improve the accuracy, robustness, and interpretability of composite indicators.

Traditional PCA offer well-established frameworks for dimensionality reduction and latent structure identification. However, their application in tourism data faces challenges such as multicollinearity, lack of explicit feature selection, and difficulties in interpreting component or factor loadings. To overcome these issues, we take into account machine learning-enhanced methodologies, including Sparse PCA, Elastic Net regularization, and non-negative PCA. These techniques allow for automatic variable selection, mitigate the effects of highly correlated indicators, and enhance the inter- pretability of the resulting composite indicators. Additionally, feature selection algorithms and hybrid machine learning models have been proposed as complementary tools to optimize variable inclusion and aggregation.

Our empirical application on Italian tourism data demonstrated the practical benefits of these methodologies. The comparison between standard and regularized approaches revealed that machine learning-enhanced methods produce more stable and interpretable composite indicators, reducing noise and redundancy while preserving essential information. The application of these techniques ensures that the constructed indices are not only statistically robust but also meaningful for policymakers and researchers seeking to assess and benchmark tourism performance across different regions.

Findings contribute to both the methodological literature on composite indicator construction and the practical implementation of data-driven tourism analysis. Future research can expand on this work by exploring alternative machine learning techniques, such as deep learning models for feature extraction, or investigating dynamic methods that account for temporal changes in tourism data. Moreover, incorporating uncertainty and sensitivity analyses can further enhance the reliability of composite indicators in policy applications.

In conclusion, integrating machine learning with traditional statistical techniques represents a promising approach to improving composite indicators. By refining variable selection, enhancing inter- pretability, and ensuring methodological robustness, this study provides a valuable framework for the development of more accurate and data-driven tourism performance measures. These advancements have broad implications not only for tourism analysis but also for other domains where composite indicators are widely used to support evidence-based decision-making.

References

Aiello, F. and M. Attanasio (2008): “Alcune considerazioni sul processo di costruzione degli indicatori composti,” in Dottor DIVAGO: Discernere, Valutare e Governare la nuova Università, ed. by V. Capursi and G. Ghellini, Milano: Franco Angeli, Collana dell’Associazione Italiana di Valutazione: 1900.2.4 – Teoria, metodologia e ricerca, 123–138.

Asmelash A. G. and Kumar S. (2019). Assessing progress of tourism sustainability: Developing and validating sustainability indicators. *Tourism Management*, 71:67–83, 2019.

Bonaccolto, G., M. Caporin, and S. Paterlini (2019): “Decomposing and backtesting a flexible specification for CoVaR,” *Journal of Banking & Finance*, 108, 105659.

Greco, S., A. IshizaKa, M. Tasiou, and G. ToRRisi (2018): “On the methodological framework of composite indices: a review of the issues of weighting, aggregation, and robustness,” *Social Indicators Research*, 141, 61–94.

Ko T. G. (2005). Development of a tourism sustainability assessment procedure: a conceptual approach. *Tourism management*, 26(3):431–445.

Kozic I. and Mikulic J. (2014). Research note: Measuring tourism sustainability: An empirical comparison of different weighting procedures used in modelling composite indicators. *Tourism economics*, 20(2):429–437.

Manente M. (2024). Sustainable tourism development: indicators and thresholds. In Italian Association of Regional Science-AISRE –XLV Conference –Turin, 6th September 2024, 2024.

Mazziotta, M., & Pareto, A. (2024). Principal component analysis for constructing socio-economic composite indicators: theoretical and empirical considerations. SN Social Sciences, 4(6), 114.OECD (2008): *Handbook on Constructing Composite Indicators: Methodology and User Guide*, OECD Publishing.

Sigg, C. D. and J. M. Buhmann (2008): “Expectation-maximization for sparse and non-negative PCA,” in Proceedings of the 25th international conference on Machine learning - ICML ’08, ACM Press, ICML ’08, 960–967.

Shwartz, M., J. F. Burgess, and D. Beriowitz (2009): “Benefit-of-the-doubt approaches for calculating a composite measure of quality,” Health Services and Outcomes Research Methodology, 9, 234–251.

Tibshirani, R. (1996): “Regression shrinkage and selection via the Lasso,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58, 267–288.

Zou, H., T. Hastie, and R. Tibshirani (2006): “Sparse Principal Component Analysis,” *Journal of Computational and Graphical Statistics*, 15, 265–286.