Social Indicators Research
https://doi.org/10.1007/s11205-023-03234-2

ORIGINAL RESEARCH

Spatial Analysis to Investigate the Relationship Between Tourism and Wellbeing in Italy

Najada Firza[1,2] · Laura Antonucci[3] · Corrado Crocetta[4] · Francesco Domenico d'Ovidio[1] · Alfonso Monaco[5,6]

Accepted: 2 October 2023

Abstract
The level and variety of services offered by tourist destinations are intricately linked to the overall health and condition of its area. We would like to investigate the existence of a possible connection between tourism and the social, economic, and environmental well-being of a territory. The tourism industry can improve the general well-being of a specific area by promoting consumption, reducing the income gap, and improving infrastructures. However, the well-being of the territory through enhancing the specific features of the local context and its factors of excellence can also influence tourism.

In this context, we applied Machine Learning methods to investigate the relationship between tourism and well-being in Italy. The analysis used Italian BES indicators at the provincial level, referred to a time window of 17 years (2004–2020). We developed a Machine Learning algorithm based on a hybrid (unsupervised and supervised) approach to study 51 well-being indexes and 9 tourism indicators. We found a close connection (80% of accuracy) between tourism and well-being. We also selected a group of tourism indicators that have a strong effect on this connection. Using eXplainable Artificial Intelligence (XAI) methods, we detected that tourism in low season periods ranks first for importance followed by the spread of farms business and urban green areas density. Our research suggests that improved social, economic, environmental, and health well-being can positively spill over the effect on tourism arrivals and revenues in the long period.

Keywords Well-being · Tourism · Machine learning · eXplainable Artificial Intelligence

1 Introduction

The attractiveness of a tourist destination is closely linked to the social, economic and environmental well-being of its territory and it arises from the dynamic and systemic interaction between local actors who directly or indirectly participate to produce economic, social, cultural and environmental benefits (Crouch et al., 1999). These benefits must lead to a substantial improvement in the current and perspective level of well-being of the local com-

Extended author information available on the last page of the article

munity. But how to assess the well-being of a territory? Should social and environmental factors also be considered in addition to economic variables?

For more than 60 years, well-being has been measured through the Gross Domestic Product (GDP), the main economic indicator that measures the well-being of a territory by focusing on economic variables.

GDP evaluates a country's economic growth and well-being both over time and relative to other countries but presents some limitations that should be highlighted. In fact, GDP does not consider essential factors such as environmental, social, human and health aspects that are an integral part of economic transactions. Several studies showed the limitations of GDP to evaluate the level of a country well-being. Talberth et al., reported that there is a threshold level beyond which the increase in GDP no longer leads to an increase in well-being (Talberth, Cobb et al., 2007) due to factors that GDP does not consider as the impoverishment of natural resources, the increase in income inequality among citizens and the low quality of free time (Max-Neef, 1995). The inclusion of these factors becomes pivotal to ensure the community continuous growth under the banner of life quality and long-term well-being (Costanza et al., 2009). Several solutions have been proposed over the years able to measure the well-being of a country not only by means of economic indicators but also using indexes that assess qualitative variables such as social and psychological indicators (Bleys, 2011).

Since 2013 the Italian National Institute of Statistics (ISTAT) has developed a system of indicators of fair and sustainable well-being (BES) on a regional or provincial scale with the aim of integrating economic indicators with additional indicators highlighting the multidimensionality nature of well-being.

BES is divided into 12 domains further divided into 130 indicators. Every year since 2015, ISTAT provides a composite index for each domain and draws up an annual report on the well-being situation in Italy that helps the government's economic planning for public policies of well-being and sustainability.

Therefore, BES to measure the well-being of a territory integrates and completes the GDP by giving an overview of the state of health of a territory.

Sustainable tourism is merged into a macro-system where economic, social and environmental factors interact closely and support each other to create benefit and well-being for the area (Gibson et al., 2005).

In literature, sustainable tourism (which originates from responsible tourism) incorporates the social well-being of the local community (XIII Assembly of the UNWTO – United Nations World Tourism Organization, 2001). Consequently, we recognize the importance of the link between the well-being of a local community and its tourism (Buhalis, 2000) but we do not know how the two phenomena can be linked to each other.

As a result, we do not know how to measure this link. In the construction of the sustainable tourism indicators (ETIS toolkit, 2016) of the European Union, an important part is played by the socio-cultural impact indicators of the local community of destination. Even these indicators, however, are not real indicators of the well-being of the territory and highlight how attractive the territory is.

For this reason, it is not yet clear how much the well-being of the territory expressed through the BES is connected to the attractiveness of a territory and therefore what is the connection between tourism and well-being.

Hence the need to build a statistical model to verify the existence of this connection.

In the present work, we focused on broad data, applying methods of machine learning to find a link between well-being variables and tourism factors. We used data collected from the Italian national statistical institute data repository, referred to Italian provinces, from 2004 to 2020. We followed a hybrid approach based on supervised and unsupervised machine learning techniques. Specifically, we outlined at a provincial scale the existence of general spatial patterns for different well-being factors through the fuzzy c-means method; then, we applied the Random Forest algorithm to investigate and better understand the association between the well-being drivers and tourism variables. Among the considered tourism features, we investigated the most influential ones in determining the connection with well-being and social aspects on Italian territory through two different feature importance approaches: (i) a global one based on Random Forest; (ii) a local one by means of eXplainable Artificial Intelligence (XAI) an innovative framework that allows to improve the interpretability and transparency of Machine Learning techniques.

The article is organized as follows: in section "Materials and method" we presented details on data collection and processing and on the machine learning and XAI techniques used in our research. In section "Results" we reported the main findings of our research. Finally, in section "Discussion" we discussed our results underling their relationship with previous literature.

2 Connection Between Tourist Destinations, Sustainable Tourism, Well-being and Well-being of the Territory

The strong instability due to the COVID-19 pandemic, has brought out the fragility of global tourism with all its facets. On the other hand, this emergency has highlighted the "resilience" and desire for transformation of this sector (Lemmi, 2020).

The transformation process that starts from the concept of "resilience" (resilience means the desire to live, redeem oneself after a crisis and therefore manifest one's desire to live through actions that express resilience: the desire to anchor oneself to life even stronger than the crisis itself), finds its daily bread in rediscovering and consequently revaluing or enhancing the authenticity that the territory offers.

Retrain local resources in terms of competitiveness after having become more aware of their importance.

As a result, a series of basic changes have been made in the sector after an awareness of the importance of territorial resources and above all their consumption (Lemmi, 2014).

The territorial resources in environmental, cultural and social key are the raw wealth of tourist destinations and nowadays the competitiveness of tourism runs through the enhancement of this wealth.

Therefore, we cannot talk about sustainability in tourism without first talking about the well-being of the territory. All the concepts just mentioned are closely connected to each other. The well-being of the territory in general and of tourist destinations in particular, must be understood in a broader vision of the term that considers well-being without being anchored only to the economic part of the same.

To highlight the connection between tourism and well-being and all the benefits that can derive from the activation of this connection, we must free ourselves from the traditional

vision of production of goods and services and emphasize what actually happens by incorporating all the factors involved.

Starting from this assumption, the publication in question, deals with the well-being of the territory under three fundamental aspects: Social, Environmental and Economic (Gibson et al., 2005).

Tourist destinations are beginning to innovate and revisit their idea of tourism, which initially becomes responsible tourism and then sustainable tourism and embraces the social and environmental values of the local community (Assante et al., 2012).

Travellers and hosts interact in a synergistic way and the closer the interaction between the two, the more benefit is created by increasing the well-being of the territory and self-nourishing sustainable tourism (Lemmi, 2020).

Even territorial development and marketing policies have increasingly paid special attention to the close connection between the well-being of the territory and its citizens and tourist attractiveness.

So, what are those variables that can detect and measure the well-being of the territory in tourist destinations?

The quality of services in the area is an important indicator of well-being that contains important services for the community of reference but also for the tourist himself. In fact, services such as "the coverage of public transport in the territory", "the beds available in hospitals", "the number of medical specialists in the territory" and so on, are indicators of social well-being of the places that welcome the tourist.

The ability of territories to connect to their environment and support its sustainability with actions such as "the use of energy generated from renewable sources", "generate responsible urban waste and through separate collection" and so on, are indicators relevant to the environmental well-being of the territory.

And finally, "employment rate", "youth employment rate", "per capita income" and so on, are indicators of the economic well-being of the territory.

Tourism indicators, on the other hand, are closely linked to indicators of territorial well-being. Cultural destinations, for example, to increase their competitiveness and extend new tourist scenarios of territorial regeneration through the combination of culture and tourism, focus on the close interconnection between museums and their territory so that they become interactive places.

The availability and increase of urban green in cities are another point of attractiveness of urban tourism that regenerates the concept of tourist cities.

The spread of agritourist companies is linked to rural tourism, slow tourism that solicits the ecotourism. No new concept but revisiting and emphasizing territoriality.

3 Materials and Methods

Not all the considered variables have values in all 17 years analysed. To overcome this inconsistency of the analysed database, we carried out a spatial analysis considering the average of the BES indicators on the reference time window, and we found a spatial link between BES and tourism.

To assess the link between well-being and tourism indicators, we implemented a learning framework based on supervised and unsupervised approaches, as summarized in Fig. 1.

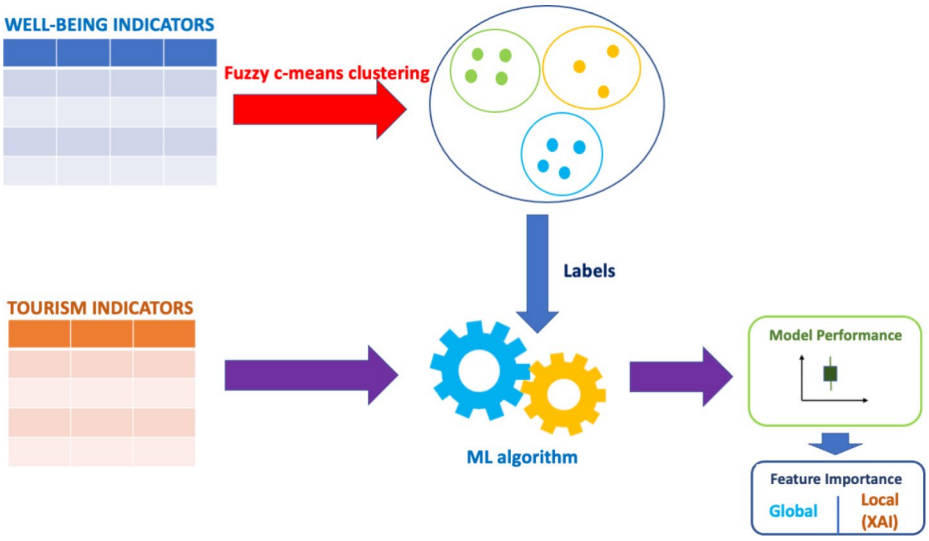

Fig. 1 The flowchart of the proposed methodology

After a data collection phase and a preliminary analysis implemented by means of Canonical Correlation Analysis, we used a cluster method to assign a membership to each Italian province based on well-being factors. Using these memberships as labels and 9 tourism indicators as features, we fed a supervised learning model. The idea behind our procedure is that if there is a relationship between well-being and tourism indicators the accuracy of the supervised model should be high. Finally, we implemented a feature importance procedure using a global approach based on Random Forest internal functionalities and a local one that takes advantage of the Shapley (SHAP) values method.

3.1 Data Collection

We collected data for the years 2004–2020 from the ISTAT public data base. We considered 51 well-being factors and 9 Tourism indicators (https://www.istat.it/en/well-being-and-sustainability/the-measurement-of-well-being/bes-at-local-level, Dataset BES at local level 2021 edition). For further details on well-being database see table S1 in the supplementary information section.

Well-being indicators are divided into 10 categories: Health, Education and training, Work and life balance, Economic well-being, Social relations, Politics and institutions, Safety, Environment, Innovation – research and creativity, Quality of services.

There are some indicators, within the BES data, that notoriously belong to tourism indicators. To avoid data redundancy and bias in the analyses we preferred to place these indicators in the tourism indicators group. Specifically, the following three indicators were therefore removed from the BES data: “Landscape and Cultural Heritage”, the Environmental indicator “Availability of urban green” and the indicator of Innovation, Research and Creativity “Employees in cultural enterprises”.

Thus, below we report the indicators on tourism:

1. Tourism rate: Built as days of presence of Italians and foreigners in the complex of accommodation establishments per inhabitant.
2. Tourism in the non-summer months: Built as days of presence of Italians and foreigners in the complex of accommodation establishments in the non-summer months per inhabitant.
3. Density and relevance of museum heritage: Built as Number of permanent exhibition structures per 100 km2 (museums, archaeological sites and monuments open to the public), weighted by the number of visitors.
4. Spread of agritourism farms: Built as number of farms per 100 km^2.
5. Density of historic green: Built as an area in m^2 of the areas of historic green and urban parks of considerable public interest in the provincial capital municipalities, per 100 m^2 of urbanized area (centres and inhabited nuclei) detected by the Population Census (2011).
6. Employees in cultural enterprises: Constructed as a percentage of employees in local units of enterprises carrying out a cultural economic activity out of the total number of employees in local units of enterprises.
7. Availability of urban green: Built as square meters of urban green per inhabitant in the provincial capital municipalities.
8. Cultivated areas: Percentage of land used for crops. Data estimated through the land cover classification gridded map available on the Copernicus Climate Data platform (https://cds.climate.copernicus.eu/portfolio/dataset/satellite-land-cover). The land cover classification data categorizes the land surface into 22 classes using the FAO Land Cover Classification System (LCCS). Cultivated areas are the percentage of gridded data (native resolution – 300 m) classified as "croplands; mosaic cropland (>50%) / natural vegetation (tree, shrub, herbaceous cover) (<50%); mosaic natural vegetation (tree, shrub, herbaceous cover) (>50%) / cropland (<50%); tree cover, broadleaved, deciduous, closed to open (>15%)" with respect to the total number of gridded points falling within the administrative boundary of the province.
9. Urban areas: Percentage of land used for urban areas. As for "Cultivated areas" but selecting the land class if "Urban areas".

3.2 Canonical Correlation Analysis

As a preliminary analysis, before applying our machine learning-based procedure, we used the Canonical Correlation Analysis (CCA) (Hotelling, 1992) to investigate the existence of a link, between two following sets of data: the first composed by 51 equitable and sustainable well-being indicators and the second with 9 tourism variables. Given two independent datasets A and B, the CCA detects the pairs of linear combinations, one of each dataset, most closely related to each other. Specifically, CCA finds the best possible correlation between two independent datasets. A and B contain N instances and *a* and *b* features, respectively. CCA can discover k pairs of canonical variables for each instance, where k represents the smallest number between *a* and *b*. Each pair includes a combination of A's features as the first element and a combination of B's features as the second element. These pairs are ranked in descending order based on their correlation values, with the first pair having a higher correlation than the second, and so on.

Once we established the existence of a relationship between the two datasets analysed in this work, we applied a procedure based on supervised and unsupervised machine learning techniques to assess this relationship.

3.3 Clustering Analysis

We applied clustering algorithms to the well-being indicator dataset to find natural groupings in the data. Firstly, we used k-means with Euclidean distance as an exploratory analysis and fuzzy c-means (FCM) as a confirmatory analysis to obtain more robust results. In both cases, heterogeneous variables were normalized by Range method 0–1, instead by Z-score, because in the preliminary analyses our data revealed different discriminant power due to their different variance. FCM assigns by randomly a weight of membership u_{ij} to each element x_i which belongs to each cluster J that we want to find. By means of an iterative process, an objective function composed by sum of the distances of each point from each cluster centre is minimized by moving dynamically the cluster centres c_j. This function is defined as (Bezdek JC et al., 1974):

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \|x_i - c_j\|^2 \tag{1}$$

where N is the number of instances in the dataset; C is the desired number of clusters imposed by the researcher; m is the fuzzifier parameter, a hyper-parameter that controls the "fuzzy degree" of the cluster. We used the Euclidean distance metric to compute the distances between the cluster centres and each observation in the sample. Since FCM is an explicit method, in which the number of clusters is chosen by the researcher, we computed the silhouette coefficient that measures how well an instance is clustered through the proximity of each point in a cluster to points in neighbouring clusters (Bezdek et al., 1974).

The clustering procedure assigns to each Italian province a tag representing the membership in a particular cluster. We used this membership as a label to train two classifiers: logistic regression and Random Forest.

3.4 Supervised Learning Framework

We started with logistic regression (LR) and then we used a machine learning approach based on Random Forest to improve the model performances. Specifically, the Random Forest (RF) algorithm, introduced by Breiman in 2001 (Breiman, 2001), is a widely used supervised machine learning technique. It involves a collection of binary classification trees (CART) and is known for its versatility and simplicity in tuning, requiring only two parameters - the number of trees in the forest (D) and the number of randomly selected features (F) for each split. RF creates the forest through a training process that utilizes a bootstrap method and a feature randomization procedure. These techniques make RF well-suited for modelling multimodal data and mitigating overfitting issues. Additionally, RF can evaluate the importance of each feature through an internal feature importance procedure. Furthermore, thanks to its out-of-bag (OOB) estimation RF generates an unbiased estimate of the generalization error.

Since the RF algorithm derives by CART techniques, enhancing their classificatory power without significant changes in statistical constraints. Fielding et al. (Fielding et al., 1973) suggest that analyses on the whole population (not on samples) and the data are from high level units (that is, complex units containing elementary units), random errors compensate each other and relatively few units can be used.

In our work we trained RF model though a dataset composed by 107 Italian provinces and 9 tourism indicators in which the label of each province is the membership assigned by FCM, as previously mentioned. Furthermore, we evaluated the feature importance through the mean decrease impurity and used a RF standard configuration with D = 500 trees and F= $\sqrt{H}$ where H is the number of input features.

To increase the robustness of our framework we implemented a 5-fold cross validation (CV) framework. In this model the initial dataset containing data of 107 Italian provinces is randomly divided into 5 subsets without re-insertion: we used $5-1$ subsets for the training phase and the remaining part for validation. We repeated this procedure 100 times so the average of the 100 performance values is a reliable indicator of the overall model accuracy. In the same way, RF feature importance is evaluated by means of 100 CVs, and overall feature importance is computed by averaging.

3.5 Explainable Artificial Intelligence and Shapley Values

The Explainable Artificial Intelligence (XAI) techniques were developed to improve the transparency and interpretability of Machine Learning models (Jiménez-Luna et al., 2020; Miller, 2019). These techniques encompass a range of approaches that aim to enhance properties of AI models such as informativeness, uncertainty estimation, generalization, and transparency, which refers to the ability to make decision-making as understandable as possible (Flach, 2019; Vollmer et al., 2020).

In our analysis, we implemented the SHAP local explanation method to assess features importance for the RF classification, described in previous section. Unlike the feature importance evaluated entirely by Random Forest which gives global information of the machine learning algorithm on the whole training set, SHAP provides the contribution made by each feature in the prediction of the single observation.

In this way we can compare the findings obtained through two independent methods.

The SHAP method utilizes the Shapley values, which are based on cooperative game theory (Lundberg et al., 2017) and offers a local and model-independent explanation tool. This tool uses interpretable linear models for individual samples and focuses on the impact of each feature on the prediction. The SHAP value for a feature j is calculated as the variation in the model's output prediction with and without that feature, while considering all possible subsets F of the total feature set S ($F \subseteq S$).

The SHAP value of the j-th feature for the observation x is measured by considering the addition of the j-th feature to all possible subsets,

$$\sum_{F \subseteq S-\{j\}} \frac{|F|!\,(|S|-|F|-1)!}{|S|!} \left[f_x(F \cup j) - f_x(F)\right] \tag{2}$$

where $|F|!$ is the number of permutations of features positioned before the j-th feature; $(|S|-|F|-1)!$ is the number of permutations of feature values that appear after the j-th fea-

ture value; $|S|!$ is the total number of feature permutations; $f_x(F)$ represents the prediction f of a model for the observation x, given a subset F that does not include the j-th feature; $f_x(F \cup j)$ is the prediction of the same model when the j-th feature is included.

For our analysis, we used the R framework version 3.6.1 (https://www.r-project.org/) with packages: CCA version 1.2.1 to perform CCA analysis, Factoextra version 1.0.6 to perfom K-means, e1071 version 1.7.2 to perform FCM, randomForest version 4.6.14 to perform Random Forest classifier, RColorBrewer version 1.1.3 to plots our results, DALEX version 2.4.2 to perform SHAPE and Metrics version 0.1.4 to compute accuracy.

4 Results and Discussion

We investigated the relationship between well-being and tourism of Italian territory combining clustering method, machine-learning procedure and XAI approach.

First, we verified the existence of a link, between well-being and tourism indicators through the CCA with a linear correlation found of 0.90 between the first two pairs of canonical variables.

Since the dataset of tourism indicators contains nine variables, CCA returns only 9 pairs of canonical variables; with the first to be preferred correlation as it yields a stronger correlation between the linear combination of the nine tourism indicators and the linear combination of the 51 well-being variables. We found a linear correlation for the first pair equal to 0.90 that indicates the existence of a strong relationship between the two analysed databases (or at least between some variables). Having obtained this indication, we can implement our procedure based on clustering algorithms as described in section Methods.

4.1 Clustering Algorithms

The silhouette analysis performed over 51 well-being variables suggested the existence of 2 clusters of provinces, with a maximum value of silhouette greater than 0.3 (see Table 1), as the optimal outcome.

The result of fuzzy c-means clustering for general well-being category is reported in the upper panel of Fig. 2. The two found clusters contain 65 and 42 provinces respectively.

To further investigate the result of the fuzzy c-means we divided the well-being indicators into 3 macro categories (Costanza et al., 2009; Gibson et al., 2005; Alaimo, 2018).

The economic macro-area is part of a broader and more inclusive system and is closely interdependent with two other macro areas: Environment and Society. These three aspects are equally important for measuring a country's well-being. For this reason, we divided our indicators into: Environment (6 features), Economy (11 features) and Society (34 features).

For these three macro categories we repeated the clustering procedure of the Italian provinces through silhouette coefficient and fuzzy c-means. The results are reported in the lower panel of Fig. 2.

Table 1 Cluster validation: Average silhouette coefficients for different value of k in general economic, social and environmental dimensions

Well-being and its dimensions	k=2	k=3	k=4
General	0.32	0.19	0.17
Economic	0.50	0.29	0.25
Social	0.33	0.31	0.15
Environment	0.30	0.31	0.30

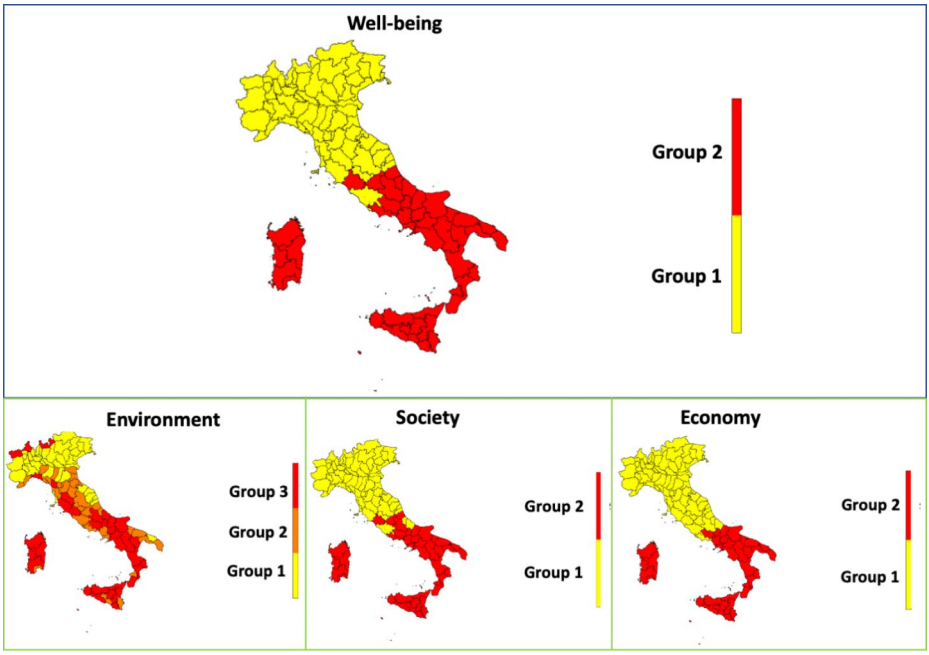

Fig. 2 Fuzzy c-means results for general well-being category (upper panel) and for the subcategories: Environment, Society and Economy (lower panel)

As we can see from Table 1, the average silhouette coefficients, indicate a clearer clustering for the Economic macro-area. This is not a new finding, in fact, it turns out that the differences in terms of well-being between North and South Italy are mainly economic.

Using the k-means algorithm we obtained an output very similar to FCM with strong clusters overlap (an average difference of about 8% of the provinces on the main category and the three sub-categories). Therefore, for the sake of simplicity, hereafter we used only the results obtained through FCM.

As shown in the upper panel of Fig. 2, the Italian provinces, according to the fuzzy c-means algorithm fed with the whole database of well-being indicators, are distributed over two geographical clusters: one group belonging to Northern Italy and the other one belonging to Southern Italy (D'Urso et al., 2022).

This indicates that well-being in Italy has a clear geographical location. This trend is also confirmed by the results of the clustering algorithm applied to the economy and society subcategories, in which the 2-class pattern is confirmed with some southern provinces passing into the northern group (Alaimo et al., 2020).

So also, about social and economic well-being Italy is divided into two parts. As regards the analysis of environmental data, the silhouette coefficient suggests the existence of 3 clusters. The associated environmental risk grows with the number of clusters (minimum for cluster 1, maximum for cluster 3). Also, for the environmental data it is evident that provinces with the greatest environmental risk are in Southern Italy, confirming a greater environmental well-being in Northern Italy.

4.2 Supervised Learning Framework

We used the outcome of FCM clustering for the general category of well-being factors as labels to train logistic regression and Random Forest algorithm (we fed the two models with the 9 tourism indicators reported in section Materials). We performed a repeated (100 times) 5-fold cross-validation procedure (described in Material and method section) and obtained a mean classification accuracy equal to 0.80 ± 0.02 and 0.69 ± 0.02 for RF and LR respectively. Then the best-performing method was RF in terms of accuracy.

Figure 3 shows the OOB error of RF, for different trees, averaged over 100 rounds of cross-validation. RF reached an OOB average error about 20%.

The results of the feature importance procedure performed through the RF algorithm are summarized in Fig. 4, where each distribution was computed through a 5-fold cross-validation process repeated 100 times. As previous mentioned, this represents a global importance assessed on the training sample. It is possible to detect that "tourism in the non-summer months" indicator, "the diffusion of farms" and "the urban green dispersion" tend to have a consistently high importance.

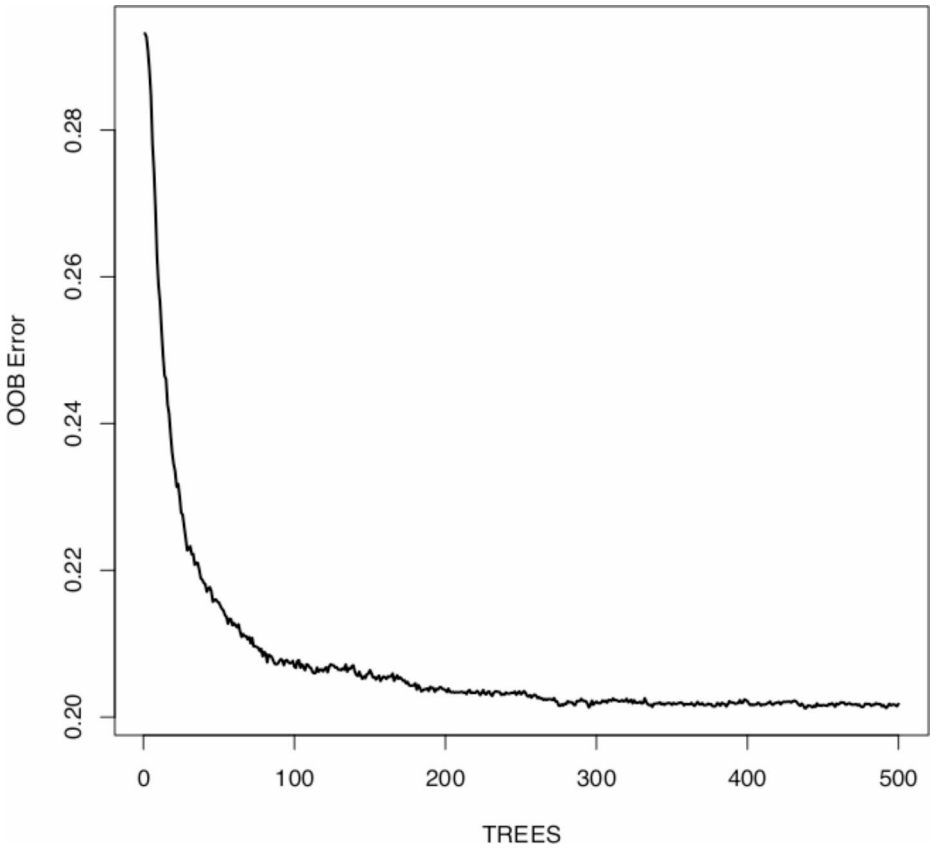

Fig. 3 OOB error of RF for the different trees. The error over 100 rounds of cross-validation has been averaged

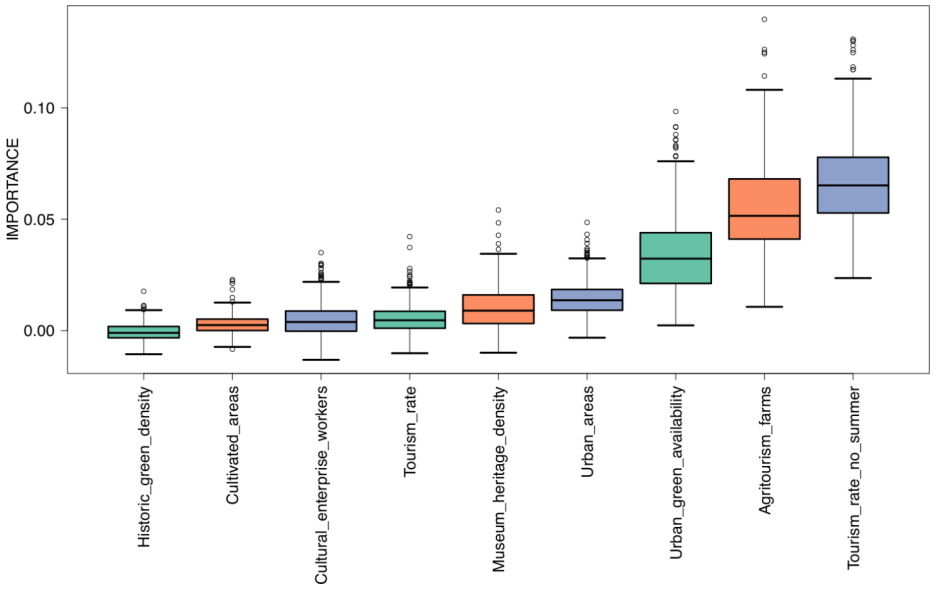

Fig. 4 Feature importance computed by means of RF. Each distribution was computed through a 5- fold cross validation procedure repeated 100 times

We computed the SHAP values associated to the RF prediction for each given province. Figure 5 shows the distribution of SHAP values for each Italian province for the group one and two (Panel A and B respectively). In these plots the features are ordered in terms of importance in the model.

As can be seen from the results, our paper measures with good accuracy the connection between BES and Tourism.

Many works in the literature have highlighted the existence of such a connection. The novelty of our work is to have measured and quantified the link between tourism and BES, through a robust and multivariate model that considers within it the various facets of well-being and tourism.

From the feature importance of RF (shown in Fig. 4), we found that the three most important tourism features for classifying the provinces in the two groups of well-being level are respectively: “tourism rate in the non-summer months”, “the number of business farms”, “the availability of green zone in urban areas”. Therefore, it is worth to note that in the regions belonging to group 2 (especially in South of Italy) the pivotal tourism is connected to the summer period, thanks to the seaside tourism, while winter and cultural tourism is more marginal. These findings are also confirmed by the XAI analysis shown in Fig. 5. The Shap-plots show that high values of these three features characterize belonging to cluster one (Panel A), while low values address the provinces towards cluster two (Panel B). The museum density indicator reports high values for group one and low values for group two, confirming that South of Italy lacks a long-term marketing strategy for cultural tourist attractions.

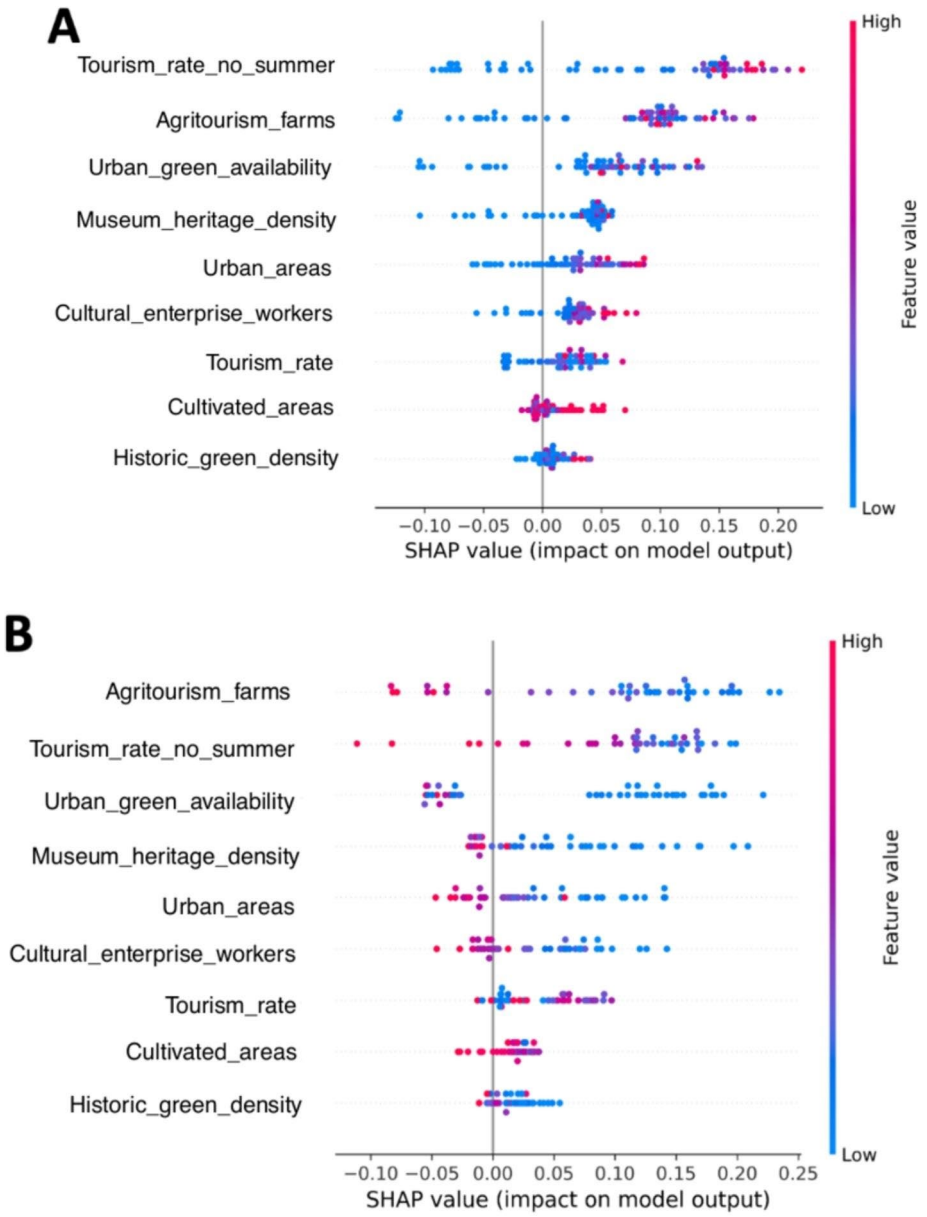

Fig. 5 SHAP distribution values of the most influential features for the group 1 (Panel A) e for the group 2 (Panel B). Each point in the same row corresponds to a different province

The differences that emerge between the two clusters found by the fuzzy c-means seem to derive from a deep disparity of available infrastructures that enhance the territory and from a not very effective development policy linked to tourism in southern Italy.

Our findings suggest the existence of a robust connection between well-being and tourism, furthermore high value of well-being (social, economic, environmental) support a more heterogeneous and less sectoral tourism and more distributed during the year.

Through our results, local and national authorities can understand where to intervene at the political and infrastructural level to improve the level of service delivery and to implement a radical transformation of tourism.

From our analysis, for example, the southern provinces are linked to a purely summer seasonal tourism. So, by intervening also on services (they are not strictly related to seaside tourism), such as the quality of museums and archaeological sites and marketing on cultural tourism, a different type of tourism can be improved by expanding and diversifying the tourist offer of the territory.

5 Conclusion

In this work we studied the connection between tourism and well-being at Italian provincial level in a specific time window (2004–2020). To do this we have implemented an analysis based on both clustering methods and supervised ML methods on data extracted from databases publicly provided by ISTAT.

We applied the fuzzy c-means method to well-being dataset to group each Italian province and we used these memberships as labels of the Random Forest algorithm fed with 9 tourism factors. We obtained a good classification performance (0.80 ± 0.02), by means of CV framework, to testify the close connection among well-being and tourism. Finally, we investigated which features were best at characterizing this connection through a global approach, with Random Forest internal functionalities and local one, based on the Shapley values method.

Random Forest results highlighted that the tourism indicators most closely linked to well-being are “Tourism in the non-summer months”, “Density and relevance of museum heritage”, “Spread of agritourism farms”, “Availability of urban green”. The SHAP analysis confirms the RF feature importance both for the group of southern and northern provinces, underlining an opposite impact on the tourism sector (negative for South and positive for North). The southern provinces show deficiencies at a structural level, highlighting an insufficiency of services and a low attractiveness of tourist destinations in the non-summer months.

Our clustering results show that well-being in Italy is well distributed geographically, with the northern provinces presenting a better condition than the southern ones. These differences are mainly economic and social. Instead, the situation seems to be more complex for the environmental field. In fact, even some provinces in central (and some northern) Italy have a critical environmental situation, which is mainly due to pollution and a difficult hydrogeological condition.

Our analysis not only quantifies a strong connection between well-being and tourism but also indicates the factors that characterize it, suggesting to local authorities where they can intervene to improve the quality of services and the quality of tourism connected to them.

Efficient tourism policies cannot ignore the well-being of the territory. Only through sustainability is it possible to build a virtuous circle in which well-being and responsible tourism can develop and feed each other (Global Code of Ethics for tourism, 1999).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11205-023-03234-2.

Authors' Contribution Statements Conceptualization NF, AM, CC; Methodology NF, AM; Formal analysis NF, AM; Writing (Original Draft) NF, AM; Writing (Review Editing) NF, AM, LA, CC, FD; Data Curation NF; Software NF, AM; Visualization NF, AM; Validation NF, AM, CC, FD. All authors have read and agreed to the published version of the manuscript

Funding The authors did not receive support from any organization for the submitted work.
Open access funding provided by Università degli Studi di Bari Aldo Moro within the CRUI-CARE Agreement.

Data Availability Data used in this analysis are available upon request.

Code Availability R codes used to perform this analysis are available upon request.

Declarations

Competing Interests The authors have no competing interest to declare that are relevant to the content of this article.

Disclosure of Potential Conflicts of Interest No conflict of interest to declare.

Ethics Approval This study not involving humans or animals.

References

Alaimo, L. S. (2018). Sustainable development and national differences: An european cross-national analysis of economic sustainability. RIEDS-Rivista Italiana di Economia, Demografa e Statistica-Italian Review of Economics. *Demography and Statistics*, *72*(3), 101–123.

Alaimo, L. S., & Maggino, F. (2020). Sustainable development goals indicators at Territorial Level: Conceptual and methodological issues—the italian perspective. *Social Indicators Research*, *147*, 383–419. https://doi.org/10.1007/s11205-019-02162-4.

Assante, L. M., Wen, H. I., & Lottig, K. (2012). An empirical assessment of residents' attitudes for sustainable tourism development: A case study of O'ahu, Hawai'i. *Journal of Sustainability and Green Business*, *1*, 1–27.

Bezdek, J. C. (1974). Cluster validity with fuzzy sets. *Journal of Cybernetics*, *3*, 58–73. https://doi.org/10.1080/01969727308546047.

Bleys, B. (2011). Beyond GDP: Classifying Alternative Measures for Progress. *Social Indicators Research*, *109*(3), 355–376. https://doi.org/10.1007/s11205-011-9906-6.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 32–45. https://doi.org/10.1023/A:1010933404324.

Buhalis, D. (2000). Marketing the competitive destination of the future. *Tourism Management*, *21*(1), 97–116.
Costanza, R., Hart, M., Posner, S., & Talberth, J. (2009). "Beyond GDP: The Need for New Measures of Progress." Pardee Paper No. 4, Boston: Pardee Center for the Study of the Longer-Range Future.
Crouch, G. I., & Ritchie, J. R. B. (1999). Tourism, competitiveness, and societal prosperity. *Journal of Business Research*, *44*(3), 137–152.
D'Urso, P., Alaimo, L. S., De Giovanni, L., et al. (2022). Well-being in the italian regions over Time. *Social Indicators Research*, *161*, 599–627. https://doi.org/10.1007/s11205-020-02384-x.
ETIS toolkit (2016). : http://ec.europa.eu/growth/sectors/tourism/offer/sustainable/indicators/index_en.htm.
Fielding, A., & Shepherd, J. W. (1973). The Sampling Stability of the Automatic Interaction Detector Technique. Contributed paper to the 16th Session of the International Statistical.
Flach, P. (2019). Performance evaluation in machine learning: The Good, the bad, the Ugly, and the Way Forward. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 9808–9814.
Gibson, B., Hassan, S., & Tansey, J. (2005). *Sustainability assessment: Criteria and processes*. Earthscan.
Global (1999). Code of Ethics for tourism.
Hotelling, H. (1992). *Relations between two sets of Variates in Breakthroughs in Statistics*. Springer.
Jiménez-Luna, J., Grisoni, F., & Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nat Mach Intell*, *2*, 573–584.
Lemmi, E., & Deri, M. G. (2020). A New Model for the 'Tourism Renaissance': The Case Study of the Tuscan Village of San Pellegrino in Alpe. *Almatourism - Journal of Tourism Culture and Territorial Development*, *11*(22), 19–43. https://doi.org/10.6092/issn.2036-5195/12345.
Lemmi, E., & Siena Tangheroni, M. (2014). "The Importance of Place Names in the Sustainable Tourist Development of the Inland Areas of Tuscany: Toponyms along the Via Francigena", in Proceedings of the XXIVICOS (International Congress of Onomastic Sciences). Barcelona, 5–9 september 2011, Barcelona, Department of Catalunya, 1869–1879.
Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model pre- dictions, in: Proceedings of the 31st international conference on neural information processing systems, pp. 44768–4777.
Max-Neef, M. (1995). Economic growth and quality of life: A threshold hypothesis. *Ecological Economics*, *15*(2), 115–118.
Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.
Talberth, D. J., Cobb, C., et al. (2007). *The genuine Progress Indicator 2006: A Tool for Sustainable Development*. Redefining Progress.
Vollmer, S., et al. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *Bmj*, *368*, l6927.

Authors and Affiliations

Najada Firza[1,2] · Laura Antonucci[3] · Corrado Crocetta[4] · Francesco Domenico d'Ovidio[1] · Alfonso Monaco[5,6]

✉ Najada Firza
najada.firza@uniba.it

[1] Department of Economics and Finance, University of Bari "Aldo Moro", Largo Abbazia Santa Scolastica 53, Bari, Italy

[2] Catholic University Our Lady of Good Counsel, Tirana, Albania

[3] Department of Clinical and Experimental Medicine, University of Foggia, Via L. Pinto, Foggia, Italy

[4] Department of Humanities Research and Innovation, University of Bari "Aldo Moro", Palazzo Ateneo - Piazza Umberto I 1, Bari, Italy

[5] Dipartimento Interateneo di Fisica M. Merlin, Università degli Studi di Bari Aldo Moro, Via

G. Amendola 173, Bari 70125, Italy

[6] Sezione di Bari, Istituto Nazionale di Fisica Nucleare (INFN), Via A. Orabona 4, Bari, Italy