



Gender and Age Matter! Identifying Important Predictors for Subjective Well-being Using Machine Learning Methods

Lasare Samartzidis^{1,2} · Marco Quatrosi³ · Angelika von Dulong⁴

Accepted: 30 May 2025
© The Author(s) 2025

Abstract

Subjective Well-Being (SWB) has emerged as a key measure in assessing societal progress beyond traditional economic indicators like GDP. While SWB is shaped by diverse socio-economic factors, most quantitative studies use limited variables and overlook non-linearities and interactions. We address these gaps by applying random forests to predict regional SWB averages across 388 OECD regions using 2016 data. Our model identifies 16 key predictors of regional SWB, revealing significant non-linearities and interactions among variables. Notably, the sex ratio among the elderly, a factor underexplored in existing literature, emerges as a predictor comparable in importance to average disposable income. Interestingly, regions with below-average employment and elderly sex ratios show higher SWB than average, but this trend reverses at higher levels. This study highlights the potential of machine learning to explore complex socio-economic systems, connecting data-driven insights with theory-building. By incorporating multidimensionality and non-linear interactions, our approach offers a robust framework for analyzing SWB and informing policy design.

Keywords Well-being · Random forest · Regions · Human development

1 Introduction

The pursuit of understanding, quantifying and promoting human well-being has become a paramount concern for international organizations, policy-makers, and governments (Durand, 2015; European Commission: Directorate-General for Employment, 2021; Mahoney, 2023; O'Neill et al., 2018). While economic activity measured through Gross Domestic Product

✉ Lasare Samartzidis
l.samartzidis@ioer.de

Marco Quatrosi
marco.quatrosi@unipa.it

Angelika von Dulong
angelika.v.dulong@gmail.com

¹ Leibniz Institute of Ecological Urban and Regional Development, Weberplatz 1, 01217 Dresden, Germany

² Faculty of Business and Economics, Technische Universität Dresden, Helmholtzstr. 10, 01062 Dresden, Germany

³ Department of Law, University of Palermo, Palermo, Italy

⁴ Aurora Energy Research, Berlin, Germany

(GDP) is the focal point for policy-makers to approximate well-being, it is increasingly acknowledged that it ignores the entwined nature of economic, social, and environmental determinants that collectively shape what constitutes the perception of a good life for people (Boarini et al., 2006; Diener & Suh, 1997; Durand, 2015). Not only is GDP a poor measure of how economic activity translates into well-being (Stiglitz et al., 2017), it can even exhibit a partially negative relationship with certain aspects of well-being, such as environmental health and leisure time (Costanza et al., 2009, 2014; Otero et al., 2020). The critique of the reliance on GDP received a renewed resurgence in the public debate since the work of the Stiglitz-Sen-Fitoussi Commission issued by the French Presidency (Bleys, 2012; Brand-Correa et al., 2022; Hoekstra, 2019; Stiglitz et al., 2017).

Three main alternative approaches emerged to overcome GDP-dependence: 1) combining economic, monetary factors with other monetized objective factors via indices (Cook & Davíðsdóttir, 2021; Talberth & Weisdorf, 2017; Kenny et al., 2019; Hayden & Wilson, 2018; Zhang et al., 2018), 2) promoting subjective measures of well-being as a policy target (Kahneman & Krueger, 2006; Cavalletti & Corsi, 2018), and 3) weighted composite indices of objective factors and subjective measures (Barrington-Leigh & Escande, 2018; Diener & Suh, 1997; Durand, 2015). Objective factors cover a wide range of elements, including air and water quality, availability of land and amenities, political engagement, housing, healthcare, and security. These components influence the so-called exterior dimension of well-being, which is related to the fulfillment of a certain set of psychological, material, and social needs, as well as the degree of freedom and resources to meet individual desires (Atkinson & Mourato, 2015; Diener & Suh, 1997). While objective indicators provide valuable insights, they are incomplete without considering the lens through which individuals perceive and experience their lives. Subjective well-being (SWB) encompasses self-reported life satisfaction, the realization of the individual's potential, and personal evaluation of one's life. All those components pertain to the interior dimension of well-being (Atkinson & Mourato, 2015; Chaaban et al., 2016). Descriptive social indicators can partially capture this subjective perspective, yet the actual interplay between objective and subjective dimensions remains challenging to empirically delineate (Atkinson & Mourato, 2015; Diener & Suh, 1997; Dolan et al., 2008; Michalos et al., 2011; Stiglitz et al., 2017).

Insights from research on the objective and subjective influences on SWB are useful in orienting policies to promote SWB. However, effectively capturing and comparing SWB across diverse contexts presents a formidable challenge. Its multidimensionality implies that no single indicator can provide a comprehensive assessment. Attempts to condense the myriad dimensions into composite indicators can facilitate cross-country comparisons but risk oversimplification and loss of nuanced information (Ciommi et al., 2017). Conversely, employing many indicators for analysis may yield a wealth of information but hinder comparability across countries and over time (Durand, 2015). The intricate web of individual, household, and societal levels further complicates the task of measurement (Durand, 2015). The OECD's "Better Life Index," for example, encompasses a wide array of dimensions spanning material conditions and quality of life. Yet, even such holistic frameworks may fall short of capturing the full spectrum of well-being (Chaaban et al., 2016; OECD, 2019). Moreover, the existing analytical frameworks often neglect spatial interdependencies and regional variations, a crucial omission given the profound influence of local context on SWB (Ballas & Thanis, 2022; Hoogerbrugge et al., 2022; Pacione, 1982, 2003).

The challenge lies in reconciling these different concerns and determining when and how to measure specific aspects of SWB while simultaneously considering its multidimensionality and non-linearity. The elements of multidimensionality and non-linearity potentially pose difficulties for the scientific method solely based on human induction, as human thinking

tends to follow linear reasoning (De Langhe et al., 2017). For example, the existing literature stresses the multidimensional nature of SWB but typically limits the number of variables analyzed to a modest range and employs linear regression models (Balestra et al., 2018; OECD, 2017). In the era of big data (Luhmann, 2017) and recent advances in data analytics, the question arises whether a more data-driven approach can offer novel perspectives between SWB and its exterior dimensions, contributing to the step of exploration and consequent induction in the theory building process. More specifically, machine learning's ability to recognize non-parametric patterns in large amounts of data when predicting outcomes may help to formulate novel and testable hypotheses, an approach explored in other fields of social science already (Shrestha et al., 2021; Van Lissa, 2023).

This paper's main contribution is to offer a workflow that generates hypotheses based on machine-learning informed induction to the scientific method when studying outcomes of complex socio-economic systems, such as SWB. We argue that machine learning methods are better suited for these exploratory tasks than traditional econometric methods, as they inherently allow to account for multidimensionality based on flexible functional forms including non-linearities and interactions among predictors¹, while at the same time enable the inclusion of a large set of predictors (Mullainathan & Spiess, 2017). We develop a multiple step workflow to achieve this. First, we extend the predictors used in the OECD regional well-being database, those defining the Better Life Index, with available predictors from other, more general OECD regional databases. Second, we employ machine learning to identify the most important predictors of SWB across all OECD regions. Third, we investigate the functional forms and interactions in our final machine learning model via methods of interpreting machine learning models (Molnar, 2022) and thoroughly compare our gathered insights to the traditional literature on the topic. Finally, we derive hypotheses for future causal studies on the topic. To our knowledge, no prior studies employ machine learning techniques to investigate regional well-being.

By applying the developed workflow to SWB, our study seeks to answer the following questions: How similar or divergent are the results of algorithm-based estimates of SWB at the regional level to previous studies and current frameworks, such as the OECD Better Life Index? Are there non-linear relationships and interactions between SWB and its predictors? Are there any novel hypotheses that emerge out of our results?

Our results highlight that the most important predictors of SWB are related to socio-economic and subjective factors, such as perception of corruption and perception of social support. Those findings are in line with previous contributions in the literature (Ali et al., 2020; Cavalletti & Corsi, 2018; De Neve & Sachs, 2020; Dolan et al., 2008; Glatz & Eder, 2020; Peiró-Palomino, 2019; Puntischer et al., 2015; Schmidt et al., 2023; Steiner et al., 2015). However, compared to the established literature, our results also offers a novel important predictor of regional SWB in the sex ratio among the elderly population. Given this novelty, we derive hypotheses for future causal research on this predictor.

Through its regional focus, our study offers a diverse perspective on SWB which holds crucial implications for regional policy-making. Most of the studies at the regional level (Di Paolo & Ferrer-i-Carbonell, 2022; Peiró-Palomino, 2019; Puntischer et al., 2015) have analyzed SWB considering only one or two dimensions between economic, social, and environmental. This study provides a multidimensional framework considering all three dimensions. However, the representativeness of our results might be limited by the choice of our sam-

¹ Variables are usually referred to as predictors in the machine learning literature. For this reason, we will stick to the term predictor for the rest of the manuscript.

ple. The data we employed for the analysis come from OECD regions. Hence, we cannot generalise our results.

The use of machine learning techniques in this work aims at spotting new angles of the relationship between SWB and relevant predictors. Machine learning algorithms are constructed to deal with the most common pitfalls of standard econometrics. However, the algorithm we have chosen for the sake of this analysis does not deliver any causal claim. To further clarify this point in this work we will refer to “predictors” or “features” and not “determinants” or “covariates” of SWB (Bartram et al., 2024).

The remainder of the paper is structured as follows: In Section 2, we review the literature on the determinants of SWB. In Section 3.1, we present our methods and describe the data in Section 3.2. Section 4 provides results and their discussion, whereas, in Section 5, we draw conclusions.

2 Literature Review

We categorise the literature on subjective well-being (SWB) into three types, individual, national and, regional. These categories have subsequently emerged and influence each other. Early studies analysed SWB on the individual scale, mostly analysing how SWB changed across the lifespan of individuals (Kahneman & Krueger, 2006). Later, data collection on SWB at the individual level strove for representative national scale samples to compare how national policies can shape SWB and possibly use it as an alternative to GDP. Only more recently, representative regional scale samples are available to study. The studies on the individual scale established the association of SWB with absolute income, relative income, health, personal and community relationships, and employment status (Dolan et al., 2008; Steiner et al., 2015). More recent studies on individual SWB showed the inverse effects of regional (Di Paolo & Ferrer-i-Carbonell, 2022) and national (Glatz & Eder, 2020) unemployment rates, positive effects of social and institutional trust (Glatz & Eder, 2020), the impact of household structures and associated feelings of loneliness (Mínguez & Vírveda, 2019), and reactions of SWB to extraordinary event like the financial crisis of 2008 (Deaton, 2012). Furthermore, the literature explored place-specific determinants on individual SWB, finding positive effects of green and natural environments, with heterogeneous effects for different gender and age groups (MacKerron & Mourato, 2013), the impact of spatial structures (Hoogerbrugge et al., 2022), a positive effect, similar in size to income, for the regional bird species richness (Methorst et al., 2021), and living in a city that hosts the European Capital of Culture (Steiner et al., 2015).

Insights from the individual scale have shaped the initial explanations of variations between countries and informed the first analyses of country differences in SWB. However, the literature studying national variations also extended the scope of analysis in SWB towards more context specificity. Different studies on the national scale are not easily comparable, as they include various sets of countries, indicators, and methodologies in their analyses. Methodological differences between studies can lead to contradictory results in the literature, as shown for the link between inequality and SWB (Schneider, 2016). The positive relationship with SWB established for GDP and income per capita on the individual scale was confirmed at the national scale (Ali et al., 2020; Cavalletti & Corsi, 2018; De Neve & Sachs, 2020; Schmidt et al., 2023). Results for health indicate positive correlations with SWB (Cavalletti & Corsi, 2018; De Neve & Sachs, 2020) and variations in health contribute to a large share of variance of SWB (De Neve & Sachs, 2020). However, regression results in Cavalletti and Corsi (2018) suggest no to a slightly negative relationship of a composite indi-

cator of health with SWB. The social reality faced by individuals explains differences in SWB between countries well. The corresponding factors are the level of spending on social protection and social safety nets, both with a positive relationship, the share of the population over 65 years with a negative relationship (Schmidt et al., 2023), and demographic differences more generally (Cavalletti & Corsi, 2018). The importance of employment for individual well-being translates into a negative relationship between unemployment and SWB on the national scale (Ali et al., 2020; Cavalletti & Corsi, 2018; Schmidt et al., 2023). Also, larger shares of the resource-based sectors in the total economic activity of a country predict lower SWB scores (Ali et al., 2020).

The literature on estimating regional averages of SWB is scarce. However, some studies analysing the individual scale include regional characteristics in their estimations. Prior work confirms results from the national scale estimating a negative relationship of regional unemployment rates with regional averaged SWB (Puntscher et al., 2015) and individual SWB (Di Paolo & Ferrer-i-Carbonell, 2022). Furthermore, using the same data as this study, regions with a growing population, a lower share of the population above 65 years, higher levels of ethnic fractionalization, higher regional expenditure, and higher institutional quality have been estimated to have higher regional SWB (Peiró-Palomino, 2019).

In conclusion, the prior literature on SWB offers a point of reference to compare our data-driven results. We seek to find the established predictors from the literature reflected in our final set of predictors from the data-driven reduction process. Aligning results will prove our confidence in our methodological approach. Furthermore, they provide the baseline for possible novel insights coming from our analysis.

3 Methods & Data

3.1 Methods

Our methodological approach reflects the multidimensionality of well-being and the exploratory nature of our research questions. Thus, two aspects are of particular interest. First, we want to include as many predictors as possible for the data-driven analysis. Second, we do not wish to make any assumption about the functional form between predictors and subjective well-being (SWB), explicitly allowing for non-linearities and interactions among predictors. These two characteristics are well captured through algorithm-based machine-learning approaches and their main advantage compared to regression analyses.

The methods described below reflect the step-wise approach of our analysis. First, we introduce the random forest (RF) algorithm (see Section 3.1.1) and methods for its interpretability (see Section 3.1.2). Then, we describe the process of attaining our final RF model by selecting the optimal group of predictors (see Section 3.1.3).

3.1.1 Random Forests

While machine learning algorithms and their consequent prediction models are framed as black-box models due to the inability to understand how they reached their results (Breiman, 2001b), recent advances in the interpretability of these models allow to reconstruct their innate functioning (Molnar, 2022). However, there are differences between different machine learning algorithms in their interpretability. Due to their complexity, (deep) neural network may still be seen as black box models, whereas there are many available techniques to interpret tree based models (Molnar, 2022).

For this study, we choose the tree-based Random Forest (RF) algorithm which is one of the most widespread machine learning approaches (Breiman, 2001a). This choice is mainly based on two aspects: its ability to include non-linearity and interactions when predicting outcomes and the wide availability of methods to interpret RF models (Molnar, 2022).

As an ensemble regression technique, RF utilises many regression trees to predict an outcome variable y_i . Tree-based methods make predictions through a step-wise approach (James et al., 2013). Predictions are made by splitting the sample based on a predictor's mean value and taking the average of the outcome variable in the resulting sub-samples to predict the outcome variable. The next step splits the sub-samples again by the mean of a predictor in that sub-sample. The maximum number of splits is usually restricted; otherwise, the resulting model will overfit the data. The predictor used to split the (sub-)sample at each split is chosen based on the maximum decrease in the residual sum of squares (RSS) of the resulting sub-samples prediction. One consequence of these splitting rules is the non-linear and interdependent functional relationships between the predictors and the outcome variable. Single regression trees are easy to interpret but their predictions compare poorly to predictions made from many ensembled trees. Through randomization of the used predictors, the method of RF decorrelates trees in the ensemble and enhances the out-of-sample predictions (James et al., 2013).

There are different parameters to choose before running a random forest algorithm, including the minimum number of nodes in a tree, the number of trees and the number of randomly chosen predictors in each split. We choose the optimal configuration of these parameters through a hypertuning process. Hypertuning is the process of running the algorithm with different parameter combinations and selecting the resulting model that has the lowest mean squared error (MSE) in predicting the outcome variable.

Based on the exploratory nature of our work, we will focus on finding the most relevant predictors of SWB and explore their functional forms and interactions.

3.1.2 Model Interpretability

There are several methods to understand the behavior of RF models and to learn about their included predictors, including the importance of single predictors, their functional forms and interactions. A predictor X_j is considered important in predicting the outcome variable Y if breaking a link between X_j and Y through randomizing the values of X_j increases the prediction error, measured in the MSE (Breiman, 2001a).

One of the strengths of the RF model is its ability to find different functional forms in the relationship between the outcome variable and predictors. A common way to investigate the functional form is through partial dependency plots (PDP) that plot average predictions over the marginal distribution of a predictor. However, this is problematic as PDPs assume the independence of the corresponding predictor (Molnar, 2022, Chapter 8.1). In other words, PDPs assume uncorrelated predictors, and their calculation involves unreasonable combinations of the corresponding predictor values with other predictors used in the RF model that are not found in the original data. In our case, computing PDPs would include imaginative and unrealistic combinations of regional characteristics. To overcome this issue, we employ Accumulated Local Effect (ALE) plots, which show how the machine learning model predicts the outcome variable over different values of a predictor (Apley & Zhu, 2020). ALE plots average and *accumulate* local effects over the conditional distribution. i.e. the distribution of a predictor conditional on the values of other predictors. In other words, ALE plots calculate changes in the predictions, not the prediction itself, over the conditional distribution

within a small window of values and, thus, consider only realistic combinations of regional characteristics found in the data.

We compute the interaction effects between predictors using the H-statistic (Molnar, 2022; Friedman & Popescu, 2008). It is based on the partial dependency function, which measures how the prediction changes over the range of values of a predictor. The H-statistic measures how much of the variance in the partial dependency function of a predictor in the model can be explained by the change in another predictor or all other predictors in the model. The statistic ranges from 0, no interaction, to 1; all the variance can be attributed to the interaction. We compute the H-statistic for the total interactions strength of a predictor and for the two-way interactions between two predictors.

3.1.3 Variable Selection

The final number of available predictors from the OECD databases is 134 (see Section 3.2 for details) which necessitates to reduce this number to improve the performance and interpretability of RF models. To avoid a researcher bias and to be consistent with our data-driven approach, we apply machine learning to optimally reduce the number of predictors.

One recently introduced method is to select variables via the least absolute shrinkage and selection operator (LASSO) before running the RF algorithm (Tan et al., 2018). However, we chose to select the variables for our final RF model via a method based on RF, too. The rationale for this choice is related to OLS being the inherent optimization function used in the LASSO method (Tibshirani, 1996). Our main motivation in this study is to employ machine learning to account for the non-linearities and interactions when predicting the highly multidimensional concept of SWB. Thus, not accounting for this when choosing our final set of predictors would have introduced a linearity bias to our final variable selection method.

We reduce the number of predictors included in the final RF model through recursive feature elimination (Gregorutti et al., 2017). This data-driven technique is mainly employed when there are many predictors and a relatively small number of observations. Recursive feature elimination exploits an RF algorithm that selects predictors according to their permutation importance. Furthermore, the selection criterion based on variable importance also considers correlations. Correlated predictors show smaller importance scores and, consequently, are more likely to be excluded in the process of recursive feature elimination (Gregorutti et al., 2017). The recursive feature elimination ranks all the predictors according to their permutation importance. After that, the algorithm eliminates a fraction, in our case 20%, of predictors deemed least important by the computation. Then, a new forest is trained on the remaining predictors. The process is repeated until only one predictor remains. We compute the so-called Out-Of-Bag error for each RF model in the process, which is the error when predicting out-of-the sample used to train the RF model, and choose the final set of predictors based on the lowest Out-Of-Bag error of the corresponding RF model (Breiman, 2001a; Gregorutti et al., 2017).

3.1.4 Workflow of the Analysis

We will conduct our analysis on a script, a consequential step-wise approach, that we call workflow, corresponding to the code provided. After compiling the dataset (see Section 3.2 for details), we use the variable selection method presented in the previous subsection. Then, using the most appropriate set of predictors, we randomly divide the sample into a training

set with 75% and a test set with 25% of its observations. After hypertuning to find the best combination of parameters, we build our final random forest model.

With the final established random forest model, we compare its performance of predicting the test set of SBW values to those of an Ordinary Least-Squares (OLS) model and XGBoost, another widely used tree-based algorithm, Chen and Guestrin (2016) by the coefficient of determination R^2 and the root mean squared error $RMSE$. Finally, we apply the model interpretability methods to learn about the RF model's behaviour and gather knowledge on the most important predictors, their functional forms and their interactions when predicting SWB. We focus our investigation on the model's predictors that are most important and show the strongest interaction effects. Furthermore, we focus on those results differing from the established literature on SWB. This focus corresponds to this study's primary motivation for using machine learning. To ensure the robustness of our results, we contrast results from the RF estimation to those of a XGBoost model and show the results in the [Appendix](#).

3.2 Data

The well-being and regional databases offered by the OECD allows the analysis of harmonized regional data across different continents. Our data stems exclusively from the 2016 edition of OECD Regional Statistics (OECD, 2019). The core of our dataset is the *Regional well-being (Edition 2016)* database, consisting of 14 indicators covering material conditions (jobs and housing), quality of life (education, health, environment, safety and access to services), and mean subjective life satisfaction scores for 388 OECD regions². Regional data on SWB stem from the Gallup World Poll, gathered between 2006 and 2014, that were reweighted for regional representativeness (OECD, 2016). To extend the variables for the RF algorithm and to test whether there are predictors of SWB different to those gathered in the OECD Regional Well-being Statistics, we combined it with five other OECD regional databases on regional demography, regional economy, regional education, regional labor markets, and regional social and environmental indicators. Then, we included the most recent available data points from our database that were no older than 2006. This decision is based on the years when the data on subjective life satisfaction originates.

The combined database includes 362 indicators for 388 OECD regions. However, there are missing entries for some regions and indicators. As we needed a dataset with no missing entries to run the RF algorithm and wanted to dismiss any possible selection bias of the non-reporting regions, we chose the indicators that have no missing entries for our 388 regions and ended up with 134 different indicators, including subjective well-being.³ The data collected have different units of measurement. To allow comparability and reduce the computational time, we centered all data and scaled it by the respective standard deviation with $X_{scaled} = \frac{X_i - \bar{X}}{\sigma(X)}$ before running the RF algorithm.

Our data represent different social, environmental, and economic aspects of people's living environment. They includes, among others, data on air quality, labor statistics, life expectancy, income, and innovative capacities. The outcome variable for our analysis is a

² The database gathers information for 391 OECD regions. However, we erased three regions, as their information for all 14 well-being indicators was incomplete. These were Helsinki-Uusimaa in Finland, Los Rios, and Arica Y Parinacota in Chile.

³ From this follows a trade-off between including as many different indicators or regions as possible in a final estimation. For example, excluding the 120 regions with the highest number of missing entries across all indicators increases the number of indicators with full information to 202. Still, it reduces the number of available regions to 268.

measure of SWB. In this case, subjective well-being is the average score of self-evaluated life satisfaction for each region. The index ranges from 0 to 10, where zero represents the minimum, and ten is the maximum level of life satisfaction reported. The average score at the regional level is based on estimates provided by the single National Statistical Offices. Those estimates come from broadly comparable questions that follow the *OECD Guidelines on Measuring Subjective Well-Being* issued in 2013.

4 Results

4.1 Selection of Predictors

The first step of our analytical strategy involves the selection of predictors. We thus ran the feature elimination algorithm on the 134 indicators collected from the OECD database. Figure 1 depicts the OOBs for the process of recursive feature elimination. For the first rounds of elimination, going left to right, the OOB error steadily decreases along a decreasing number of predictors included in an RF model to predict SWB. Then, there is a minimum when the RF model uses 20 predictors and an OOB error of 0.230. After that minimum, the error stays relatively constant before increasing steeply when the RF model uses only a few predictors for prediction. We continue our analysis with the 16 predictors solution, as this solution does not increase the OOB much but decreases the final number of predictors to interpret and, thus, reduces the possible correlation between predictors. Table 1 shows the summary statistics for our final predictors. It is important to note here that our predictors include only nine of the original fourteen OECD Regional Well-being predictors, and seven predictors stem from other regional databases of the OECD.

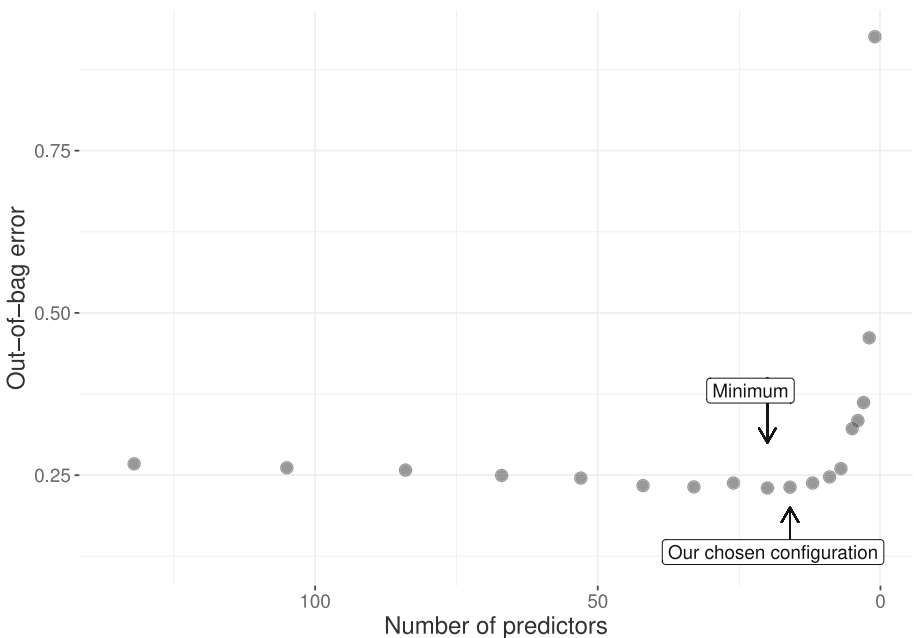


Fig. 1 Results of the recursive predictor selection process. In each step, a number of predictors remain along with the corresponding OOB error of an RF model using these predictors to predict subjective well-being

Table 1 Summary statistics

Predictors	Database	Mean	SD
Predictor			
Employment rate	Wellbeing	66.34	10.26
Sex ratio 65+ (male/female)	Demographic	77.88	8.66
Disposable income	Wellbeing	17909.62	10393.23
Unemployment rate	Wellbeing	8.63	5.98
Perceived social network support	Wellbeing	88.89	7.69
Subjective perceived corruption	Wellbeing	60.47	19.69
Population density growth index (2001 vs. 2014/2015)	Demographic	109.32	12.36
Rooms per capita	Wellbeing	1.75	0.56
Air pollution (PM2.5)	Wellbeing	10.57	5.29
Death rate (male)	Demographic	8.77	2.30
Population growth index (2001 vs. 2014/2015)	Demographic	109.53	11.30
Child-to-woman ratio	Demographic	25.71	6.03
Youth dependency rate	Demographic	27.87	8.09
Labour force with at least secondary education (in %)	Wellbeing	74.15	18.96
Household broadband access	Wellbeing	69.84	18.99
Sex ratio total population (male/female)	Demographic	97.02	3.68
Outcome variable			
Subjective Wellbeing	Wellbeing	6.66	0.81

Notes: Employment rate is the share of people currently employed out of the workforce (aged 15-64). The Sex ratio 65+ (male/female) is the ratio of male over female population aged more than 65. Disposable income is the income available to households from wages and salaries, self-employment and unincorporated enterprises, pensions and other social benefits, and financial investments. The Unemployment rate is the share of the population of the working-age population who are without work, are available for work, and have taken specific steps to find work. Perceived social network support represents the share of the population who reported having relatives or friends they can count on. Subjective perceived corruption is a score ranging from 0 if the regional institutions are perceived as highly corrupt to 100 if they are perceived as very clean. The Population density growth index represents the ratio between the growth rate of the population in 2001 and the growth rate of the years 2014/2015. Rooms per capita is the average number of rooms in a dwelling divided by the number of persons living in a dwelling. Air Pollution (PM 2.5) is the average level of PM 2.5 ($\mu\text{g}/\text{m}^3$) experienced by the population of a region. Death rate (male) is the average number of deaths among the male population divided by the number of total deaths. Child-to-woman ratio represents the average number of children aged 0-4 years and the number of females aged 15-49. Youth dependency rate is the number of persons aged less than 15 as % of the population aged 15-64. The Labour force with at least secondary education (in %) is the share of working-age people with an upper secondary education level. Household broadband access is the average share of households with broadband internet connection. The Sex ratio total population (male/female) is the share of male over female population in a region. Subjective well-being is the average score of self-evaluated life satisfaction for each region, ranging from 0 (minimum) to 10 (maximum)

The resulting predictors can be expected to be correlated. To check for any correlations that need to be considered when interpreting our results, we provide a visualised correlation matrix in the appendix (see Fig. 5).

4.2 Random Forest Estimations

After hypertuning the parameters of the model⁴, we use the 16 predictors to predict subjective well-being. Utilizing our RF model compared to an Ordinary Least-Squares (OLS) model

⁴ The final model parameters are six randomly chosen predictors to possibly split at in each node, two as the minimum number of nodes, and 1085 trees.

and a hypertuned XGBoost model with the same predictors increases the share of variance predicted, called the coefficient of determination, R^2 from $R^2_{OLS} = 0.666$ and $R^2_{XGBoost} = 0.731$ to $R^2_{RF} = 0.780$ and decreases the root mean squared error $RMSE$ from $RMSE_{OLS} = 0.621$ and $RMSE_{XGBoost} = 0.552$ to $RMSE_{RF} = 0.521$. Thus, our RF model yields a significant improvements compared to a linear OLS estimation and slightly better prediction compared to a XGBoost model. We attribute the improvements compared to OLS to including non-linearity between the outcome variable and the predictors and interactions between predictors.

Figure 2 shows the predictor importance for all employed predictors as measured by the loss of mean squared error (MSE). The importance of the predictors steadily decline. However, with a significant gap, the most important predictor in the model is the employment rate. It is followed by subjective perceived corruption, disposable income, subjective perceived social support, sex ratio of the elderly population, and the unemployment rate. As mentioned in the Section 3.1.3, predictors with higher correlations are punished when computing their importance score. We see this effect when considering the correlation matrix visualised in

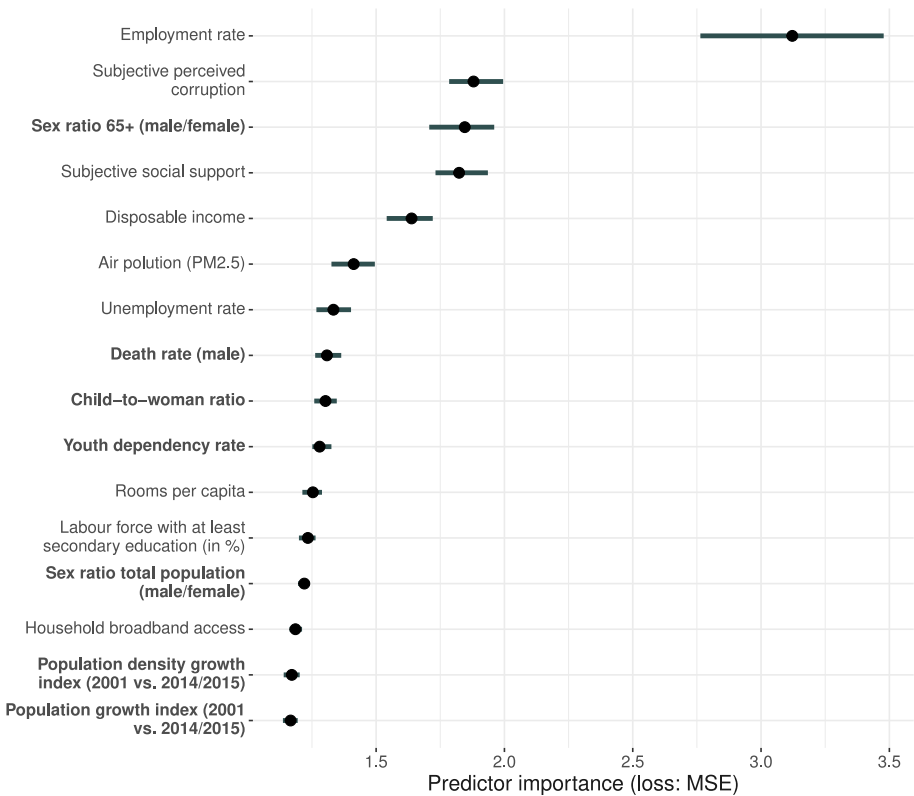


Fig. 2 Importance of all predictors in the final model. The importance of a predictor is evaluated based on a higher loss of mean squared error (MSE) for shuffled values when predicting SWB. The dot represents the median value for 100 repetitions of the shuffling process, and the bar depicts the range of its 5% and 95% quantiles. Variables that do not stem from the well-being database are marked in bold

the appendix in Fig. 5. The male death rate and the household broadband access show particularly high correlations with several other predictors. Therefore, their importance scores are relatively low. Furthermore, through randomising which predictors are used in each split, results from the random forest algorithms are not as susceptible to correlations of the included predictors as linear regressions.

The ALE plots of the predictors complement their importance as they show the functional form of the relationship between SWB and the main predictors. Figure 3 shows the results of the ALE plots for each predictor. The predictors are ordered based on their importance in predicting SWB (Fig. 2). Most predictors show an S-curve-like shape that signifies certain levels of those predictors up to which SWB does not further increase or increase with diminishing magnitude. The ALE plot for subjective social support, the male death rate, and the child-to-woman ratio have slight U-shaped curves.

Figure 4 shows the emerging interactions in the RF model that are useful to assess the influence of interactions among predictors in the model on SWB predictions. Due to all

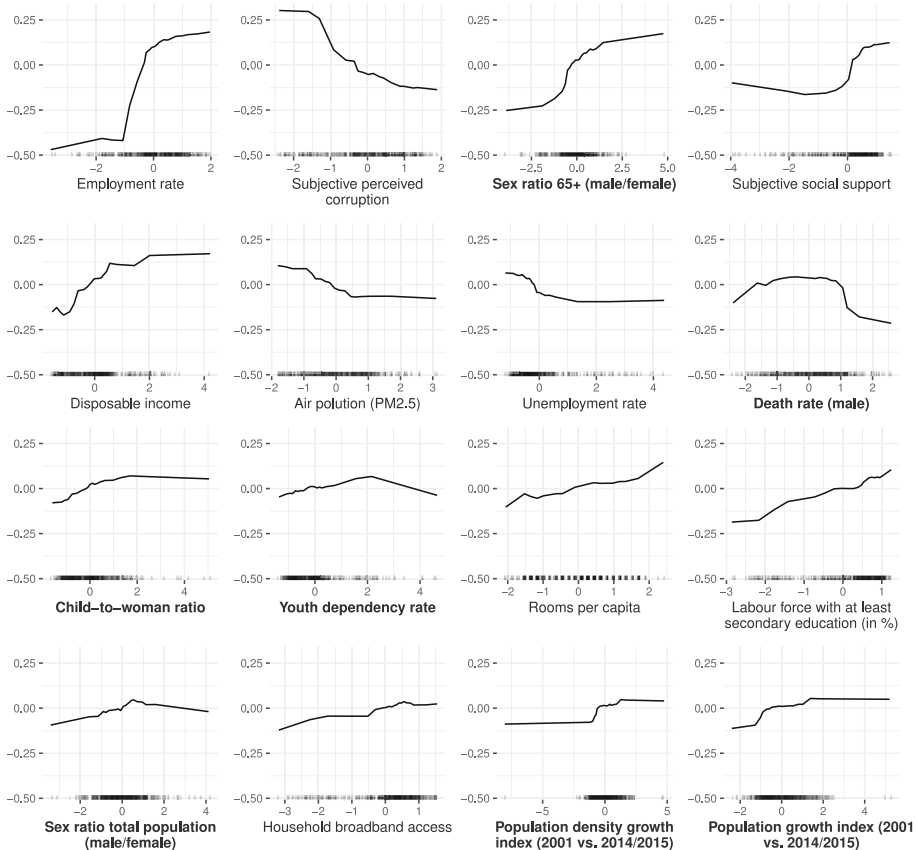


Fig. 3 ALE plots of all predictors. The y-axis depicts changes in the predicted Accumulated Local Effects (ALE) for subjective well-being (SWB) at different values of the corresponding predictor. The x-axis is scaled in units of the standard deviation of the corresponding predictor, with zero representing its sample mean. Steeper ALE curves, indicating more substantial changes in the prediction as the value of the predictor changes, suggest that the predictor is more important. Variables that do not stem from the well-being database are marked in bold

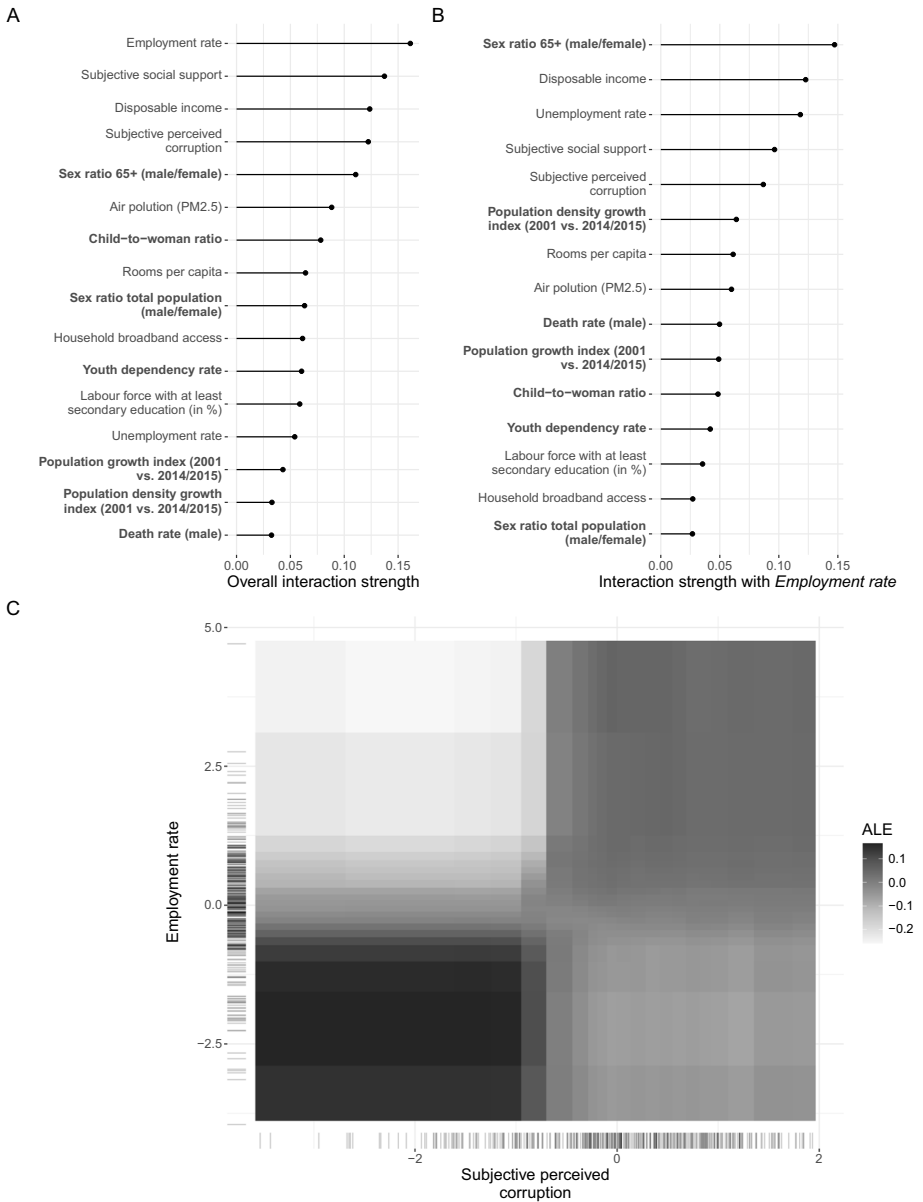


Fig. 4 Selection of interactions in the model. Panel A: shows the overall interaction strength of all predictors in our final prediction model, measured by the H-statistic. Panel B: shows the contribution of each two-way interaction with the employment rate to the variance from the previous step (Panel A); Panel C: shows the functional form of varying values for the interaction of employment rate and sex ratio of people over 65, i.e., the share of male over female for people aged above 65 years, through the Accumulated Local Effect (ALE) of predicting SWB. Darker shades of grey depict higher predicted SWB values. Variables that do not stem from the well-being database are marked in bold

the possible interactions among the predictors, we only engage with the major interactions in our model to showcase the potential of such an analysis. Among all the predictors, the employment rate shows the highest interaction effect. The interactions with other predictors

account for more than 15 % of its prediction variation of SWB (Fig. 4A). As defined in our workflow in Section 3.1.4, we further investigate the interactions of the employment rate. Figure 4B shows the two-way interactions of the employment rate. The largest two-way interactions are the sex ratio of people over 65 that accounts for around 8 % of the prediction variation of the employment rate, the disposable income, and the unemployment rate (both around 8 %).

Among all the two-way interactions of the employment rate, the sex ratio of people over 65 is the strongest. Moreover, the latter is a novel predictor that emerged from our data-driven variable selection procedure. For both reasons, we decided to further investigate the functional form of this two-way interaction to visualise how the predicted SWB changes in response to changes in the employment rate and the sex ratio of people over 65 (Fig. 4C). The figure shows that predicted SWB is lower in regions with high employment rates and a lower-than-average elderly sex ratios, as represented by the upper-left quadrant. The same applies to the bottom-right quadrant, where a higher-than-average elderly sex ratio combined with low employment predicts lower values of SWB. Conversely, in the bottom-left quadrant, a lower-than-average employment rate and a lower-than-average elderly sex ratio are associated with higher SWB. The same is true for the upper-right quadrant, where higher-than-average employment and a higher-than-average elderly sex ratio are associated with higher predicted SWB.

We check whether our results depend on the RF algorithm and contrast them with the results of the XGBoost algorithm. We present the contrasting results in the appendix. They show that our main results hold for both algorithms, namely the importance of the employment rate, its strong interaction with the elderly sex ratios, and the functional form of this interaction.

5 Discussion and Conclusion

The subjective perception of well-being reflects the multidimensionality of the "good life" in a society. In this study, we predict subjective well-being (SWB) in OECD regions by socio-economic and environmental factors with machine learning techniques, thus accounting for its highly multidimensional nature. Our results indicate that these data-driven approaches can meaningfully reduce many factors, what we call predictors, to an essential set for predicting outcomes of a complex socio-economic system. The data-driven reduction is meaningful, as the final predictors correspond well to established frameworks while still offering novel insights previously unexplored in the literature. Also, our results reveal non-linear relationships between predictors and SWB and the interactions between predictors.

In line with previous contributions in the literature, our results highlight socioeconomic predictors as crucial to SWB. However, a novel insight into our approach stresses the importance of the regional gender composition of older people. Nonetheless, a residual component of SWB remains elusive, emphasizing the need for further investigation. Nine out of 16 predictors that emerged from our data-driven approach stem from the well-being database of the OECD. The previous literature established the importance of the employment and unemployment rate (Ali et al., 2020; Cavalletti & Corsi, 2018; Dolan et al., 2008; Puntischer et al., 2015; Schmidt et al., 2023; Steiner et al., 2015), subjective social network support and perceived corruption, as proxied by social and institutional perception/trust (Glatz & Eder, 2020; Peiró-Palomino, 2019), and disposable income (Ali et al., 2020; Cavalletti & Corsi, 2018; De Neve & Sachs, 2020; Schmidt et al., 2023), all among the most important predictors

in our machine learning model. Also, the general importance of demographics on SWB has been established in the previous literature (Cavalletti & Corsi, 2018) and connects to the seven predictors from the demographic OECD regional database in our final model.

A previously unexplored aspect in the literature relates to our finding that the sex ratio among the elderly population is as significant a predictor of the regional average SWB as the average disposable income. Higher inequalities in sex ratios indicate that more elderly females live alone, possibly facing financial stress due to pension-gender gaps (Ebbinghaus, 2021). Also, this translates into higher care demand from other, mostly female members of society, decreasing their financial and time resources and, thus, negatively affecting general well-being in a region. This interpretation is also in line with our findings on the interaction between the employment rate and the sex ratio among the elderly, which shows that our model predicts higher SWB in regions with a lower-than-average employment rate and a lower-than-average sex ratio among the elderly compared to both average levels, as the lower employment rate translates into more available time resources to care for the single elderly. We view these findings with their tentative interpretations as the results from the here presented inductive exercise and as hypotheses to be substantiated by future causal studies.

By reaching well-aligned results with the previous literature despite using a different method, we consider our data-driven approach a fruitful methodological contribution to the study of SWB, especially to untangle the multidimensional aspects of the complex socio-economic concept of SWB to raise novel research questions. Furthermore, we advance the field by raising novel hypotheses for future causal studies.

An important point of reflection relates to the general usage of data-driven approaches and how their results crucially hinge upon the input data to this process. We tried to avoid a selection bias in the selected regions and only included predictors covering the full range of regions for which SWB averages were available. Hence, the input data to the machine learning algorithm only included the well-being database and most entries from the demographic database, restricting our possible results to these two databases. Furthermore, few prior studies analysed regional averages of SWB, and compared to country scale studies, indicators on the regional scale show higher variation compared to national averages, thus allowing a more nuanced analysis of indicators that shape collective SWB. We believe the more nuanced regional input data combined with the exploratory nature of our analysis including non-linearities and interactions, contributed greatly to our novel findings of predictors for SWB.

The empirical strategy chosen for this work is well-suited to deal with analyses involving a high number of features. The process to deal with this higher amount of variables is fully data-driven with limited interventions by the researchers. In general, and in particular, for the relationship between SWB and its predictors, the choice of which variable to include in the model might influence results significantly (Bartram, 2024; Bartram et al., 2024). The issue has also been raised with respect to variables related to gender and their interactions with other socio-economic controls (Bartram, 2022). We thus acknowledge that our results are influenced by the set of variables used in the model and that another setting would have led to different outcomes.

Our work naturally comes with limitations. First, our sample includes regions in OECD countries, consisting of only a tiny subset of all countries worldwide and representing mostly industrialized countries. Consequently, we cannot easily generalize our results to regions in low-and middle-income countries. A more general limitation concerns the framework of our analysis, which does not allow for causal claims. The machine learning techniques

chosen only serve as a predictive tool. Nevertheless, the insights of this study serve as a valuable source of different hypotheses for future causal works, that could include methods of instrumental variables and propensity score matching. Further, even though our predictors were chosen based on the predictor importance measure, which punishes highly correlated predictors, our final list still contains correlated predictors, and future studies should consider this limitation when referencing our results.

From the current study, we identify several promising directions for future research. Future work should include regional data from low- and middle-income countries in our analysis, which would contribute to their generalisability. However, the lack of harmonised data is a caveat to such an effort. Though machine learning models derive results for prediction purposes, future work needs to establish our results in a causal framework, especially regarding the link between SWB and the sex ratio of the elderly population. Our results align well with prior research and are an encouraging sign for applying machine learning approaches to study socio-economic systems. For instance, the study of individual SWB and its socio-economic determinants could utilise the workflow of the present study to generate novel, previously unexplored hypotheses. Finally, the empirical literature on SWB would benefit from considering spatial interlinkages when estimating SWB, as socio-economic outcomes tend to show spatial dependencies across neighbouring regions.

We draw specific tentative policy implications regarding our results on the sex ratio of the elderly population and general implications from our methodology for studying complex socio-economic outcomes. Policy implications from our results on the sex ratio of the elderly population can only be tentative as they need to be verified by future causal studies. Following our results, policies should promote more equal health outcomes for women and men so that elderly sex ratios equalize more effectively in the long term, have a strong focus on greater equality of distributive outcomes such as incomes, working time and pensions to reduce the financial dependencies of the female part of the population and promote the importance of the societal role of elderly care, especially regarding needs of elderly female parts of the population. However, in many OECD countries, demographic shifts, such as the progressive aging of the population and a declining generational exchange in the job market between younger and older workers, pose significant challenges to the sustainability of welfare systems. In the future, there may be fewer workers to support a growing number of elderly individuals reliant on pensions. This imbalance could undermine the viability of redistributive pension schemes designed to promote more equitable pension outcomes. Additionally, another critical factor affecting pensions, particularly for women, is the *gender pay gap*. Women consistently earn less than men, which directly translates into lower pension benefits in later life.

The dimension of gender ratios and their role in regional well-being is overlooked in most well-being indicators and constitutes one of the novelties in our findings. Our study also implies the usefulness of exploratory data-driven approaches in the empirical research of multidimensional, complex socio-economic outcomes to generate novel hypotheses and contribute to a step-wise theory-building process (Shrestha et al., 2021). In the spirit of open science practices, the workflow and corresponding code established in this paper are readily available for adaptation to studying other socio-economic outcomes. With an ever-increasing coupling of social and economic processes in an ever-increasingly connected world that generates lots of data, social scientists should embrace the ability of data-driven approaches to contribute to novel hypotheses in the scientific inquiry of the outcomes of this complex system.

Appendix

Correlations of our Final set of Predictors

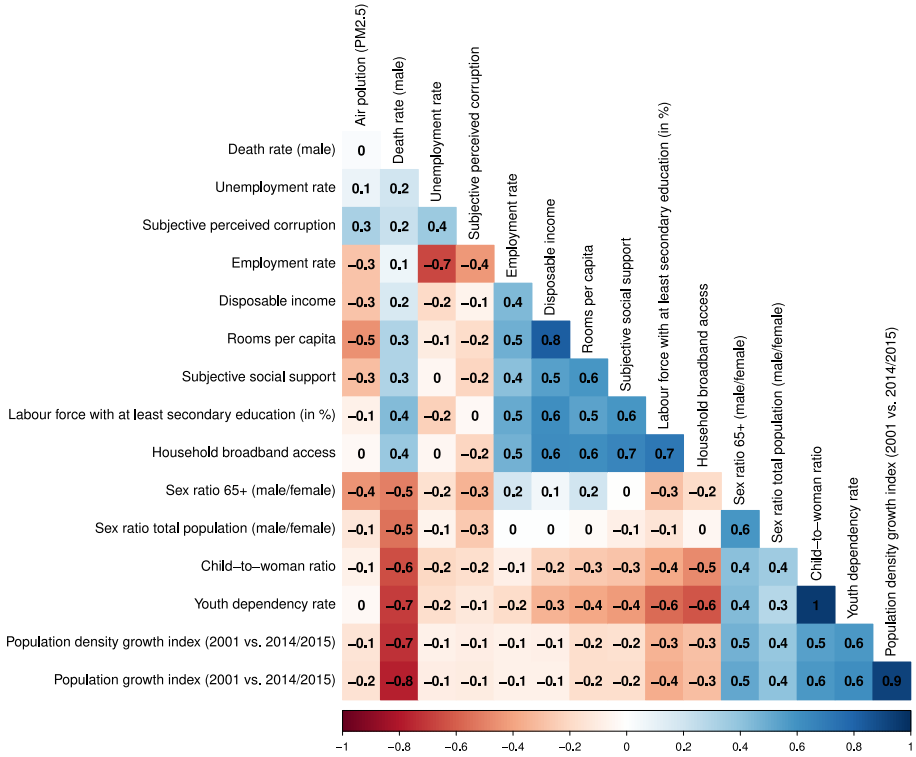


Fig. 5 Correlation matrix visualized of the predictors used in our final RF model. The plot shows pairwise Pearson correlation coefficients (colored cells) for each predictor in the dataset. Predictors are ordered based a hierarchical clustering. Coefficients are displayed in black text with one decimal point. The color scale reflects the strength of the correlations

Robustness Check - Random Forest - XGBoost

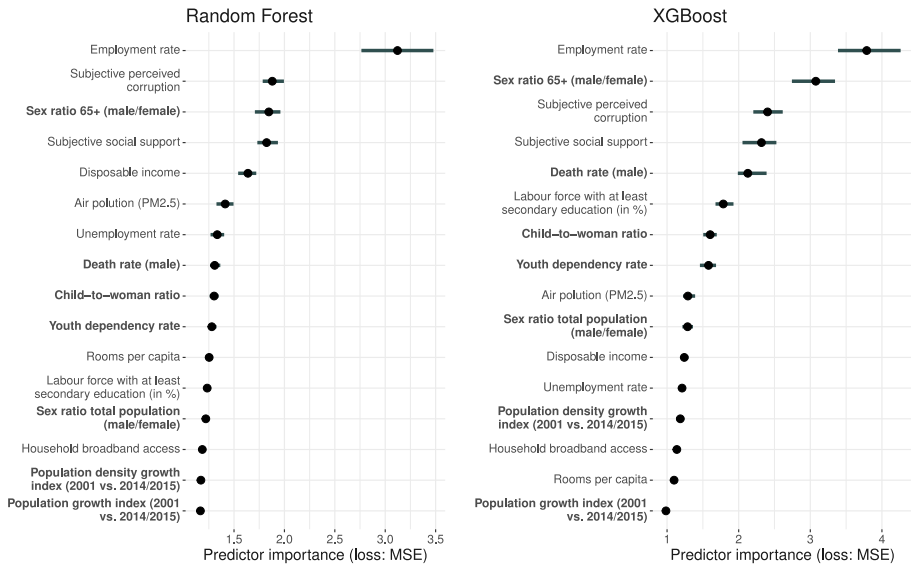


Fig. 6 Importance comparison between Random Forest and XGBoost model. Along the lines of the importance plot in the main manuscript, this plot shows the importance scores for the predictors for the Random Forest and the XGBoost model

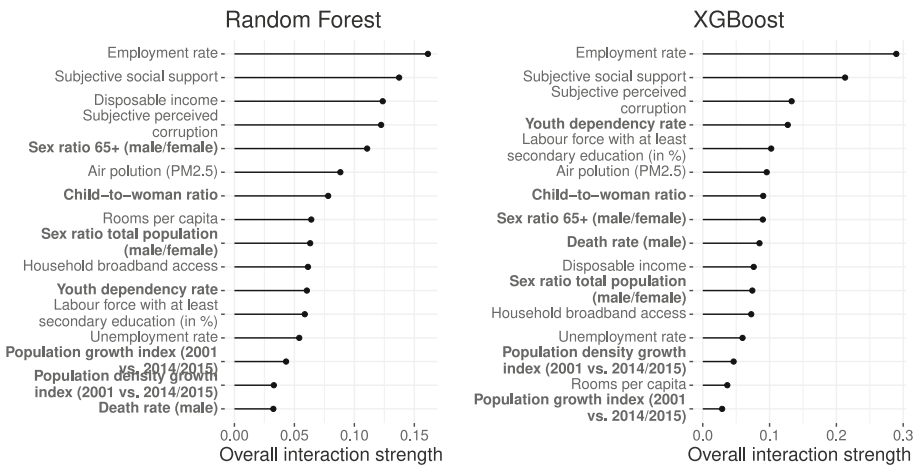


Fig. 7 Interaction comparison between Random Forest and XGBoost model. Along the lines of the interaction plot in the main manuscript, this plot shows the interaction scores of the predictors for the Random Forest and the XGBoost model

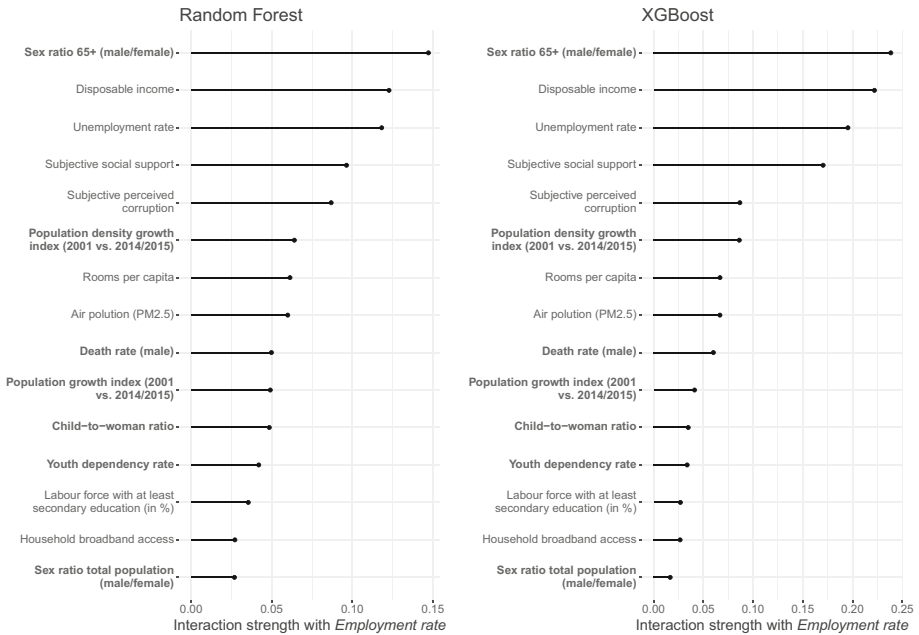


Fig. 8 Interaction with employment rate comparison between Random Forest and XGBoost model. Along the lines of the interaction plot in the main manuscript, this plot shows the interaction scores of the predictors with the Employment Rate for the Random Forest and the XGBoost model

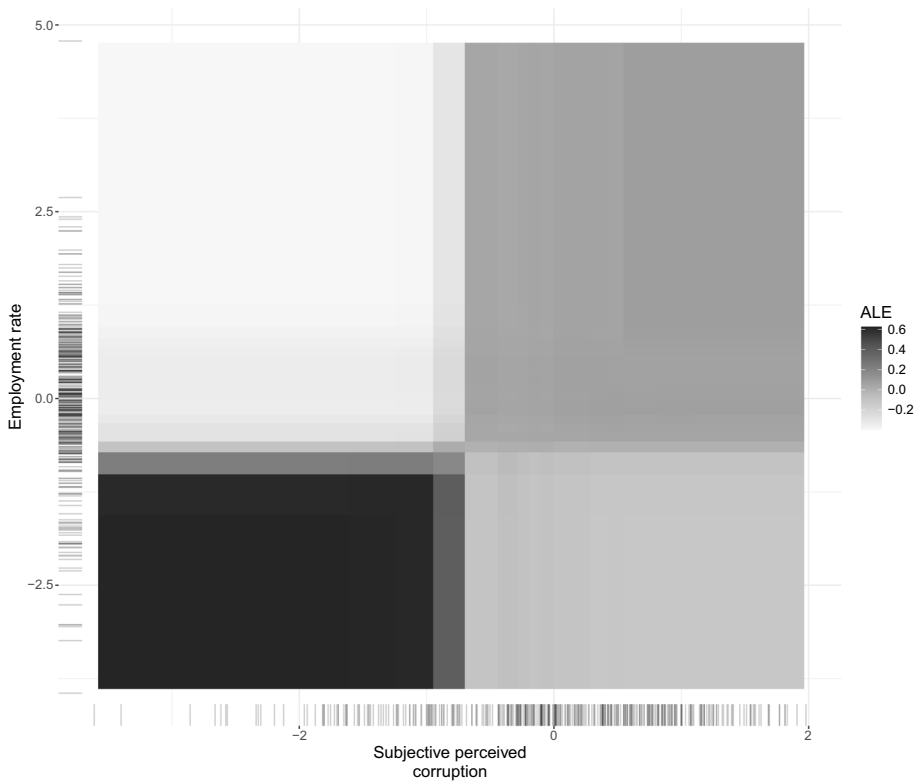


Fig. 9 ALE plot for the interaction of the employment rate with the elderly sex ratio of the XGBoost model. Along the lines of the ALE plot in the main manuscript, this plot shows the ALE plot and thus the functional form of the interactions between the employment rate with the elderly sex ratio scores for the XGBoost model

Acknowledgements Lasare Samartzidis would like to thank Artem Korzhenevych, Gerd Lintz and Mathias Jehling for valuable discussions, inputs and comments on previous versions of the manuscript. Further valuable comments were received from participants of the 10th International Degrowth Conference/15th Conference of the European Society for Ecological Economics, Pontevedra (ES), 2024. A previous version of this work has been awarded the [Young Environmental Economist Award](#) at the 11th Annual Conference of the Italian Association of Environmental and Resource Economists - IAERE 23-24 February 2023, Naples (Italy). Marco Quatrosi acknowledges the support of the Next Generation EU project GRINS-Growing Resilient, Inclusive Sustainable PE00000018 CUP B73C22001260006.

Author Contributions Conceptualization: LS, MQ, AvD, methodology: LS, MQ, literature review: MQ LS, formal analysis, data compilation, software, visualization: LS, open science implementation: LS, resources and writing - draft preparation: LS, MQ, validation, and revisions: LS, MQ, review, and editing: LS, MQ.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The datasets analysed during the current study and code to reproduce all results, figures, and tables are available in a Zenodo repository from <https://doi.org/10.5281/zenodo.15535818>.

Declarations

Ethical Approval and Informed Consent Not applicable to this study as only data from secondary sources were analyzed.

Competing interests The authors declare no competing financial interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ali, S., Murshed, S. M., & Papyrakis, E. (2020). Happiness and the Resource Curse. *Journal of Happiness Studies*, 21(2), 437–464. <https://doi.org/10.1007/s10902-019-00080-3> <http://link.springer.com/10.1007/s10902-019-00080-3>
- Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4), 1059–1086. <https://doi.org/10.1111/rssb.12377> <https://onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12377>
- Atkinson, G., & Mourato, S. (2015). *Cost-Benefit Analysis and the Environment*. <https://www.oecd-ilibrary.org/content/paper/5jrp6w76tstg-en>. Paris:OECD
- Balestra, C., Boarini, R., & Tosetto, E. (2018). *What matters most to people? Evidence from the OECD better life index users' responses*. *Social Indicators Research*, 136, 907–930. Publisher: Springer.
- Ballas, D., & Thanis, I. (2022). *Exploring the geography of subjective happiness in europe during the years of the economic crisis: A multilevel modelling approach*. *Social Indicators Research*, 1–33. Publisher: Springer.
- Barrington-Leigh, C., & Escande, A. (2018). Measuring progress and well-being: A comparative review of indicators. *Social Indicators Research*, 135(3), 893–925. <https://doi.org/10.1007/s11205-016-1505-0> <http://link.springer.com/10.1007/s11205-016-1505-0>
- Bartram, D. (2022). The 'Gender Life-Satisfaction/Depression Paradox' is an artefact of inappropriate control variables. *Social Indicators Research*, 164(3), 1061–1072. <https://doi.org/10.1007/s11205-022-02986-7> <https://link.springer.com/10.1007/s11205-022-02986-7>
- Bartram, D. (2024). To evaluate the age-happiness relationship, look beyond statistical significance. *Journal of Happiness Studies*, 25(1–2), 22. <https://doi.org/10.1007/s10902-024-00728-9> <https://link.springer.com/10.1007/s10902-024-00728-9>
- Bartram, D., Alaimo, L. S., Avery, E., Bardo, A., Bella, E. D., Binder, M., & Tani, M. (2024). Towards the next fifty years of social indicators research: Some guidance for authors. *Social Indicators Research*, 174(1), 1–17. <https://doi.org/10.1007/s11205-024-03401-z> <https://link.springer.com/10.1007/s11205-024-03401-z>
- Bleys, B. (2012). Beyond GDP: Classifying alternative measures for progress. *Social indicators research*, 109, 355–376.
- Boarini, R., Johansson, A., & d'Ercole, M. M. (2006). *Alternative measures of well-being*. Publisher: Oecd.
- Brand-Correa, L., Brook, A., Büchs, M., Meier, P., Naik, Y., & O'Neill, D. W. (2022). Economics for people and planet-moving beyond the neoclassical paradigm. *The Lancet Planetary Health*, 6(4), e371–e379. [https://doi.org/10.1016/S2542-5196\(22\)00063-8](https://doi.org/10.1016/S2542-5196(22)00063-8) <https://linkinghub.elsevier.com/retrieve/pii/S2542519622000638>
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199–231.
- Cavalletti, B., & Corsi, M. (2018). "Beyond GDP" effects on national subjective well-being of OECD countries. *Social Indicators Research*, 136(3), 931–966. <https://doi.org/10.1007/s11205-016-1477-0>
- Chaaban, J., Irani, A., & Khoury, A. (2016). The composite global well-being index (CGWBI): A new multi-dimensional measure of human development. *Social Indicators Research*, 129, 465–487.

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: Association for Computing Machinery. <https://dl.acm.org/doi/10.1145/2939672.2939785>
- Ciommi, M., Gigliarano, C., Emili, A., Taralli, S., & Chelli, F. M. (2017). A new class of composite indicators for measuring well-being at the local level: An application to the Equitable and Sustainable Well-being (BES) of the Italian Provinces. *Ecological Indicators*, 76, 281–296. <https://doi.org/10.1016/j.ecolind.2016.12.050> <https://www.sciencedirect.com/science/article/pii/S1470160X16307555>
- Cook, D., & Davíðsdóttir, B. (2021). An appraisal of interlinkages between macro-economic indicators of economic well-being and the sustainable development goals. *Ecological Economics*, 184, 106996. <https://doi.org/10.1016/j.ecolecon.2021.106996> <https://www.sciencedirect.com/science/article/pii/S0921800921000549>
- Costanza, R., Hart, M., Posner, S., & Talberth, J. (2009). *Beyond GDP: The Need for New Measures of Progress*. (MAG ID: 2159593953)
- Costanza, R., Kubiszewski, I., Giovannini, E., Lovins, H., McGlade, J., Pickett, K. E., & Wilkinson, R. G. (2014). Development: Time to leave GDP behind. *Nature*, 505(7483), 283–285. <https://doi.org/10.1038/505283a> (MAG ID: 2051147555)
- Deaton, A. (2012). The financial crisis and the well-being of Americans: 2011 OEP Hicks Lecture *. *Oxford Economic Papers*, 64(1), 1–26. <https://doi.org/10.1093/oenp/gpr051>
- De Langhe, B., Puntoni, S., & Larrick, R. (2017). Linear thinking in a nonlinear world. *HARVARD BUSINESS REVIEW*, 95(3), 130–139, (tex.orcid-numbers: Larrick, Richard/0000-0002-0491-426X De Langhe, Bart/0000-0002-7913-5279 Puntoni, Stefano/0000-0002-3259-2325 tex.researcherid-numbers: De Langhe, Bart/ISB-0198-2023 Larrick, Richard/B-8944-2011 tex.unique-id: WOS:000399968200029)
- De Neve, J. E., & Sachs, J. D. (2020). The SDGs and human well-being: A global analysis of synergies, trade-offs, and regional differences. *Scientific Reports*, 10(1), 15113. <https://doi.org/10.1038/s41598-020-71916-9> <https://www.nature.com/articles/s41598-020-71916-9>
- Diener, E., & Suh, E. (1997). Measuring quality of life: Economic, social, and subjective indicators. *Social indicators research*, 189–216. Publisher: JSTOR.
- Di Paolo, A., & Ferrer-i-Carbonell, A. (2022). Regional borders, local unemployment, and life satisfaction. *Journal of Regional Science*, 62(2), 412–442. <https://doi.org/10.1111/jors.12573> <https://onlinelibrary.wiley.com/doi/10.1111/jors.12573>
- Dolan, P., Peasgood, T., & White, M. (2008). Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. *Journal of Economic Psychology*, 29(1), 94–122.
- Durand, M. (2015). The OECD better life initiative: How's life? And the measurement of well-being. *Review of Income and Wealth*, 61(1), 4–17.
- Ebbinghaus, B. (2021). Inequalities and poverty risks in old age across Europe: The double-edged income effect of pension systems. *Social Policy & Administration*, 55(3), 440–455. <https://doi.org/10.1111/spol.12683> <https://onlinelibrary.wiley.com/doi/10.1111/spol.12683>
- European Commission: Directorate-General for Employment, S.A.a.I. (2021). *The European pillar of social rights action plan*. Publications Office.
- Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3), 916–954. <https://doi.org/10.1214/07-AOAS148> <https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-3/Predictive-learning-via-rule-ensembles/10.1214/07-AOAS148.full> (Publisher: Institute of Mathematical Statistics)
- Glatz, C., & Eder, A. (2020). Patterns of Trust and Subjective Well-Being Across Europe: New Insights from Repeated Cross-Sectional Analyses Based on the European Social Survey 2002–2016. *Social Indicators Research*, 148(2), 417–439. <https://doi.org/10.1007/s11205-019-02212-x> <https://doi.org/10.1007/s11205-019-02212-x>
- Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, 27(3), 659–678. <https://doi.org/10.1007/s11222-016-9646-1>
- Hayden, A., & Wilson, J. (2018). Taking the First Steps beyond GDP: Maryland's Experience in Measuring "Genuine Progress." *Sustainability*, 10(2), 462. <https://doi.org/10.3390/su10020462> <http://www.mdpi.com/2071-1050/10/2/462>
- Hoekstra, R. (2019). *Replacing GDP by 2030: Towards a common language for the well-being and sustainability community (First Edition ed.)*. New York: Cambridge University Press.
- Hoogerbrugge, M. M., Burger, M. J., & Van Oort, F. G. (2022). Spatial structure and subjective well-being in North-West Europe. *Regional Studies*, 56(1), 75–86. <https://doi.org/10.1080/00343404.2021.1873261> <https://www.tandfonline.com/doi/full/10.1080/00343404.2021.1873261>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

- Kahneman, D., & Krueger, A. B. (2006). Developments in the Measurement of Subjective Well-Being. *Journal of Economic Perspectives*, 20(1), 3–24. <https://doi.org/10.1257/089533006776526030> <https://www.aeaweb.org/articles?id=10.1257>
- Kenny, D. C., Costanza, R., Dowsley, T., Jackson, N., Josol, J., Kubiszewski, I., & Thompson, J. (2019). Australia's Genuine Progress Indicator Revisited (1962–2013). *Ecological Economics*, 158, 1–10. <https://doi.org/10.1016/j.ecolecon.2018.11.025> <https://linkinghub.elsevier.com/retrieve/pii/S0921800917317445>
- Luhmann, M. (2017). Using Big Data to study subjective well-being. *Current Opinion in Behavioral Sciences*, 18, 28–33. <https://doi.org/10.1016/j.cobeha.2017.07.006> <https://www.sciencedirect.com/science/article/pii/S2352154617300360>
- MacKerron, G., & Mourato, S. (2013). Happiness is greater in natural environments. *Global Environmental Change*, 23(5), 992–1000. <https://doi.org/10.1016/j.gloenvcha.2013.03.010> <https://www.sciencedirect.com/science/article/pii/S0959378013000575>
- Mahoney, J. (2023). Subjective well-being measurement (17). <https://doi.org/10.1787/4e180f51-en> <https://www.oecd-ilibrary.org/content/paper/4e180f51-en>
- Methorst, J., Rehdanz, K., Mueller, T., Hansjürgens, B., Bonn, A., & Böhning-Gaese, K. (2021). The importance of species diversity for human well-being in Europe. *Ecological Economics*, 181, 106917. <https://doi.org/10.1016/j.ecolecon.2020.106917> <https://www.sciencedirect.com/science/article/pii/S0921800920322084>
- Michalos, A., Smale, B., Labonté, R., Muhtarjane, N., Scott, K., Moore, K., et al. (2011). *The Canadian index of well-being*. Waterloo: Canadian Index of Wellbeing.
- Mínguez, A. M., & Vírveda, J. A. V. (2019). Well-being and living arrangement of elderly people from European comparative perspective. *The Social Science Journal*, 56(2), 228–242.
- Molnar, C. (2022). *Interpretable machine learning. A Guide for Making Black Box Models Explainable* (2nd ed.). <https://christophm.github.io/interpretable-ml-book>
- Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), 87–106. <https://doi.org/10.1257/jep.31.2.87> <https://www.aeaweb.org/articles?id=10.1257>
- OECD. (2016). *OECD regions at a glance 2016*. Paris: OECD Publishing. https://doi.org/10.1787/reg_glance-2016-en
- OECD. (2017). *How's life? 2017: Measuring well-being*. Paris: OECD Publishing. https://doi.org/10.1787/how_life-2017-en
- OECD. (2019). *OECD regional statistics*.
- O'Neill, D. W., Fanning, A. L., Lamb, W. F., & Steinberger, J. K. (2018). A good life for all within planetary boundaries. *Nature Sustainability*, 1(2), 88–95. <https://doi.org/10.1038/s41893-018-0021-4> <https://www.nature.com/articles/s41893-018-0021-4>
- Otero, I., Farrell, K.N., Pueyo, S., Kallis, G., Kehoe, L., Haberl, H., & Pe'er, G. (2020). Biodiversity policy beyond economic growth. *Conservation Letters*, 13(4), <https://doi.org/10.1111/conl.12713> <https://onlinelibrary.wiley.com/doi/10.1111/conl.12713>
- Pacione, M. (1982). The use of objective and subjective measures of life quality in human geography. *Progress in Human Geography*, 6(4), 495–514.
- Pacione, M. (2003). Urban environmental quality and human wellbeing—a social geographical perspective. *Landscape and Urban Planning*, 65(1–2), 19–30.
- Peiró-Palomino, J. (2019). Regional well-being in the OECD. *The Journal of Economic Inequality*, 17(2), 195–218. <https://doi.org/10.1007/s10888-018-9398-6>
- Puntscher, S., Hauser, C., Walde, J., & Tappeiner, G. (2015). The impact of social capital on subjective well-being: A regional perspective. *Journal of Happiness Studies*, 16(5), 1231–1246. <https://doi.org/10.1007/s10902-014-9555-y>
- Schmidt, E., Pacek, A.C., & Radcliff, B. (2023). The welfare state and human well-being around the world: A cross-national analysis. *Applied Research in Quality of Life*. <https://doi.org/10.1007/s11482-023-10247-z> <https://link.springer.com/10.1007/s11482-023-10247-z>
- Schneider, S. M. (2016). Income inequality and subjective wellbeing: Trends, challenges, and research directions. *Journal of Happiness Studies*, 17(4), 1719–1739. <https://doi.org/10.1007/s10902-015-9655-3>
- Shrestha, Y. R., He, V. F., Puranam, P., & von Krogh, G. (2021). Algorithm supported induction for building theory: How can we use prediction models to theorize? *Organization Science*, 32(3), 856–880. <https://doi.org/10.1287/orsc.2020.1382> <https://pubsonline.informs.org/doi/full/10.1287/orsc.2020.1382>. (Publisher: INFORMS)
- Steiner, L., Frey, B., & Hotz, S. (2015). European Capitals of Culture and life satisfaction. *Urban Studies*, 52(2), 374–394. <https://doi.org/10.1177/0042098014524609> <http://journals.sagepub.com/doi/10.1177/0042098014524609>

- Stiglitz, J.E., Sen, A., & Fitoussi, J. P. (2017). *Report by the commission on the measurement of economic performance and social progress*.
- Talberth, J., & Weisdorf, M. (2017). Genuine Progress Indicator 2.0: Pilot Accounts for the US, Maryland, and City of Baltimore 2012–2014. *Ecological Economics*, 142, 1–11. <https://doi.org/10.1016/j.ecolecon.2017.06.012> <https://www.sciencedirect.com/science/article/pii/S0921800917304184>
- Tan, J., Liu, H., Li, M., & Wang, J. (2018). A prediction scheme of tropical cyclone frequency based on lasso and random forest. *Theoretical and Applied Climatology*, 133(3), 973–983. <https://doi.org/10.1007/s00704-017-2233-3>
- Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Van Lissa, C. J. (2023). Developmental data science: How machine learning can advance theory formation in Developmental Psychology. *Infant and Child Development*, 32(6), e2370. <https://doi.org/10.1002/icd.2370> <https://onlinelibrary.wiley.com/doi/abs/10.1002/icd.2370>
- Zhang, S., Zhu, D., Shi, Q., & Cheng, M. (2018). Which countries are more ecologically efficient in improving human well-being? An application of the Index of Ecological Well-being Performance. *Resources, Conservation and Recycling*, 129, 112–119. <https://doi.org/10.1016/j.resconrec.2017.10.015> <https://linkinghub.elsevier.com/retrieve/pii/S0921344917303348>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.