NOWCASTING SUBJECTIVE WELL-BEING WITH GOOGLE TRENDS: A META-LEARNING APPROACH

OECD PAPERS ON WELL-BEING AND INEQUALITIES
https://doi.org/10.1787/4ca48f7c-en

WORKING PAPER No.27

Fabrice Murtin
Max Salomon-Ermel

Organisation for Economic Co-operation and Development

WISE(2024)6

English text only

WELL-BEING, INCLUSION, SUSTAINABILITY AND EQUAL OPPORTUNITY CENTRE

Nowcasting subjective well-being with Google Trends

A meta-learning approach

JEL Classification: C1, C45, C53, D60, I31
Keywords: well-being, poverty, spatial inequality

All OECD Working Papers on Well-being and Inequalities are available at:
https://www.oecd.org/wise/papersandbriefs.

Fabrice Murtin: fabrice.murtin@oecd.org

OECD Papers on Well-being and Inequalities

www.oecd.org/wise/papersandbriefs

The OECD Papers on Well-being and Inequalities Series – managed by the OECD Centre on Well-being, Inclusion, Sustainability and Equal Opportunity (WISE) – features working papers on the measurement agenda for well-being, inclusion, sustainability and equal opportunity as well as papers seeking to deepen the understanding of the drivers of these issues, the ways in which they interact and how they evolve. These papers are prepared by OECD staff, external experts or by outside consultants working on OECD projects.

OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed are those of the author(s).

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcomed, and may be sent to wellbeing@oecd.org.

This series is designed to make available to a wider readership selected labour market, social policy and migration studies prepared for use within the OECD. Authorship is usually collective, but principal writers are named. The papers are generally available only in their original language – English or French – with a summary in the other.

The release of this working paper has been authorised by Romina Boarini, Director of the OECD Centre on Well-being, Inclusion, Sustainability and Equal Opportunity (WISE).

Abstract

This paper applies Machine learning techniques to Google Trends data to provide real-time estimates of national average subjective well-being among 38 OECD countries since 2010. We make extensive usage of large custom micro databases to enhance the training of models on carefully pre-processed Google Trends data. We find that the best one-year-ahead prediction is obtained from a meta-learner that combines the predictions drawn from an Elastic Net with and without interactions, from a Gradient-Boosted Tree and from a Multi-layer Perceptron. As a result, across 38 countries over the 2010-2020 period, the out-of-sample prediction of average subjective well-being reaches an R^2 of 0.830.

Résumé

Ce document applique des techniques d'apprentissage automatique aux données Google Trends afin de fournir des estimations en temps réel du bien-être subjectif moyen national dans 38 pays de l'OCDE depuis 2010. Nous faisons un usage intensif de grandes bases de données micro pour améliorer l'apprentissage des modèles sur des données Google Trends soigneusement prétraitées. Nous constatons que la meilleure prédiction à un an est obtenue à partir d'un méta-apprentissage qui combine les prédictions tirées d'un réseau élastique avec et sans interactions, d'un arbre boosté par gradient et d'un perceptron multicouche. En conséquence, pour 38 pays sur la période 2010-2020, la prédiction hors échantillon du bien-être subjectif moyen atteint un R2 de 0,830.

Table of contents

Tables

Figures

1 Introduction

1. Easily surveyable measures like subjective well-being (SWB) can help governments better understand people's life experiences and anticipate their behaviour. Moreover, this type of measures also allows policy-makers to understand reactions to policy, shocks and crises (Kaiser, Oswald and Easterlin, 2022[1]). The past two French presidential elections are good examples as results suggested that voting for far-left and far-right candidates was negatively correlated with SWB (Algan et al., 2018[2]). More generally speaking SWB measures are highly relevant from a policy perspective (OECD, 2020[3]) and they gained influence to monitor population's well-being during the COVID-19 (OECD, 2021[4]).

2. Despite the several applications and increasing demand of SWB measures, there are many cases where potentially useful data on SWB is hardly available in a satisfactory manner: first, the frequency of large national or cross-national surveys can be too low to allow us to identify shifts precisely and in a timely fashion. Second, many countries are only included episodically in large cross-national surveys, creating issue for long-term tracking of evolutions.

3. In an attempt to solve these two challenges, this paper uses Machine Learning techniques applied to Google Trends data to build nowcasting models of SWB among OECD countries. In particular, this paper provides an extensive description of the complex data engineering required to deliver proper predictions at a two-year horizon. All steps in the proposed methodology are tested and supported by intermediary experiments, which highlight the high degree of sophistication that is needed to achieve good forecast performance and is intended to guide practitioners in their methodological choices in a similar context. In particular, the paper sheds light on important intermediary steps of Machine Learning estimation, such as data treatment, dimension reduction, optimization of hyper-parameters, criterion selection, validation method and models combination through an ensemble method.

4. The scope of this paper is placed on life satisfaction (also called life evaluation), which is a specific dimension of SWB (Diener et al., 1999[5]), and is often used as a proxy in the literature. Evaluations can be done with regards to the overall lives of the respondents, or focused on certain domains like wealth, health, or relationships (Cheung and Lucas, 2014[6]). The latter, which are more resource-intensive and with a lower response-rate than the general question, are unfortunately rarer in surveys (Kahneman and Krueger, 2006[7]). The study of subjective evaluations has become increasingly relevant over time for social scientists and policy-makers alike. As argued by (Diener and Seligman, 2009[8]), for developed countries, SWB may be a more important metric to maximise than growth in the long-run, given that life satisfaction has stagnated over the last decades in some countries despite rapid increases in GDP, although this view has been contested ((Stevenson and Wolfers, 2008[9]); (Böckerman, Laamanen and Palosaari, 2016[10])). In any case, the SWB of individuals varies with changes of circumstances and environments, such as widowhood, long-term disabilities, but also with exogenous shocks like natural disasters or economic crises ((De and Thamarapani, 2022[11]); (Hariri, Bjørnskov and Justesen, 2016[12]); (Lucas, Dyrenforth and Diener, 2008[13])). Therefore, tracking changes in SWB within populations or sub-populations makes sense from a policy-making perspective as it allows governments to notice the impacts of policies and external shocks, and potentially take action to remedy the situation.

5. Yet, despite the potential applications of SWB tracking, countries are unequal in their access to timely information, which can be crucial especially at the onset of a crisis. The past episode of the

coronavirus pandemic illustrates the importance of having timely SWB data to assess the need and scope for policy interventions. As an example, (Foa, Gilbert and Fabian, 2020[14]), (Foa, Fabian and Gilbert, 2022[15]) and (Brodeur et al., 2021[16]) have developed trackers of SWB during the COVID-19 pandemic to disentangle the impacts of the lockdowns and those of the pandemic, while using various timely data sources.

6. In the same vein, we advocate in this paper for the use of Google Trends time series as a way to bypass issues related to the timeliness of the reporting on SWB, whether in national or cross-national contexts. This methodology is not novel and Google Trends data have already been used successfully in the past for many applications: (Algan et al., 2019[17]) track variations in the search frequency of multiple queries to predict changes in the average SWB of the USA, both nationally and at the state level. Similarly, (Woloszko, 2020[18]) uses similar data to nowcast throughout OECD countries the changes in GDP at a high level of temporal granularity. The richness of the literature on using Google Trends for social sciences is not surprising. Indeed, Google Trends data enable researchers to track changes in the search frequency of keywords, themes, and concepts on Google, over time and across countries. Moreover, this data is open-access and timely. Furthermore, given that Google Search users are usually not aware that their queries are collected and aggregated, Google Trends data can be compared to naturalistic observations of people’s interest.

7. Nevertheless, the novel use of this tool also implies new methodological challenges. Specifically, researchers on Google Trends can obtain time series of the Search Volume Index (SVI) for specific expressions, topics, and categories of queries, for specific locations. The SVI can be summarized as the relevant share of total searches, in a given area, in a given population, normalized between 0 and 100, for a certain expression or topic. Given the nature of the SVI, its custom scaling and privacy requirements, there is a certain degree of opacity as to how the SVI is calculated and what it represents. Moreover, as described by (Bokelmann and Lessmann, 2019[19]), many Google Trends time series, even when not thematically related, can behave similarly and show spurious associations. These challenges require researchers to be prudent in their interpretation of the relationships they analyse, but also in the way they pre-process their data.

8. In this paper, we attempt to overcome these limitations and to show the potential of using Google Trends to nowcast SWB within OECD countries. Two-years ahead prediction is achieved thanks to the help of a sophisticated Meta-Learning methodology that includes: i) an extensive use of large custom micro databases for enhanced training; ii) detailed investigation of pre-processing methods for Google Trends data; ii) Bayesian optimization of hyper-parameters and usage of a country-weighted MSE criterion for out-of-sample predictions; iii) use of walk-forward validation instead of k-fold cross-validation; iv) choice of different base-learners, in practice an Elastic Net with and without interactions, a Gradient-Boosted Tree and a Multi-layer Perceptron; v) estimation of a meta-learner with country fixed-effects interactions with the base learners predictions. As a result, across 38 countries over the 2010-2020 period, the out-of-sample prediction of average SWB reaches a R2 of 0.830, with 95% of true country-month averages being within 0.664 of our final out-of-sample predictions.

9. The paper is structured as follows. Section 2 describes the data and methods, Section 3 presents the main results and Section 4 concludes.

2 Data

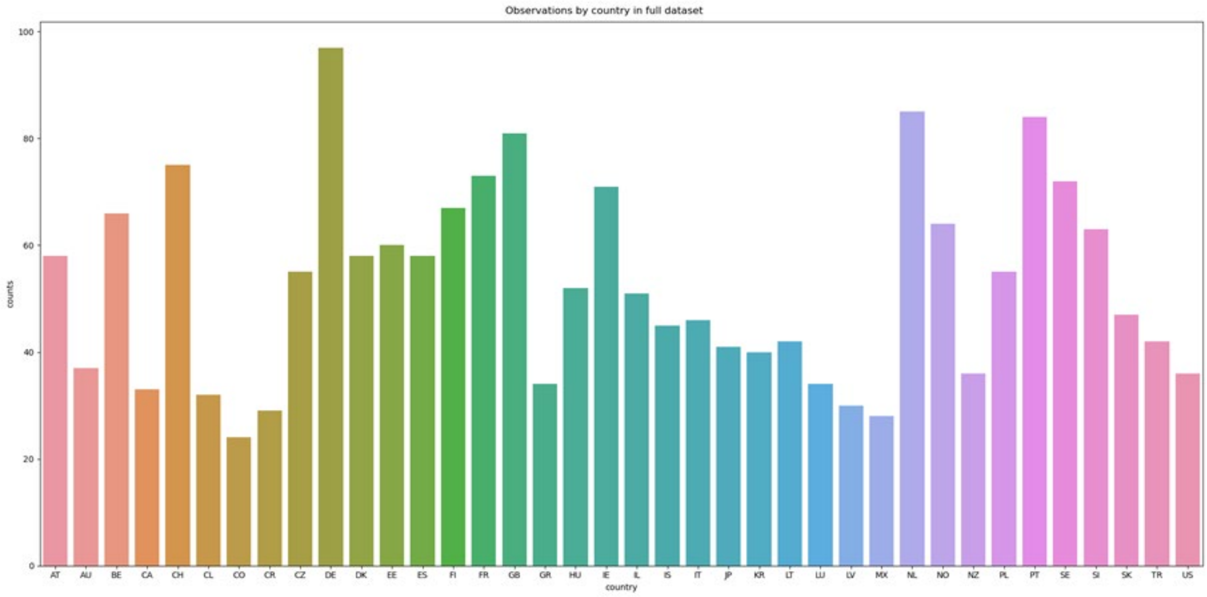
2.1. National average SWB

10. One of the innovative aspects of this paper is the creation of a large custom dataset of individual measures of general SWB. Specifically, we harmonize and combine multiple panel surveys, which include the European Social Survey (ESS), the Gallup World Poll (GWP), the World Value Survey (WVS), and the European Value Survey (EVS).

11. All of these surveys feature in the multiple waves and countries the standard single-item measure of SWB, rated with comparable question wording and Likert scales. Moreover, the surveys also share comparable survey-weighting schemes, with sample weights showing similar distributions (Figure A.1 in Supplementary Information, henceforth SI, available from the authors upon request). Individual observations are aggregated from these four surveys, keeping information on countries, year and month of response, self-reported SWB, and individual sample weight, to finally obtain a large sample of 966,577 observations covering the 38 OECD member countries. These observations are used to compute time series of weighted monthly average SWB for each country, resulting in 2,212 country-month observations (see Table A.1 in SI). On a second step, the database is filtered based on different minimum number of survey respondents per country-month. After visual inspection, we settle on using at least 25 respondents for each country-month observation, in order to reduce noise and non-representativeness of monthly estimates, while at the same time retaining as many samples as possible. As an example, Figure A.2. in SI shows with the French time series that a minimum of 25 samples per month allows to align with the yearly trends, to minimize noisy outliers, while maximizing sample size. Overall, this exclusion corresponds to removing about 10% of the observations from the data set as can be seen on the plot of the distribution of monthly sample sizes (Figure A.3 in SI). Using this threshold, time series retain the general yearly trends but also provide additional information despite the apparent noisiness of the data. First, they include information about seasonality that we are interested in modelling. Second, they allow us to consider the impact of punctual extreme events as can be seen in the UK time series (Figure A.4. in SI). While the time series we use is not as extreme as the unfiltered one, it still does indicate a sudden drop in 2005 that is probably related to the 2005 London Bombings and which impact of the population’s SWB has been well documented in the literature (Dustmann and Fasani, 2014[20]).

12. The final data set comprises 2001 country-month observations. It exhibits significant imbalance between countries in terms of monthly measures for each as can be seen on Figure 1. There are 52.7 observations per country on average, with a minimum of 24 for Colombia and a maximum of 97 for Germany. This imbalance has consequences for the metrics used when fitting and tuning models (like mean squared error), and thus it will be properly addressed methodologically in this paper.

Figure 1. Number of monthly observations per country



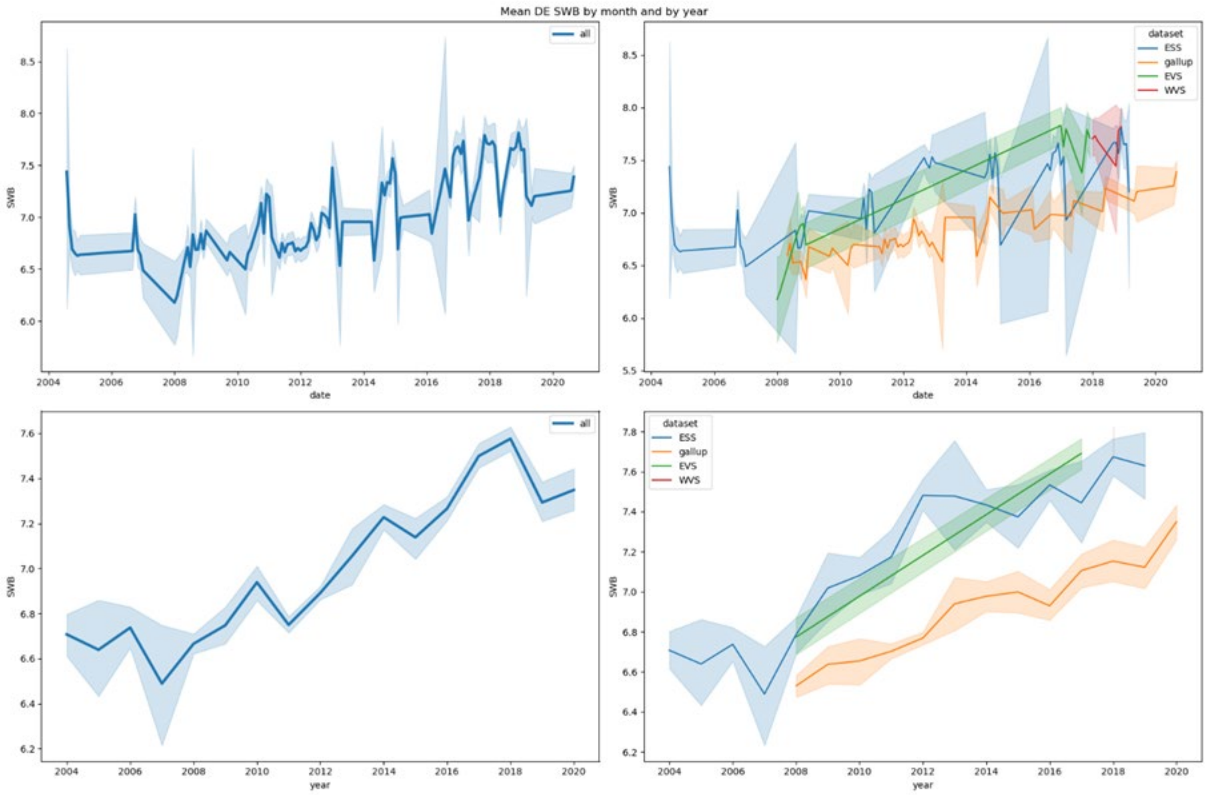
Source: OECD calculations based on the European Social Survey (ESS), the Gallup World Poll (GWP), the World Value Survey (WVS), and the European Value Survey (EVS).

13. In terms of temporal continuity, the data set contains a significant number of gaps between country-specific observations at a monthly frequency (Figure A.5. in SI). Moreover, these gaps are random rather than periodic. This has consequences for the modelling approaches that can be used. In particular, this prevents us from using methods based on analysis of month-to-month growth, or year-to-year monthly changes like (Woloszko, 2020[18]) did to account for seasonality.

14. The choice of combining multiple surveys first allows to create a much richer training dataset to train the nowcasting model. Indeed, had we decided to create our country-month SWB averages using only the Gallup World Poll, we would have ended up with about half the final number of observations, specifically 1105 country-month observations with some countries absent or under-represented. In turn, a larger training data set allows to estimate more complex models which ultimately perform better. A second motivation for aggregating these data sets is to reduce survey-specific noise: (Deaton and Stone, 2016[21]) explain that contextual effects in Gallup surveys in the US can have large consequences on the reported SWB of respondents. Specifically, the authors warn researchers of the impact the ordering of questions in surveys can have, reporting that when Gallup asked political questions before the SWB one, reported SWB was significantly lower. We also observe similar phenomena when comparing the yearly estimates of average SWB in the US, computed using either the Gallup US Dailies or the Gallup World Poll. The same observation can be made when comparing averages computed on the different surveys for the entire OECD time series or for specific countries like Germany (Figure 2). In particular, there are differences in general linear trends as well as differences in absolute value.

Figure 2. Time series of the average SWB in Germany

By month (top) and by year (bottom), for the combined (left) and the individual (right) data sets



Source: OECD calculations based on the European Social Survey (ESS), the Gallup World Poll (GWP), the World Value Survey (WVS), and the European Value Survey (EVS).

2.2. Google Trends time series

Creation of the full and condensed data sets

15. Google Trends offers monthly time series for the SVI of specific queries, topics, and categories of searches. Specific queries are what normal users typically use on Google. They allow researchers on Google Trends to be certain of the object they are tracking over time. However, these specific queries have many disadvantages compared to topics and categories: they need to be translated in the local language of each country included. Moreover, this type of queries is case-sensitive, meaning that one would miss similar concepts worded differently. This is particularly problematic in cases where a same concept is formulated in different ways across countries or across time. In comparison, topics and categories of searches are created and aggregated by Google, in order to maintain a semantic consistency across countries and across time. Hence, given the cross-national context of this study, topics are more appropriate. While categories of searches aggregate queries that are thematically related, like Women's Health, Gifts, or Christmas, topics refer to more specific, more precise and less ambiguous concepts, like Cheating in Relationships and General Cheating.

16. For the purpose of training nowcasting models, we collect a large sample of time series from Google Trends, covering many topics and categories of searches that are potentially relevant to predict SWB. We first create a large dictionary of topics and categories of searches, that is used to generate national time series. To combine a data-intensive and a theory-driven approach, this dictionary includes

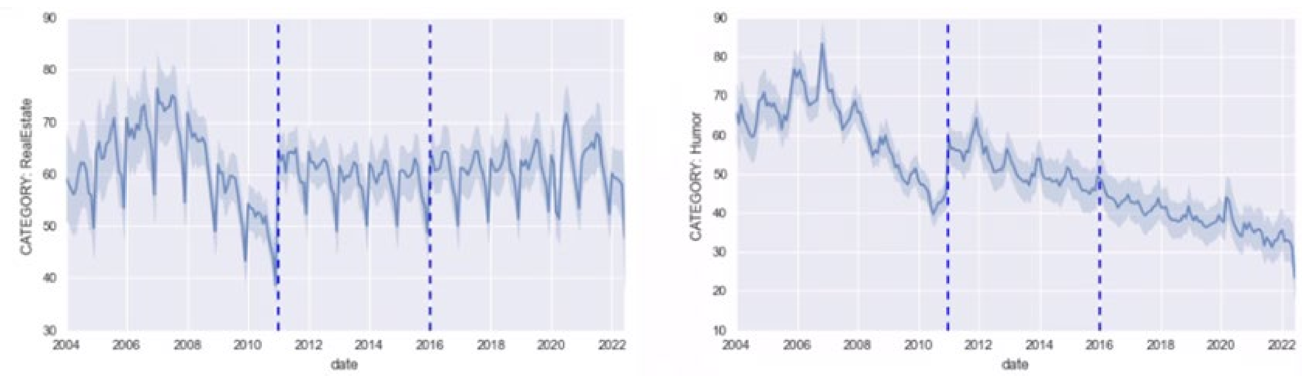
all the possible categories of searches on Google Trends (N = 1,426) and a hand-picked selection of topics (N=218) based on the existing literature on well-being. Specifically, topics are chosen based on their relevance according to the American Time Use Survey, the OECD Well-being framework (OECD, 2020[3]), and the domains of life satisfaction in happiness studies. Once the dictionary is finalized, we create a loop using the Google Trends API for Python PyTrends and iterate over the different country-keywords tuples to generate the Google Trends dataset of time series. Because of deprecation of certain references, over time or between countries, there is notable amount of loss, and the final dataset contains 1072 time series, including 914 categories and 158 topics.

17. Finally, a condensed version of the Google Trends data set has been created, where multiple time series are aggregated into subtopics, based on the facets of the OECD Better Life Index. This allows create sparser models for an exploratory econometric approach, with fewer, more interpretable, and less noisy Google Trends variables (see Figure A.7 in SI). In the end, the condensed Google Trends database features 42 composite variables.

Pre-processing the Google Trends variables

18. As mentioned in the introduction, multiple papers have used Google Trends time series as predictors in social sciences and faced similar problems during the pre-processing step given the artefacts these time series contain. In this literature, (Bokelmann and Lessmann, 2019[19]) and (Woloszko, 2020[18]) suggest some of the most thoughtful and reasonable pre-processing steps in order to better leverage Google Trends data. First, as described by both papers, the Google Trends time series tend to show similarities over time, even when they are not thematically-related, creating spurious relationships. Second, the first phenomenon is amplified across topics and categories by shared breaks, which correspond to changes in the way Google reported search queries and computed the SVI (Figure 3). As reported by Google Trends, these breaks, which occurred in 2011 and 2016, do not correspond to meaningful changes in search behaviour but instead to changes in the way the SVI was computed. Hence, these breaks also need to be accounted for during the pre-processing step. Third, many time series, like the category Science, tend to display a decay over time, with their shares of searches decreasing because of the shift over time in the uses of Google. Woloszko (2020) posits that this represents the transition of Google from a tool for academics to a tool used by the general public.

Figure 3. Time series of two unrelated categories of searches and dates of breaks in Google Trends reporting methods (dashed lines)



Source: Google Trends.

19. To address these artefacts, (Woloszko, 2020[18]) relies on a two-step approach: first, the author deals with breaks in the time series by considering that these breaks create change in the growth rates of the SVIs. Hence, he subtracts the difference between January 2011 and January 2010, and between January 2016 and January 2015, to all subsequent observations. Then, the author assumes that the remaining common component of the time series can be separated by logging the time series and subtracting from them the first component obtained through a Principal Components Analysis of the smoothed logged time series for each country. Conversely, (Bokelmann and Lessmann, 2019[19]) uses a similar but simpler approach (see Figure A.8 in SI for a comparison made on a few categories): the authors first extract for each country a common trend from the country-averaged smoothed time series, and then divide the original time series by these country-specific common trends. Part of the rationale of the authors is that this division allows them to get rid of the more nebulous part of the computation of the SVIs, namely their denominators.

20. In this paper, we take an agnostic approach as to optimal pre-processing. Indeed, given the opacity as to how exactly the SVI is computed, one considers that pre-processing is a matter that needs to be settled based on performance rather than theory. Therefore, when optimising the final nowcasting models, we let the hyper-parameter optimization procedure and the out-of-sample predictive performance determine whether to use the pre-processing steps of (Woloszko, 2020[18]), of (Bokelmann and Lessmann, 2019[19]), or none at all.

2.3. Macroeconomic control variables

21. Macro-economic control variables are included in the set of explanatory variables as they have repeatedly shown significant relationships with SWB in the literature. Namely, GDP per capita, the inflation rate, and the participation rate for individuals between 15 and 64 years old are added in the set of controls. GDP per capita, with constant prices and purchasing power parity, as well as the participation rate, are both readily available on stats.oecd.org. The inflation rate is also open-access on the website of the World Bank. For these three control variables, there are multiple missing entries for two main reasons. First, some countries do not have valid observations and measures before certain dates. Second, many indicators are measured at a quarterly frequency and not at a monthly frequency. We therefore have to impute missing values through backward and forward interpolation. This raises concerns of course as to the validity of these variables. However, given that our objective is to train nowcasting models which are accurate and rely on regularization, one prefers to include these variables in our models as they may increase prediction performance.

3 Methods

22. The quantitative analysis proceeds in 3 steps: first, we show evidence of the contribution of Google Trends data to the explanatory power of standard econometric models by fitting several multiple linear regression models. Then, we motivate the use of Elastic Nets and of forward-rolling validation methods to create our nowcaster through early experiments. Finally, we develop and optimize in a stepwise fashion nowcasting models to obtain the best out-of-sample predictive performances.

3.1. Exploratory analysis

23. Five multiple regression linear models are estimated, according to the following specifications:

$$(1): SWB_{c,t} = \beta_0 + \beta_1 . ECO_{c,t} + \varepsilon_{c,t}$$

$$(2): SWB_{c,t} = \beta_0 + \beta_1 . ECO_{c,t} + \beta_2 . COUNTRY_c + \varepsilon_{c,t}$$

$$(3): SWB_{c,t} = \beta_0 + \beta_1 . ECO_{c,t} + \beta_2 . COUNTRY_c + \beta_3 . TIME_t + \varepsilon_{c,t}$$

$$(4): SWB_{c,t} = \beta_0 + \beta_1 . ECO_{c,t} + \beta_2 . COUNTRY_c + \beta_4 . GOOGLE_{c,t} + \varepsilon_{c,t}$$

$$(5): SWB_{c,t} = \beta_0 + \beta_1 . ECO_{c,t} + \beta_2 . COUNTRY_c + \beta_3 . TIME_t + \beta_4 . GOOGLE_{c,t} + \varepsilon_{c,t}$$

24. In these models, SWBc,t is the average SWB for country c at time t, β0 is our intercept, ECOc,t is the vector of economic controls for country c at time t, COUNTRYc is a vector of country fixed effects, TIMEt is a vector of time fixed effects, GOOGLEc,t is a vector of Google Trends variables for country c at time t. Finally, εc,t is an error term. These 5 models are fitted with both yearly and monthly data, using the 3 different pre-processing of Google Trends data. Also, this exploratory analysis uses the Google Trends condensed database, while the full set of variables will be used for nowcasting at the last stage.

3.2. Training design for base learners

25. The final nowcaster has multiple key characteristics that are motivated by the following intermediary results drawn from a series of experiments: i) Google Trends data are included as they contribute to better predictions of country-specific accounts of monthly SWB; ii) Models are optimized with the minimization of the country-weighted mean squared errors (CW-MSE) instead of the simple mean squared errors (MSE) to account for country imbalance in the dataset; iii) a modified walk-forward validation is preferred to the traditional k-fold cross-validation as it is arguably a better way to measure the actual performance of the final nowcaster; iv) the final nowcaster relies on an ensemble model that stacks four different types of base models.

Base leaners

26. The four base learners are an Elastic Net regressor, an interacted Elastic Net regressor, a Gradient-Boosted Trees regressor, and a Multi-layer Perceptron regressor. The final meta-learner is a simple Elastic Net regressor (see below). This combination of linear and non-linear, and sparse and dense models is expected to generate the best predictions once stacked.

27. Elastic Net regressors refer to penalized regression models that combine both L1 and L2 penalties ((Zou and Hastie, 2005[23]). As compared to standard regressions, Elastic Net regressions can accommodate situations where the number of features is superior to the number of observations. Moreover, compared to Lasso regressions, they can also produce better solutions in cases where many relevant variables are correlated, without the need to exclude some of them. Indeed, we suspect that the best models in our situation are ones that predict SWB as a function of many variables with modest impacts, rather than models focused on finding sparse solutions. Given its two types of penalties, the Elastic Net regression can be considered as a generalization of Lasso and Ridge regressions as it allows analysts to optimise for both L1 and L2 regularization to produce the optimal solution. This allows us, by analysing the best hyper-parameters and observing whether the regularization is closer to Ridge or Lasso, to determine if our initial intuition was correct or not. In practice, we train two Elastic Net regressors, a simple one with the features and no interaction, and another one with the following first-order interactions: Country x Google variables, season x Google variables, and Economic variables x Google Variables. We also include all the possible first-order interactions between our Economic variables.

28. Aside linear regressions, we also include two non-linear models to be a part of our ensemble as we suspect that they may provide different predictions that could be beneficial for final predictions. The first non-linear model chosen is a Gradient-Boosted Trees (GBT) regression: it is an ensemble model, combining multiple regression trees via Boosting. Tree-based models combine regressions trees that sequentially split the dataset across multiple variables in order to minimize the impurity in the end leaves. The depth and complexity of the trees are controlled through regularization. GBT models tend to produce non-linear sparse solutions, which can accommodate non-linear phenomena given their highly interactive nature. For this paper, we use the implementation of GBT by the LightGBM package. One of the advantages of LightGBM over other frequently used implementations of GBT like XGBOOST is that it generates non-symmetrical trees. Although this increases the risk of overfitting, it also allows for the generation of more complex non-linear models. The second non-linear model chosen is a Multi-layer Perceptron regression. This is directly inspired by the methodology of (Woloszko, 2020[18]), who justified using a neural network regression given its stronger ability to map to the dense data-generating process of GDP, and to the many non-linear small contributions exerted by the variables his final model included. Likewise, we suspect that SWB, similar to GDP, is generated by a complex and dense data-generating process where many variables have a small impact.

Metric selection

29. CW-MSE is the average of country-specific MSE, which is preferred to the standard MSE, itself a micro-averaged metric (Japkowicz, 2013[24]).To justify the use of CW-MSE over MSE, we compare two prototypes of Elastic Net regressors, trained and optimized to either minimize CW-MSE or MSE. We emphasize in our demonstration the impact of the two metrics on the overall quality of predictions but also on the quality of country-specific predictions. Beyond performance-related motivations, this methodological choice is made for two key reasons: first, given that the data set is imbalanced in terms of the number of observations for the different countries, there is a risk of creating a biased model if one just tries to minimize out-of-sample MSE. This is why the use of this type of macro-averaged metrics, as opposed to micro-averaged ones, has been advocated for by (Japkowicz, 2013[24]) when facing imbalanced data problems. Second, given that this tool is supposed to be valuable to decision-makers in all OECD countries, our

nowcaster must perform for every country in the data set, and thus needs to be heavy penalized when it is inaccurate for certain countries despite being very accurate for others.

Validation scheme

30. K-fold cross-validation is a standard validation technique in Machine Learning that splits datasets into different segments called folds, and successively leaves one fold out as a test or validation set while using the combined remaining folds as training sets (Langford, 2005[25]). This method has the advantage over simpler train-test split validation that it maximizes the size of the training set, leading usually to better performances. Moreover, k-fold cross-validation implies testing several models on several test sets, allowing analysts to notice if the predictive performance of a model is uneven across the different test sets. Conversely, walk-forward validation can be described as a cross-validation designed for time series (Bergmeir and Benítez, 2012[26]). Specifically, instead of randomly splitting a dataset into different folds, this validation method splits data sets between training and testing sets in forward-rolling fashion. This means that models using this method are always trained on past data and tested on future data. This allows analysts to develop nowcasters and forecasters in more realistic environments before implementation (see Figure A.9 in SI for a visual description of the two strategies).

31. In this paper, the 4 base learners are trained, tested and optimized via sequential walk-forward validation on a subset containing 1,695 observations and optimised via k-fold cross-validation on the same subset (see Figure A.10 in SI for a summary of the methodology). The ensemble model is finally tested, alongside all our base-learners, on a test set which comprises the remaining 306 observations, all dated from 2010 onwards. All the splits that we operate with k-fold cross-validation and simple train-test split are stratified at the country-level. To justify the use of walk-forward validation over the traditional k-fold cross-validation, we again compare the performances of two prototypes of Elastic Net regressors which hyper-parameters are optimized using those two validation methods. In particular, we emphasize in this segment the risk of overconfidence that analysts run when they use k-fold cross-validation to train a regressor to predict time series.

3.3. Meta-learning with an ensemble model

Meta-learner model

32. The heterogeneity of base learners motivates the use of an ensemble model for the final nowcaster. The meta-learner that combines predictions is a simple Elastic Net regression model, with the following equation:

$$SWB_{c,t} = \beta_0 + \beta_1.\mathrm{BASE\_Y\_pred}_{c,t} + \Lambda_c + \Pi_c.\mathrm{BASE\_Y\_pred}_{c,t} + \varepsilon_{c,t}$$

33. where $\mathrm{BASE\_Y\_pred}_{c,t}$ are the base-learners' predictions and (Λ_c, Π_c) are country fixed-effects. Given that the 4 base-learners may not perform equally well across countries, we also include country-fixed effects and interactions between fixed-effects and the base-learners' predictions. This provides to the meta-learner with higher weights to certain base-learners for the countries on which they perform the best.

Meta-learner tuning

34. We devote a significant amount of time and consideration to the tuning of all models to be trained. We depart from (Woloszko, 2020[18]) by conducting a full-fledged optimisation of hyper-parameters, which in our view does not lead to overfitting when the validation process is specifically designed to prevent information leakage between train, validation, and test sets, as our protocol does.

35. In practice, the 4 base learners are trained, tested and optimized via sequential walk-forward validation on a subset containing 1,695 observations. The final ensemble model is trained and optimised via k-fold cross-validation on the same subset. The ensemble model is finally tested, alongside all our base-learners, on a test set which comprises the remaining 306 observations, all dated from 2010 onwards. All the splits that we operate with k-fold cross-validation and simple train-test split are stratified at the country-level.

36. Hyper-parameters are optimized through Bayesian Hyper-parameter Optimization in order to expedite the process and to allow the optimization to learn and focus more on promising areas in the high-dimensional hyper-parameters space. We specifically use Optuna for the Hyper-parameter Optimisation for its performance and flexibility, which also allows us to simultaneously optimize the nowcaster for other parameters, such as the types of pre-processing for the Google Trends data, or the type of interactions to include in our final model (Akiba et al., 2019[27]). The parameters and their associated spaces are reported for each model in Table A.2 in SI to allow for replication.

4 Results

4.1. Exploratory analysis

37. Results from the exploratory approach are promising with regards to the potential benefits of integrating Google Trends data to traditional models of SWB. For the yearly regression models, results suggest that the inclusion of composite Google Trends variables increase the adjusted R2, whether one includes time fixed-effects or not, and with country fixed-effects always included (Table 1). The effects associated with composite variables are smaller than those corresponding to macroeconomic variables, but still significant for many. The results are quite similar across the 3 types of pre-processing for the Google Trends data, with R2 and adjusted R2 suggesting that the pre-processing as per (Bokelmann and Lessmann, 2019[19]) benefits to the final regression specifications, in a modest but consistent fashion. As is frequently the case in such SWB models, the fixed-effects account for the majority of the explained variance. Finally, monthly regression models yield qualitatively similar results, where the R2 of regressions with macro variables alone equals 0.38, 0.73 when country dummies are added and 0.79 when Google Trends variables are added.

Table 1. Results of yearly regression models for the three types of pre-processing

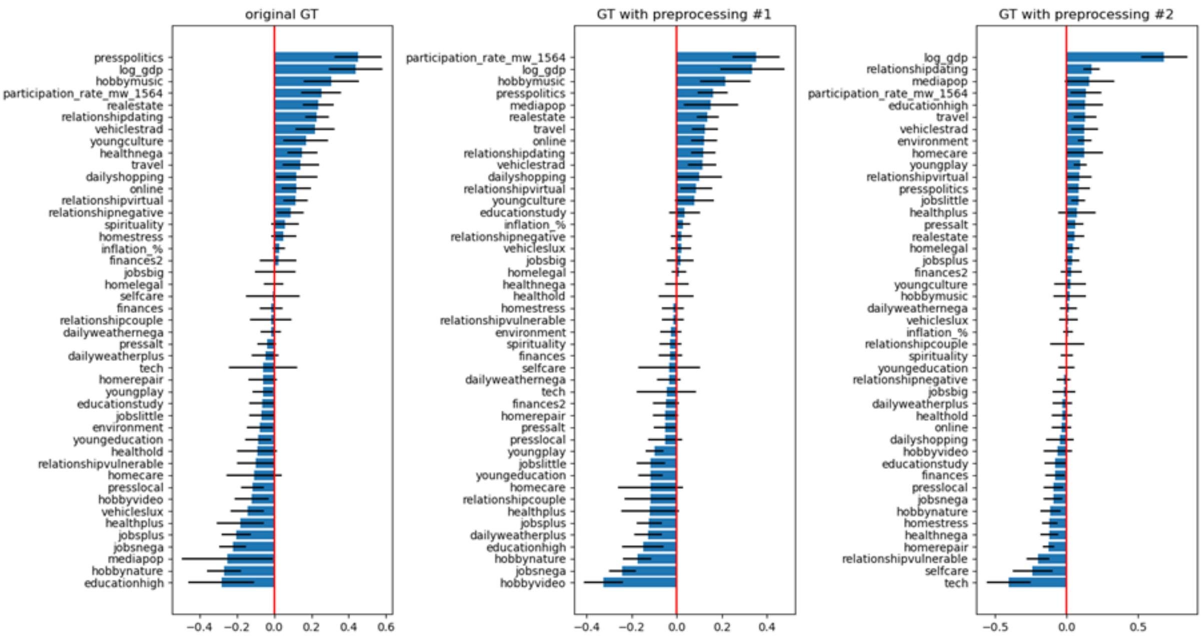
preprocessing	Original				Method 1				Method 2			
variables	country FE	country FE + Google	country FE + time FE	country FE + time FE + Google	country FE	country FE + Google	country FE + time FE	country FE + time FE + Google	country FE	country FE + Google	country FE + time FE	country FE + time FE + Google
R^2	0.830	0.886	0.858	0.892	0.830	0.889	0.858	0.895	0.830	0.878	0.858	0.885
Adjusted R^2	0.817	0.867	0.842	0.869	0.817	0.869	0.842	0.872	0.817	0.857	0.842	0.860

Source: OECD calculations. ‘Original’ refers to the raw Google Trends series, ‘Method 1’ to Bokelmann & Lessmann treatment, ‘Method 2’ to Wolozko treatment.

38. As can be seen on Figure 4, the interpretability of the coefficients on composite Google variables is encouraging. Indeed, many variables have the valence that one would expect. For instance, increases in variables like relationshipvulnerable, which combine multiple abuse-related categories and topics, are associated with lower SWB on average. This being said, the attempt to disentangle the positive and negative aspects of a specific theme failed for many composite variables: for example, jobsplus and jobsnega both end up being negatively associated with SWB. One speculative interpretation of this phenomenon is that people’s queries suggest what they are looking for, aspiring to and potentially missing at the moment. Hence, once could argue that increases in certain query terms are signs of deprivation. Moreover, a few variables’ valence switch sign across specifications, such as youngplay or environment, which likely reflects multi-collinearity effects.

Figure 4. Z-scores and standard errors of all variables of interest for the final specification of the yearly regression model

By type of pre-processing for Google Trends data



Note: Lines represent 95% confidence intervals.
Source: OECD calculations.

4.2. Training design for base learners

39. In this part, we report differences in performance across different methodological choices with a common regression model (the learner Reference ENET), or across different learners with a common methodological setup. Reference ENET corresponds to an Elastic Net regression model, which optimises its hyper-parameters, features and pre-processing method to minimize the OOS CW-MSE criterion derived from a walk-forward validation scheme.

40. The results reported in Table 2 are based on performances drawn from a large initial training split (N = 1,695), which is repeatedly split between further training and validation sets with k-fold cross-validation or walk-forward validation. Specifically, we test for differences that arise from: i) including Google Trends variables or not; ii) optimizing hyper-parameters for MSE or CW-MSE minimization for OOS prediction; iii) validating parameters through k-fold cross-validation or walk-forward validation; iv) choosing different base-learners, namely an Elastic Net with interactions, a Gradient-Boosted Tree and a Multi-layer Perceptron.

Table 2. Summary of experimental results related to the design of base learners

MODEL SPECIFICATION	OOS MSE	OOS CW-MSE	OOS R^2
Reference ENET	**0.1783**	**0.1909**	**0.7163**
Reference ENET without Google Trends variables	0.2036	0.2263	0.6761
Reference ENET optimized for MSE reduction	0.1787	0.1911	0.7158
Reference ENET optimized with K-Fold cross-validation	0.1868	0.1961	0.7028
ENET + interactions minimizing CW-MSE with Walk-Forward validation	0.2009	0.2224	0.6805
GBT minimizing CW-MSE with Walk-Forward validation	0.2339	0.2548	0.6279
MLP minimizing CW-MSE with Walk-Forward validation	0.1941	0.2088	0.6912

Source: OECD calculations. Reference ENET is an Elastic Net using Google Trends variables, the CW-MSE optimization criterion (instead of MSE) and walk-forward validation (instead of K-fold cross-validation).

41. In an out-of-sample prediction context, the results again suggest that Google Trends data can improve SWB predictions as compared to using only economic variables, resulting in a 12.3% reduction in MSE (from 0.2036 to 0.1783) and a 14.8% reduction in CW-MSE (from 0.2263 to 0.1909), see Figure B.2 in SI. Specifically, the reference model ENET performs best with a weak Lasso-type regularization, using the Google Trends variables with no complex pre-processing but for a standardization of variables and a reduction of the original features to 100 principal components via Principal Components Analysis (PCA). For all models, the best way to account for seasonality is to include quarter fixed-effects rather than month fixed-effects. Hyper-parameter tuning suggests that using weights to penalize more during training errors from under-sampled countries help reduce the country-weighted mean squared errors during validation.

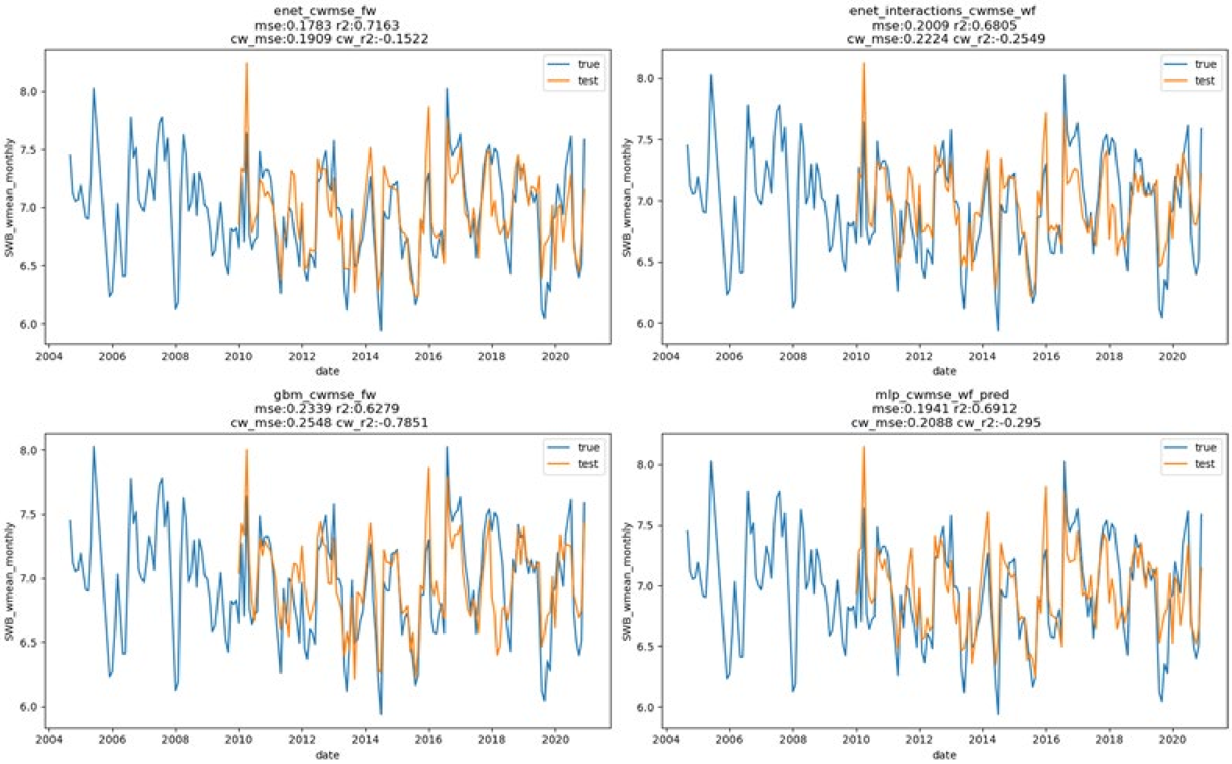
42. Results from this experiment suggest that our hyper-parameter optimization process converged towards a better solution when minimizing country-weighted mean squared error as compared to mean squared error. In particular, while the micro-averaged test metrics are quite similar between the two models, the macro-averaged ones suggest a clear advantage for the model optimized to minimize the macro-averaged country-weighted mean squared error (Figure B.3 in SI). This can be explained by the imbalanced number of observations per country in our dataset: Optimizing using country-weighted mean squared error ensures that our model is penalized if it performs well only on a few better-represented countries as compared to performing well on most countries. Hence, our reference ENET ends up probably being a better model of SWB that generalizes to more countries. This suggests that macro-averaged metrics should be considered by analysts when dealing with imbalanced data, especially in social sciences where these matters receive little attention.

43. Table 2 also shows that OOS results obtained from models optimized with k-fold cross-validation are worse than those drawn from the same model and the same parameters but with a walk-forward validation scheme (i.e. the Reference ENET). This happens even though both models had access to the same hyper-parameters space during the optimization step. This suggests that analysts should avoid relying solely on k-fold cross-validation when building nowcasting and forecasting models.

44. Finally, Table 2 reports the results for the 3 other base learners, namely the Elastic Net with interactions, the Gradient-Boosted Tree and the Multi-Layer Perceptron (Figure 5). It is striking that the reference Elastic Net regression model, without interactions, outperforms more complex models, suggesting that Google Trends data can be leveraged even using simple solutions by analysts for nowcasting and forecasting purposes.

Figure 5. Time series of true values versus out-of-sample predictions by 4 base learners – Averages over pooled sample of country-month observations

Base learners are an Elastic Net model, an Elastic Net model with interactions, a Gradient-Boosted Trees model, and a Multi-Layer Perceptron regressor (top left to bottom right)

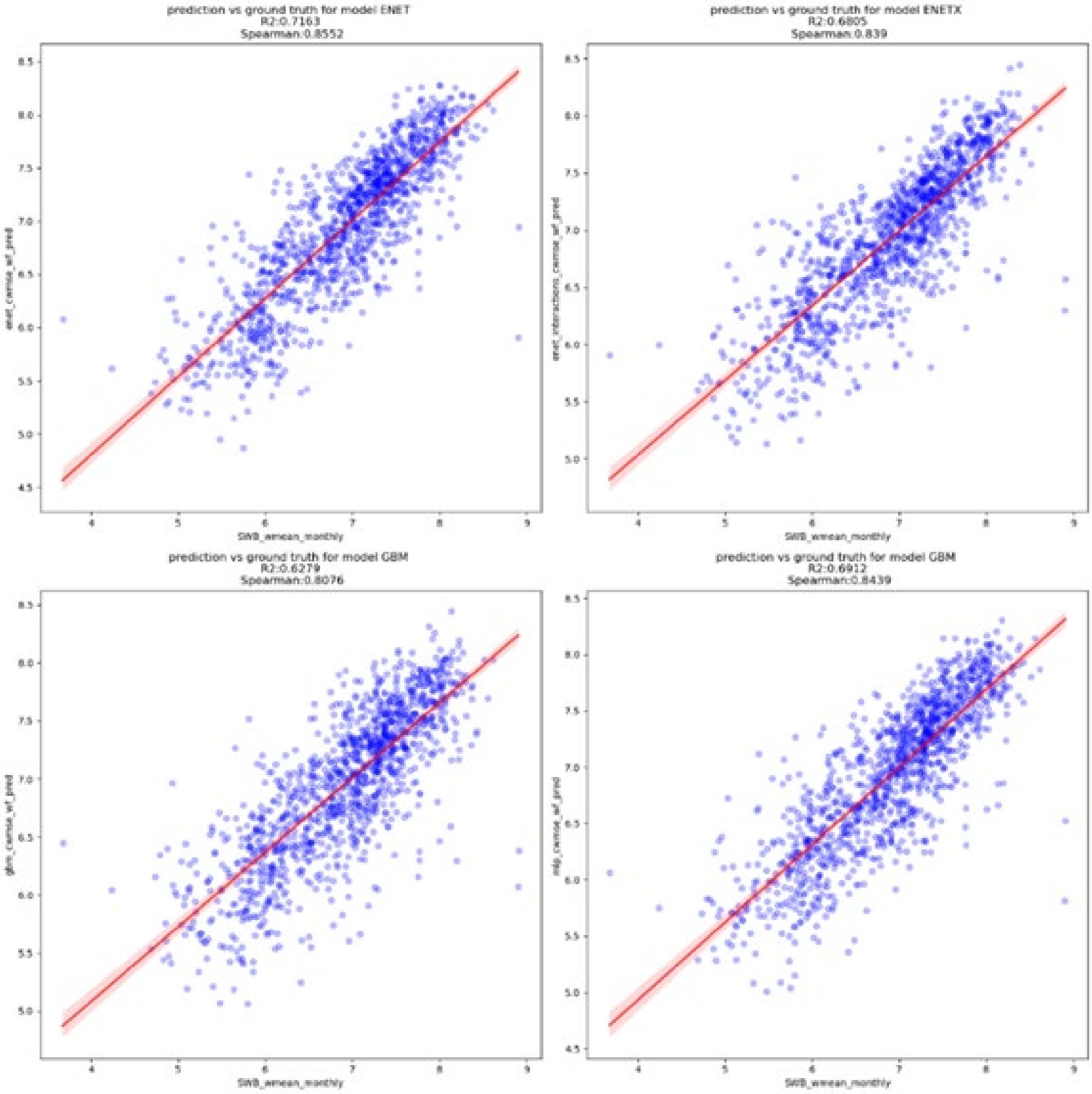


Note: Base learners are optimized via walk-forward validation to minimize CW-MSE.
Source: OECD calculations.

45. All the models still provide reasonably good predictions of SWB at the country-month level (Figure 6), with all models showing large and positive R2 and Spearman rank correlations coefficients in an out-of-sample context. Part of this finding may be explained by the fact that SWB is a stationary variable, with country fixed-effects playing a major role in models as shown initially using our simple econometric approach. However, these small changes could still translate into large variations with regards to the lived experiences of individuals, meaning that we cannot discount these differences observed over time.

Figure 6. Out-of-sample predictions of country-month observations against true values for the 4 base-learners

Base learners are an Elastic Net model, an Elastic Net model with interactions, a Gradient-Boosted Trees model, and a Multi-Layer Perceptron regressor (top left to bottom right)



Note: Base learners are optimized via walk-forward validation to minimize CW-MSE.
Source: OECD calculations.

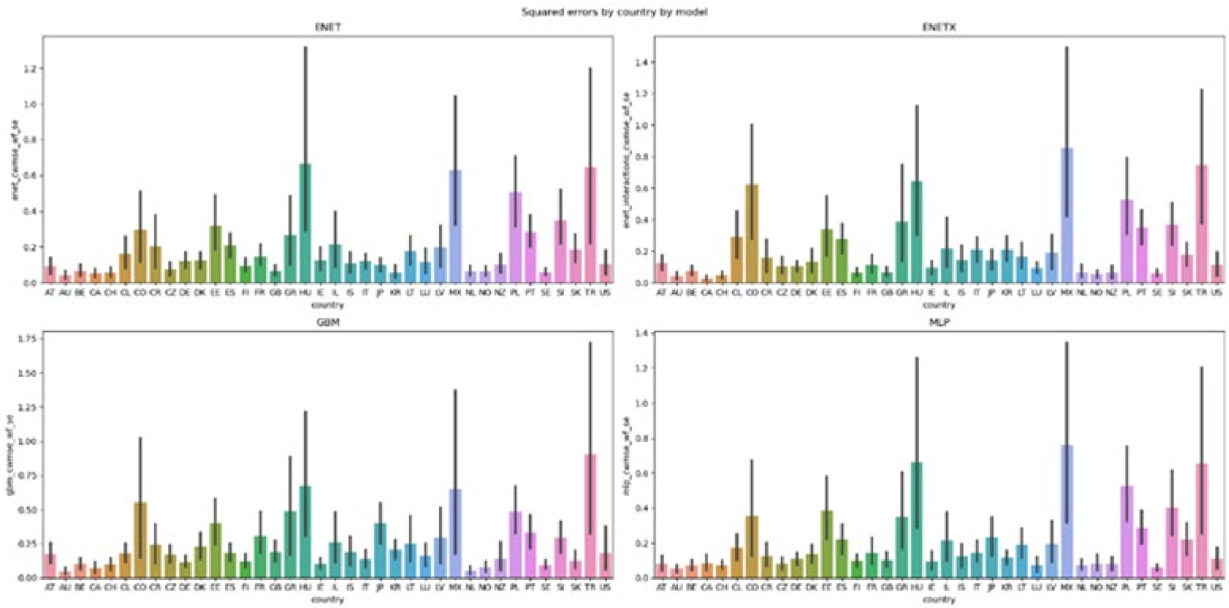
46. One last observation is that dense models tend to outperform sparse models, as shown by the two most performant models, namely an Elastic Net regression with a weak regularization as well as a neural network regression with 3 hidden layers. This means that SWB is better predicted by complex models including many variables able to interact, rather than by simpler sparser models like the GBT regression.

47. Figure 7 depicts differences in performance across countries and shows that the models perform extremely well on a majority of countries, but display MSEs that can be 5 to 10 times larger on outlier countries like Hungary, Mexico, Poland and Republic of Türkiye. There are two potential explanatory factors related to the imbalance issue mentioned several times. First, as can be seen on Figure B.5 in SI,

the squared errors of each model are negatively related to the number of country-months instances of each country in the data set. This is a promising result as it suggests that the nowcaster could be improved by either oversampling underrepresented countries, or, according to our preferred solution, by combining multiple national surveys with the aggregated database. Second, Figure B.6 in SI shows that errors are negatively correlated to the sample size used to compute the country-month average SWB values. This result again suggests that performance issues in the nowcaster are data-related and could be corrected by including additional data into the combined data set.

Figure 7. MSE of out-of-sample predictions by country and by base learner

Base learners are an Elastic Net model, an Elastic Net model with interactions, a Gradient-Boosted Trees model, and a Multi-Layer Perceptron regressor (top left to bottom right)



Note: Lines represent 95% confidence intervals.
Source: OECD calculations.

48. Finally, as can be seen on Figure B.7 in SI, the performance of the 4 models does not meaningfully change over time. This suggests that the nowcaster could potentially be successfully trained with a smaller but more balanced dataset. Moreover, this also indicates that the usefulness of Google Trends data do not change over time through changing meanings for terms for instance.

4.3. Meta-learning with an ensemble model

49. As shown on Figure 7, the performances of the 4 base models varies across countries, with some models performing better than others for certain countries. This suggests that the 4 base learners behave differently and can be potentially complementary with each other. This last point is further reinforced by Figure B.8, showing that the models' errors are not perfectly correlated with each other.

50. Out-of-sample performance of the meta-learner evaluated on the test set isolated at the beginning reveal large improvements over any single base-learner, as shown on Table 3. The meta-learner's MSE and CW-MSE are both 10% lower than the ones of the best base-learner. Looking at predictions for each country individually (Figure 8), results shows that the meta-learner has combined the strength of the

different base-learners, making it competitive against each base-learner for any specific country. Improvements remain to be made however for certain countries, as seen on Figure B.9: Republic of Türkiye is still extremely poorly predicted and predictions for Australia, Chile, Colombia, Greece, Iceland, Japan, Korea, Luxembourg, New Zealand and Poland could be further improved. The countries cited should be the ones for which national surveys should be integrated the soonest in order to improve prediction performance.

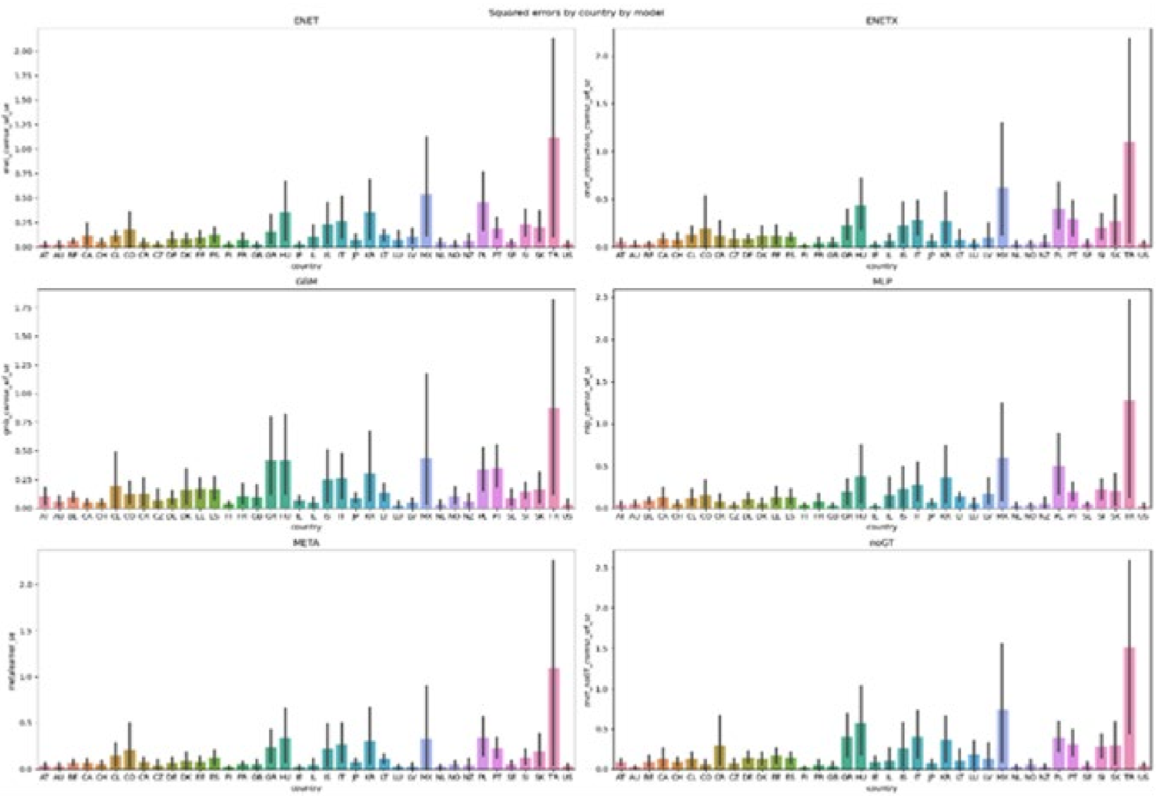
Table 3. In-sample and out-of-sample performance of base learners and meta learner

	Training set			Testing set		
	MSE	CW-MSE	R^2	MSE	CW-MSE	R^2
Elastic Net	0.1192	0.1278	0.8163	0.141	0.158	0.8018
Elastic Net with interactions	0.106	0.1128	0.8356	0.1466	0.1644	0.8093
Gradient-Boosted Tree	0.0797	0.0832	0.8712	0.1557	0.167	0.7933
Neural Network	0.133	0.1419	0.7983	0.1521	0.1716	0.7798
Meta Learner	**0.070**	**0.0727**	**0.8883**	**0.1269**	**0.1422**	**0.8304**

Note: The Meta learner is obtained from an Elastic Net applied to 4 base learners, namely an Elastic Net model, an Elastic Net model with interactions, a Gradient-Boosted Trees model, and a Multi-Layer Perceptron regressor.
Source: OECD calculations.

Figure 8. Out-of-sample MSE by country for the Meta learner

With 95% confidence interval



Note: The Meta learner is obtained from an Elastic Net applied to 4 base learners, namely an Elastic Net model, an Elastic Net model with interactions, a Gradient-Boosted Trees model, and a Multi-Layer Perceptron regressor.
Source: OECD calculations.

5 Conclusion

51. This paper provides an extensive description of the data engineering required to nowcast average SWB among a large number of OECD countries. Adequate prediction two-years ahead (with one year of data embargo) is achieved thanks to the help of a sophisticated Meta-Learning methodology that includes: i) an extensive use of large custom micro databases for enhanced training; ii) detailed investigation of pre-processing methods for Google Trends data; ii) Bayesian optimization of hyper-parameters and usage of a country-weighted MSE criterion for out-of-sample predictions; iii) use of walk-forward validation instead of k-fold cross-validation; iv) choice of different base-learners, in practice an Elastic Net with and without interactions, a Gradient-Boosted Tree and a Multi-layer Perceptron; v) estimation of a meta-learner with country fixed-effects interactions with the base learners predictions. As a result, across 38 countries over the 2010-2020 period, the out-of-sample prediction of average SWB reaches a R2 of 0.8304, with 95% of true country-month averages being within 0.664 of our final out-of-sample predictions.

52. There are several ways of improving this meta-learner. First, it would be necessary to oversample countries with poor performance in order to increase the number of country-month observations, and/or the sample size at the individual level used to compute each country-month observation. In practice, this can be done by including national social surveys in the combined data set. Second, more macro-economic control variables could have been included. Third, new ways of pre-processing the Google Trends variables could be experimented. In this regard, we found that our simple method of using only 100 Principal Components outperformed more complex methodologies described in the literature.

53. Besides, this meta-learning approach could be used to model different dependent variables, such as quantiles of the SWB distribution (i.e. people with relatively low or high SWB) or shares of people under or above a given SWB threshold. Likewise, other outcomes could be modelled, such as shares of people with negative or positive affect, trust in others or in government. Finally, this methodology could be extended to capture regional well-being as geographical information is available in both survey data and Google Trends data.

References

Akiba, T. et al. (2019), *Optuna: A Next-generation Hyper-parameter Optimization Framework*, https://doi.org/10.1145/3292500.3330701. [26]

Algan, Y. et al. (2018), *The Rise of Populism and the Collapse of the Left-Right Paradigm: Lessons from the 2017 French Presidential Election.*, https://papers.ssrn.com/abstract=3235586. [2]

Algan, Y. et al. (2019), "Well-being through the lens of the internet", *PLOS ONE*, Vol. 14/1, https://doi.org/10.1371/JOURNAL.PONE.0209562. [17]

Auer, P. and R. Meir (eds.) (2005), *The Cross Validation Problem*, Springer, Berlin, Heidelberg., https://doi.org/10.1007/11503415_47. [24]

Bergmeir, C. and J. Benítez (2012), "On the use of cross-validation for time series predictor evaluation", *Information Sciences*, Vol. 191, pp. 192–213, https://doi.org/10.1016/J.INS.2011.12.028. [25]

Böckerman, P., J. Laamanen and E. Palosaari (2016), "The Role of Social Ties in Explaining Heterogeneity in the Association Between Economic Growth and Subjective Well-Being", *Journal of Happiness Studies*, Vol. 17/6, pp. 2457–2479, https://doi.org/10.1007/S1090. [10]

Bokelmann, B. and S. Lessmann (2019), "Spurious patterns in Google Trends data: An analysis of the effects on tourism demand forecasting in Germany", *Tourism Management*, Vol. 75, pp. 1-12, https://doi.org/10.1016/J.TOURMAN.2019.04.015. [19]

Brodeur, A. et al. (2021), "COVID-19, lockdowns and well-being: Evidence from Google Trends", *Journal of Public Econonomics*, Vol. 193, https://doi.org/10.1016/j.jpubeco.2020.104346. [16]

Cheung, F. and R. Lucas (2014), "Assessing the validity of single-item life satisfaction measures: results from three large samples", *Quality of Life Research*, Vol. 23/10, pp. 2809–2818, https://doi.org/10.1007/s11136-014-0726-4. [6]

Deaton, A. and A. Stone (2016), "Understanding context effects for a measure of life evaluation: how responses matter", *Oxford Economic Papers*, Vol. 68/4, pp. 861–870, https://doi.org/10.1093/OEP/GPW022. [21]

De, P. and D. Thamarapani (2022), "Impacts of negative shocks on wellbeing and aspirations – Evidence from an earthquake", *World Development*, Vol. 154, https://doi.org/10.1016/J.WORLDDEV.2022.105876. [11]

Diener, E. (ed.) (2009), *Beyond money: Toward an economy of well-being*, Springer Science + Business Media., https://doi.org/10.1007/978-90-481-23. [8]

Diener, E. et al. (1999), “Subjective well-being: Three decades of progress”, *Psychological Bulletin*, Vol. 125/2, pp. 276–302, https://doi.org/10.1037/0033-2909.125.2.276. [5]

Dustmann, C. and F. Fasani (2014), “The effect of local area crime on mental health”, *The Economic Journal*, Vol. 126, pp. 978–1017, https://doi.org/10.1111/ecoj.12205. [20]

Foa, R., M. Fabian and S. Gilbert (2022), “Subjective well-being during the 2020-21 global coronavirus pandemic: Evidence from high frequency time series data”, *PLoS One*, Vol. 17/2, https://doi.org/10.1371/journal.pone.0263570. [15]

Foa, R., S. Gilbert and M. Fabian (2020), *COVID-19 and Subjective Well-Being: Separating the Effects of Lockdowns from the Pandemic*, https://doi.org/10.17863/CAM.90181. [14]

Hariri, J., C. Bjørnskov and M. Justesen (2016), “Economic Shocks and Subjective Well-Being: Evidence from a Quasi-Experiment”, *The World Bank Economic Review*, Vol. 30/1, pp. 55–77, https://doi.org/10.1093/WBER/LHV004. [12]

He, H. and Y. Ma (eds.) (2013), *Assessment Metrics for Imbalanced Learning. Imbalanced Learning: Foundations, Algorithms, and Applications.*, Wiley, https://doi.org/10.1002/9781118646106.CH8. [23]

Kahneman, D. and A. Krueger (2006), “Developments in the Measurement of Subjective Well-Being”, *Journal of Economic Perspectives*, Vol. 20, pp. 3–24, https://www.aeaweb.org/journals/jep. [7]

Kaiser, C., A. Oswald and R. Easterlin (2022), “The scientific value of numerical measures of human feelings”, *Proceedings of the National Academy of Sciences*, Vol. 119/42, https://doi.org/10.1073/PNAS.2210412119. [1]

Lucas, R., P. Dyrenforth and E. Diener (2008), “Four Myths about Subjective Well-being”, *Social and Personality Psychology*, Vol. 2, pp. 2001-2015., https://doi.org/10.1111/j.1751-9004.2008.00140.x. [13]

OECD (2021), *COVID-19 and Well-being: Life in the Pandemic*, OECD Publishing, Paris, https://doi.org/10.1787/1e1ecb53-en. [4]

OECD (2020), *How’s Life? 2020: Measuring Well-being*, OECD Publishing, Paris., https://doi.org/10.1787/9870c393-en. [3]

OECD (2013), *OECD Guidelines on Measuring Subjective Well-being*, OECD Publishing, Paris, https://doi.org/10.1787/9789264191655-en. [27]

Stevenson, B. and J. Wolfers (2008), “Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox”, *Brookings Papers on Economic Activity*, Vol. 39/1, pp. 1-102. [9]

Woloszko, N. (2020), “Tracking activity in real time with Google Trends”, *OECD Economics Department Working Papers*, No. 1634, OECD Publishing, Paris, https://doi.org/10.1787/6b9c7518-en. [18]

Zou, H. and T. Hastie (2005), “Regularization and variable selection via the elastic net”, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 67/2, pp. 301-320, https://doi.org/10.1111/J.1467-9868.2005.00503.X. [22]