scientific reports

Check for updates

OPEN

Product progression: a machine learning approach to forecasting industrial upgrading

Giambattista Albora[1,2], Luciano Pietronero[2], Andrea Tacchella[3] & Andrea Zaccaria[2,4]✉

Economic complexity methods, and in particular relatedness measures, lack a systematic evaluation and comparison framework. We argue that out-of-sample forecast exercises should play this role, and we compare various machine learning models to set the prediction benchmark. We find that the key object to forecast is the activation of new products, and that tree-based algorithms clearly outperform both the quite strong auto-correlation benchmark and the other supervised algorithms. Interestingly, we find that the best results are obtained in a cross-validation setting, when data about the predicted country was excluded from the training set. Our approach has direct policy implications, providing a quantitative and scientifically tested measure of the feasibility of introducing a new product in a given country.

In her essay *The Impact of Machine Learning on Economics*, Susan Athey states: "Prediction tasks [...] are typically not the problems of greatest interest for empirical research in economics, who instead are concerned with causal inference " and "economists typically abandon the goal of accurate prediction of outcomes in pursuit of an unbiased estimate of a causal parameter of interest "[1]. This situation is mainly due to two factors: the need to ground policy prescriptions[2,3] and the intrinsic difficulty to make correct predictions in complex systems[4,5]. The immediate consequence of this behavior is the flourishing of different or even contrasting economic models, whose concrete application largely relies on the specific skills, or biases, of the scholar or the policymaker[6]. This *horizontal* view, in which models are every time aligned and selected, in contrast with the *vertical* view of hard sciences, in which models are selected by comparing them with empirical evidence, leads to the challenging issue of distinguishing which models are wrong. While this situation can be viewed as a natural feature of economic and, more in general, complex systems[6], a number of scholars coming from hard sciences have recently tackled these issues, trying to introduce concepts and methods from their disciplines in which models' falsifiability, tested against empirical evidence, is *the* key element. This innovative approach, called *Economic Fitness and Complexity*[7–12] (EFC), combines statistical physics and complex network based algorithms to investigate macroeconomics with the aim to provide testable and scientifically valid results. The EFC methodology studies essentially two lines of research: indices for the competitiveness of countries and relatedness measures.

The first one aims at assessing the industrial competitiveness of countries by applying iterative algorithms to the bipartite network connecting countries to the products they competitively export[13]. Two examples are the Economic Complexity Index ECI[14] and the Fitness[7]. In this case, the scientific soundness of either approach can be assessed by accumulating pieces of evidence: by analyzing the mathematical formulation of the algorithm and the plausibility of the resulting rankings[15–18], and by using the indicator to predict other quantities. In particular, the Fitness index, when used in the so-called Selective Predictability Scheme[19], provides GDP growth predictions that outperform the ones provided by the International Monetary Fund[10,20]. All these elements concur towards the plausibility of the Fitness approach; however, a *direct* way to test the predictive performance of these indicators[21] is still lacking. This naturally leads to the consideration of further indices, that can mix the existing ones[22] or use new concepts such as information theory[23]. We argue that, on the contrary, the scientific validity of relatedness indicators can be univocally assessed, and this is the purpose of the present work.

The second line of research in EFC investigates the concept of Relatedness[24], the idea that two human activities are, in a sense, *similar* if they share many of the capabilities needed to be competitive in them[25]. Practical applications are widespread and include international trade[11,26], firm technological diversification[27,28], regional smart specialization[29,30], and the interplay among the scientific, technological, and industrial layers[31]. Most of these contributions use relatedness not to forecast future quantities, but as an independent variable in a

[1]Dipartimento di Fisica, Universitá Sapienza, Rome, Italy. [2]Centro Ricerche Enrico Fermi, Rome, Italy. [3]Joint Research Centre, Seville, Spain. [4]Istituto dei Sistemi Complessi-CNR, UOS Sapienza, Rome, Italy. ✉email: andrea.zaccaria@cnr.it

regression, and so the proximity (or quantities derived from it) is used to explain some observed simultaneous behavior. We point out, moreover, that no shared definition of relatedness exists, despite the widespread use of co-occurrences, since different scholars use different normalizations, null models, and data, so the problem to decide "which model is wrong" persists. For instance, Hidalgo et al.[26] base the goodness of their measure on its correlation with the probability that a country starts to export a product. O'Clery et al.[32] test the goodness of their relatedness measure through an in-sample logit regression; in this way models with a greater number of parameters (as provided, for instance, by the addition of fixed effects on countries and products) tend to have greater scores. Finally, Gnecco et al.[33] propose an approach to assess the relatedness based on matrix completion. Note that their test of the goodness of their approach is based on the reconstruction of the country-product pairs that have been removed from the data; the approach used here, instead, consists into looking at how good the proposed model is to guess new exports of countries after 5 years. So once again the performances are not comparable, as it is evident by looking, for instance, at the respective magnitude of the reported F1 scores.

The examples just discussed clarify why we believe that it is fundamental to introduce elements of falsifiability in order to compare the different existing models, and that such comparison should be made by looking at the performances in out-of-sample forecasting, that is the focus of the present paper. We will consider export as the economic quantity to forecast because most of the indicators used in economic complexity are derived from export data, being it regarded as a global, summarizing quantity of countries' capabilities[10,34] but also for the immediate policy implications of the capability to be able, for instance, to predict in which industrial sector a country will be competitive, say, in five years.

In this paper, we propose a procedure to systematically compare different prediction approaches and, as a consequence, to scientifically validate or falsify the underlying models. Indeed, some attempts to use complex networks or econometric approaches to predict exports exist[32,35–37], but these methodologies are practically impossible to compare precisely because of the lack of a common framework to choose how to preprocess data, how to build the training and the test set, or even which indicator to use to evaluate the predictive performance. In the following, we will systematically scrutiny the steps to build a scientifically sound testing procedure to predict the evolution of the export basket of countries. In particular, we will forecast the presence or the activation of a binary matrix element M_{cp}, that indicates whether the country c competitively exports product p in the Revealed Comparative Advantage sense[38] (see "Methods" for a detailed description of the export data).

Given the simultaneous presence in the literature of different approaches to measure the relatedness, it is natural to argue whether machine learning algorithm might play a role and build comparable or even better measures. In particular, given the present ubiquitous and successful use of artificial intelligence in many different contexts, it is natural to use machine learning algorithms to set the benchmark. A relevant by-product of this analysis is the investigation of the statistical properties of the database (namely, the strong auto-correlation and class imbalance), that has deep consequences on the choice of the most suitable algorithms, testing exercises, and performance indicators.

Applying these methods we find two interesting results:

1. The best performing models for this task are based on decision trees. A fundamental property that separates these algorithms from the main approaches used in the literature[26] is the fact that here the presence of a product in the export basket of a country can have a negative effect on the probability of exporting the target product. i.e. decision trees are able to combine Relatedness and Anti-Relatedness signals to provide strong improvements in the accuracy of predictions[39]
2. Our best model performs better in a cross-validation setting where we exclude data from the predicted country from the training set. We interpret this finding by arguing that in cross-validation the model is able to better learn the actual Relatedness relationships among products, rather than focusing on the very strong self-correlation of the trade data.

In the "Methods" section we show a detailed comparison between our machine learning based approach and some of the other definitions of relatedness we mentioned.

The present investigation of the predictability of the time evolution of export baskets has a number of practical and theoretical applications. First, predicting the time evolution of the export basket of a country needs, as an intermediate step, an assessment of the likelihood that the single product will be competitively exported by the country in the next years. This likelihood can be seen as a measure of the *feasibility* of that product, given the present situation of that country. The possibility to investigate with such a great level of detail which product is relatively *close* to a country and which one is out of reach has immediate implications in terms of strategic policies[40]. Second, the study of the time evolution of the country-product bipartite network is key to validate the various attempts to model it[41,42]. Finally, the present study represents on of the first attempts to systematically investigate how machine learning techniques can be applied in development economics, that is something still not much discussed in literature with except to very recent works[33,39,43].

Results

Statistical properties of the country-product network. A key result of the present investigation is a clear-cut methodology to compare different models or predictive approaches in Economic Complexity. In order to understand the reasons behind some of the choices we made in building the framework, we first discuss some statistical properties of the data we will analyze.

Our database is organized in a set of matrices $\mathbf{V}$ whose element V_{cp} is the amount, expressed in US dollars, of product p exported by country c in a given year. When not otherwise specified, the number of countries is 169, the number of products is 5040, and the time range covered by our analysis is 1996-2018. We use the HS1992,

6-digits classification. The data are obtained from the UN-COMTRADE database and suitably cleaned in order to take into account the possible disagreements between importers' and exporters' declarations (see "Methods"). We compute the Revealed Comparative Advantage[38] to obtain a set of RCA matrices **R** and, by applying a threshold equal to 1, a set of matrices **M** whose binary elements are equal to 1 if the given country competitively exports the given product. Here and in the following we use "competitively" in the Balassa sense, that is, $R_{cp} > 1$. In this paper we will discuss the prediction of two different situations: the unconditional presence of a "1" element in the **M** matrix and the *appearance* of such an element requiring that the RCA values were below a non-significance threshold t=0.25 in all the previous years. We will refer to the first case as the *full matrix* and to the new product event as an *activation*. The definition of the activation is somehow arbitrary: one could think, for instance, to change the threshold t or the number of inactive years. We find however our choice to be a good trade-off to have both a good numerosity of the test set and avoid the influence of trivial 0/1 flips. We point out that our final aim is to detect, as much as possible, the appearance of really new products in the export basket of countries.

In Fig. 1, left, we plot the probability that a matrix element M_{cp} in 1996 will change or not change its binary value in the future years. One can easily see that even after 5 years the probability that a country remains competitive in a product is relatively high (~ 0.64); being the probability that a country remains not competitive ~ 0.95, we conclude that there is a very strong auto-correlation: this is a reflection of the persistent nature of both the capabilities and the market conditions that are needed to competitively export a product. Moreover, the appearance of a new product in the export basket of a country is a rare event: the empirical frequency is about 0.047 after 5 years. A consequence of this persistence is that we can safely predict the presence of a 1 in the **M** matrices by simply looking at the previous years, while the appearance of a new product that was not previously exported by a country is much more difficult and, in a sense, more interesting from an economical point of view, since it depends more on the presence of suitable, but unrevealed, capabilities in the country; but these capabilities can be traced by looking at the other products that country exports. Not least, an early detection of a future activation of a new product has a number of practical policy implications. Note in passing that, since machine learning based smoothing procedures[10,44] may introduce extra spurious correlations, they should be avoided in prediction exercises, and so only the RCA values directly computed from the raw export data are considered.

On the right side of Fig.1 we plot the density of the matrices **M**, that is the number of nonzero elements with respect to the total number of elements. This ratio is roughly 10%. This means that both the prediction of the *full*, unconditional matrix elements and the prediction of the so-called *activations* (i.e., conditioning to that element being 0 and with RCA below 0.25 in all the previous years) show a strong class imbalance. This has deep consequences regarding the choice of the performance indicators to compare the different predictive algorithms. For instance, the ROC-AUC score[45], one of the most used measures of performance for binary classifiers, is well known to suffer from strong biases when a large class imbalance is present[46]. More details are provided in the "Methods" sections.

Recognize the country vs. learning the products' relations. In this section we present the results concerning the application of different supervised learning algorithms. The training and the test procedures are fully described in the "Methods" section. Here we just point out that the training set is composed by the matrices $\mathbf{R}^{(y)}$ with $y \in [1996 \ldots 2013]$, and the test is performed against $\mathbf{M}^{(2018)}$, so we try to predict the export basket of countries after $\Delta = 5$ years.

The algorithms we tested are XGBoost[47,48], a basic Neural Network implemented using the Keras library[49] and the following algorithms implemented using the scikit learn library[50]: Random Forest[51], Support Vector Machines[52], Logistic Regression[53], a Decision Tree[54], ExtraTreesClassifier[55], ADA Boost[56] and Gaussian Naive Bayes[57]. For reasons of space, we cannot discuss all these methods here. However, a detailed description can be found in[58] and references therein and, in the following sections, we will elaborate more on the algorithms based on decision trees, which result to be the most performing ones.

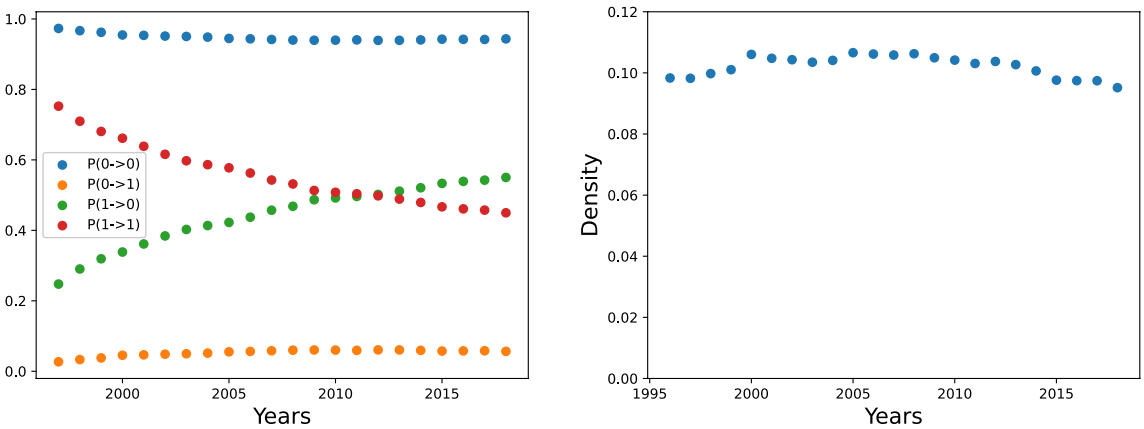

Figure 1. Left: transition probabilities between the binary states of the export matrix **M**. The strong persistency implies the importance of the study of the appearance of new products (called *activations*) with respect to the unconditional presence of one matrix element (in the following, *full matrix*). Right: the fraction of nonzero elements in **M** as a function of time. A strong class imbalance is present.

In Fig. 2 we show an example of the dynamics that our approach is able to unveil. On the left we show the RCA of Bhutan for the export of Electrical Transformers as a function of time. RCA is zero from 1996 to 2016, when a sharp increase begins. Was it possible to predict the activation of this matrix element? Let us train our machine learning algorithm XGBoost using the data from 1996 to 2012 to predict which products will Bhutan likely export in the future. The result is a set of scores, or *progression probabilities*, one score for each possible product. Each of these scores measures the feasibility, or relatedness, between Bhutan and all the products it does not export. The distribution of such scores is depicted in Fig. 2 on the right. The progression probability for Electrical Transformers was much higher than average, as shown by the arrow: this means that, already in 2012, Bhutan was very close to this product. Indeed, as shown by the figure on the left, Bhutan will start to export that specific product in about 5 years. Obviously, this is just an example, so we need a set of quantitative tools to measure the prediction performance on the whole test set on a statistical basis.

In order to quantitatively assess the goodness of the prediction algorithms, a number of performance indicators are available from the machine learning literature of binary classifiers. Here we focus on three of them, and we show the results in Fig. 3, where we show a different indicator in each row, while the two columns refer to the two prediction tasks, *full matrix* (i.e., the presence of a matrix element equal to one) and *activations* (a zero matrix element, with RCA below 0.25 in previous years, possibly becoming higher than one, that is the appearance of a new product in the export basket of a country). AUC-PR[46] gives a parameter-free, comprehensive assessment of the prediction performance. The F1 Score[59,60] is a harmonic mean of the Precision and Recall measures[61], and so takes into account both False Positives and False Negatives. Finally, mean Precision@10 considers each country separately and computes how many products, on average, are actually exported out of the top 10 predicted. All the indicators we used are discussed more in detail in the "Methods" section.

We highlight with a red color the RCA benchmark model, which simply uses the RCA values in 2013 to predict the export matrix in 2018. From the analysis of Fig. 3 we can infer the following points:

1. The performance indicators are much higher for the full matrix. This means that predicting the unconditional presence of a product in the export basket of a country is a relatively simple task, being driven by the strong persistence of the **M** matrices through the years.
2. On the contrary, the performance on the activations is relatively poor: for instance, on average, less than one new product of out the top ten is correctly predicted.
3. Algorithms based on ensembles of trees perform better than the benchmark and the other algorithms on all the indicators.
4. Thanks to the strong autocorrelation of the matrices, the RCA-based prediction represents a very strong benchmark, also in the case of the activations. However, Random Forest, ExtraTreesClassifier and XGBoost perform better both in the full matrix prediction task and in the activations prediction task.

We speculate that the machine learning algorithms perform much better in the full matrix case because, in a sense, they *recognize* the single country and, when inputted with a similar export basket, they correctly reproduce the strong auto-correlation of the export matrices. We can deduce that using this approach we are not learning the complex interdependencies among products, as we should, and, as a consequence, we do not correctly predict the new products. In order to overcome this issue, we have to use a k-fold Cross Validation (CV): we separately train our models to predict the outcome of k countries using the remaining $C - k$, where in our case $C = 169$ and $k = 13$. In this way, we prevent the algorithm to recognize the country, since the learning is performed on disjoint sets; as a consequence, the algorithm learns the relations among the products and is expected to improve the performances on the activations.

Using the cross validation procedure, we trained again the three best performing algorithms which are the Random Forest, ExtraTreesClassifier, and XGBoost. The result is that only the XGBoost algorithm improves

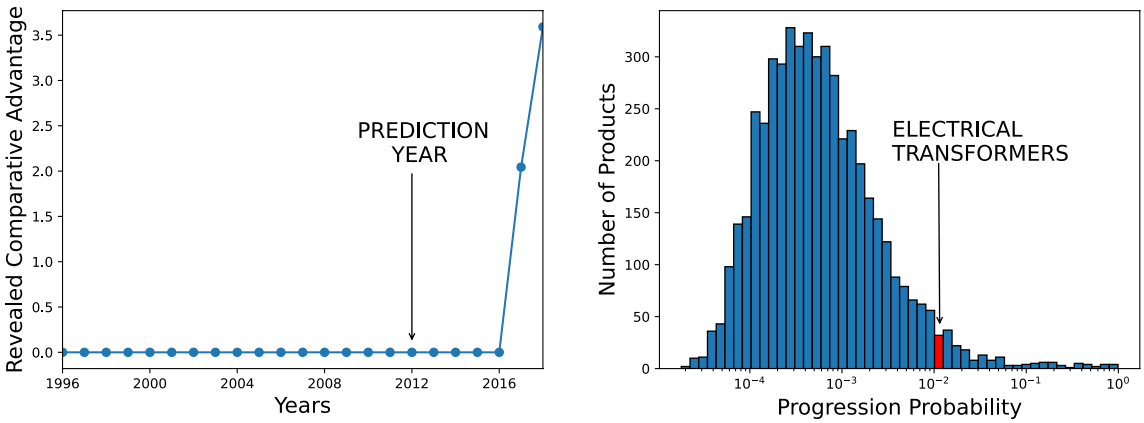

Figure 2. An example of successful prediction. On the left, the RCA of Bhutan in electrical transformers as a function of time. Already in 2012, with RCA stably below 1, the progression probability of that matrix element was well above its country average, as shown by the histogram in the figure on the right. Bhutan will start to competitively export electrical transformers after 5 years.

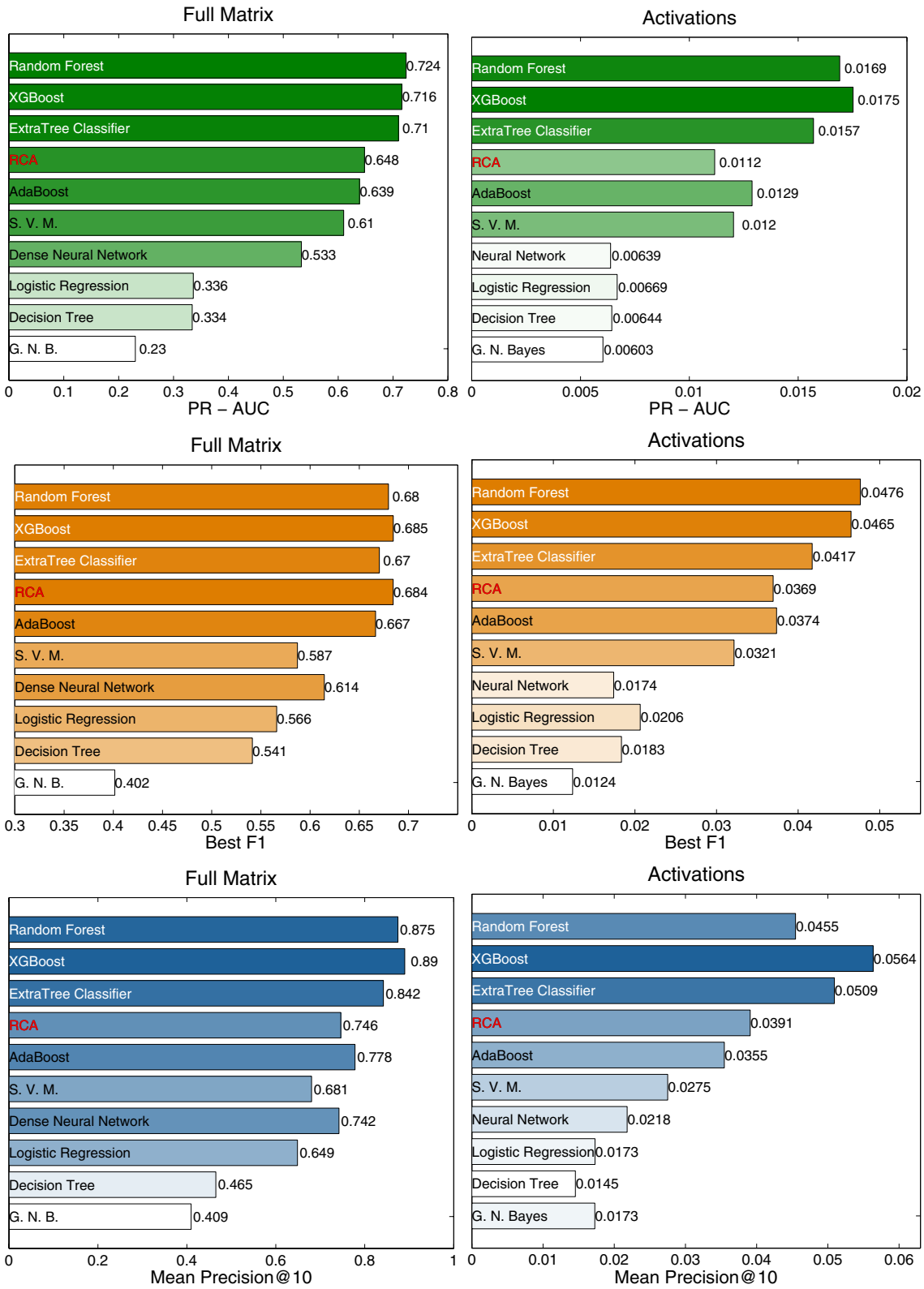

Figure 3. Comparison of the prediction performance of different algorithms using three performance indicators. Tree-based approaches are performing better; the prediction of the activations is a harder task with respect to the simple future presence of a product.

its scores, which means that in the cross-validation setting it is more capable to learn the inter-dependencies among products. So what is happening is that, if we do not perform the cross validation, the Random Forest tends to recognize the countries better than XGBoost, but if we perform the cross validation XGBoost learns the inter-dependencies among products better than the Random Forest. This step is crucial if one wants to build a representation of such interdependencies which has also a good forecasting power[39].

In Fig. 4 (left) we show the relative improvements of various performance indicators when the CV is used to train the XGBoost model and the test is performed on the activations. All indicators improve; in particular, F1-score and mean Precision@10 increase by more than 10%. On the right, we compare the cross-validated XGBoost predictions with the RCA benchmark, showing a remarkable performance although the previously noted goodness of the benchmark.

In Table 1 we report the values of the performance indicators for the non cross-validated Random Forest, the cross-validated XGBoost and the RCA benchmark model, once again tested on the activations. The last four rows represent the confusion matrix, where the threshold on the prediction scores is computed by optimizing the F1 scores.

The cross validated XGBoost gives the best scores except for the AUC-ROC and the accuracy which are influenced by the high class imbalance because of the large number of True Negatives, making these metrics unsuitable for evaluating the goodness of the predictions. However, the non cross-validated Random Forest is comparable and in any case shows better scores than the RCA benchmark, so it represents a good alternative, especially because of the much lower computational cost. Indeed, the inclusion of the cross-validation procedure increases the computational cost by about a factor 13, moreover, if compared with the same number of trees, Random Forest is 7.7 times faster than XGBoost. So, even if the cross validated XGBoost model is the best performing, the non cross validated Random Forest is a good compromise to have good predictions in less time.

In general, a desirable output of a classification task is not only a correct prediction, but also an assessment of the likelihood of the label, in this case, the activation. This likelihood provides a sort of confidence in the prediction. In order to test whether the scores are correlated or not with the actual probability of activations we

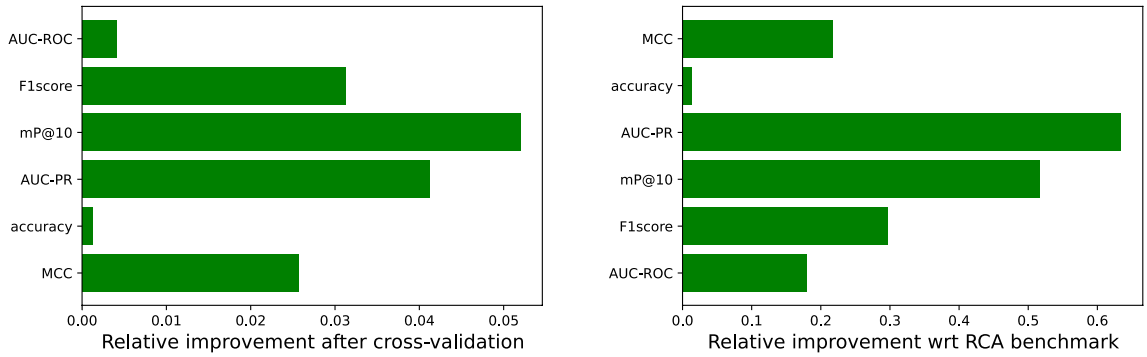

Figure 4. Left: relative improvement of the prediction performance of XGBoost when the training is cross validated. The algorithm now can not recognize the country, and so all the performance indicators improve. Right: relative improvement of the cross validated XGBoost algorithm with respect to the RCA benchmark.

Algorithm	XGBoost-CV	Random Forest	RCA
AUC-ROC	0.698	**0.724**	0.592
F1 score	**0.0479**	0.0476	0.0369
mean Precision@10	**0.059**	0.045	0.039
Precision	0.34	**0.035**	0.023
Recall	**0.079**	0.073	0.103
MCC	**0.043**	0.042	0.035
AUC-PR	**0.018**	0.017	0.011
Accuracy	0.981	**0.982**	0.967
Negative predictive value	**0.994**	**0.994**	**0.994**
TP	202	186	263
FP	5663	5063	11413
FN	2359	2375	2298
TN	403767	404367	398017
Computational cost	100	1	–

Table 1. Comparison of the predictive performance of XGBoost with cross validation, Random Forest without cross validation and the RCA benchmark for the activations using different indicators. The last row indicates the computational cost with respect to the non cross validated Random Forest; XGBoost is about 100 times slower. The highest values of each indicator are in bold.

build a calibration curve. In Fig. 5 we show the fraction of positive elements as a function of the output (i.e., the scores) of the XGBoost and Random Forest algorithms in the activations prediction task. We divide the scores into logarithmic bins and then we compute the mean and the standard deviation inside each bin. In both cases a clear correlation is present, pointing out that a higher prediction score corresponds to a higher empirical probability that the activation of a new product will actually occur. Moreover, we note that the greater is the score produced by the model, the greater is the error on the y axis; the reason is that the models tend to assign higher scores to the products already exported from a country, so if we look at the activations the values start to fluctuate, and the statistic becomes lower.

We close this section mentioning the possibility to train our algorithms by taking explicitly into account the class imbalance, as suggested in[62,63]. The results of this investigation are reported in section 2 of the Supplementary Information. We observe a mild decrease of the prediction performance.

Opening the black box. In order to qualitatively motivate the better performance of tree-based algorithms, in this paragraph we elaborate on the operation of Random Forests. As specified in the "Methods" section, in these prediction exercises we train one Random Forest model for each product, and each Random Forest contains 100 decision trees. In Fig. 6 we show one representative decision tree. This tree is obtained by putting the number of features available for each tree equal to $P = 5040$: this means that we are bootstrap aggregating, or *bagging*[64] the trees, instead of building an actual Random Forest, which considers instead a random subset of the products[51] (the decision trees may be different also in this case, since the bagging procedure extracts the features with replacement). Moreover, the training procedure is cross validated, so the number of input countries is 156×7 (156 countries and 7 years from 2007 to 2013).

The decision tree we show refers to the product with HS1992 code 854089; the description is *valves and tubes not elsewhere classified in heading no. 8540*, where 8540 stands for *cold cathode or photo-cathode valves and tubes like vacuum tubes, cathode-ray tubes and similars*.

The color represents the class imbalance of the leaf (dark orange, many zeros; dark blue, many ones, quantified in the square brackets). The root product, the one which provides the best split, is *chromium*, which is used,

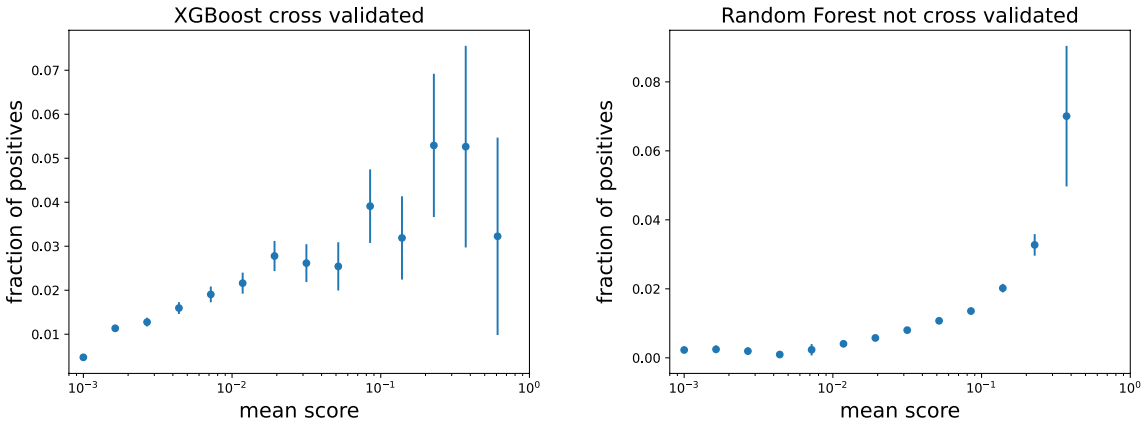

Figure 5. Calibration curves: fraction of positive elements as a function of the scores produced by XGBoost (left) and Random Forest (right) for the activations prediction task. In both cases a clear positive correlation is present, indicating that higher scores are associated to higher empirical probabilities that the activation will actually occur.

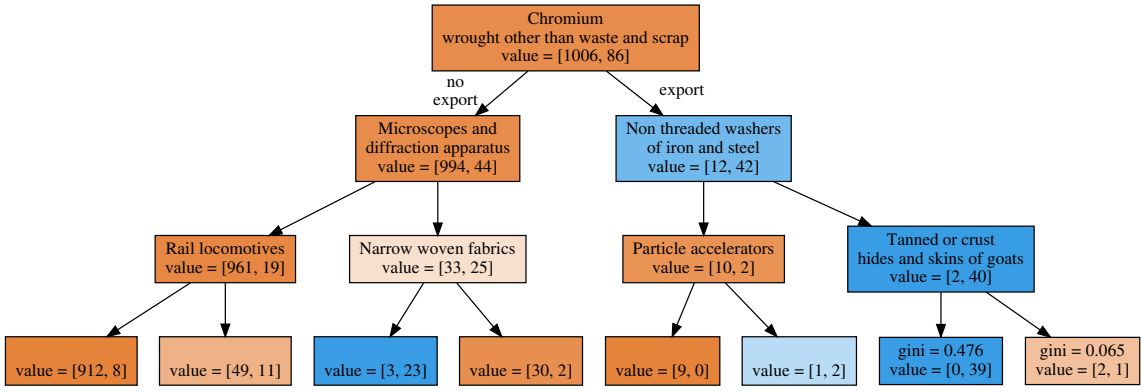

Figure 6. A representative decision tree to forecast the export of the product *valves and tubes*. The root product, *chromium*, has a well known technological relation with the target product, and in fact is able to discriminate against future exporters with high precision.

for instance, in the cathode-ray tubes to reduce X-ray leaks. So the Random Forest found a nontrivial connection between chromium and these types of valves and tubes: out of the 1006 couples country-year that do not export valves and tubes, 994 do not export chromium either (note the negative association). We can explore the network considering that the no-export link is always on the left. Looking at the export direction we find the cut on washers of iron and steel that works very well: only 2 of the 12 couples country-year that do not export valves and tubes do export washers and only 2 of the 42 countries that export valves and tubes do not export washers.

Looking at the other splits we find some of them more reasonable, like the one on particle accelerators, and some that seem coincidental, like the one on hides and skins of goats.

From this example it is clear that the decision tree is a natural framework to deal with a set of data in which some features (i.e., products) may be by far more informative than others, and so a hierarchical structure is needed to take into account this heterogeneous feature importance.

Feature importance may be evaluated by looking at the normalized average reduction of the impurity at each split that involves that feature[50]. In our case, we are considering the Gini impurity. In Fig. 7 we plot this assessment of the feature importance to predict the activation of valves and tubes. One can easily see that the average over the different decision trees is even more meaningful than the single decision tree shown before, even if the each one of the former sees fewer products than the latter: all the top products are reasonably connected with the target product and so it is natural to expect them to be key elements to decide whether the given country will export valves and tubes or not.

Time dependence. In the procedure discussed above we used a time interval Δ_{model} equal to 5 years for the training, and we tested our out-of-sample forecasts using the same time interval Δ. Here we investigate how the choice of the forecast horizon Δ affects the quality of the predictions. To make this analysis we used XGBoost models trained with the cross validation method and a lower $\Delta_{model} = 3$. The machine learning algorithms are trained using data in the range $y \in [1996 \ldots 2008]$ and their output, obtained giving $\mathbf{RCA}^{(2008)}$ as input, will be compared with the various $\mathbf{M}^{(2008+\Delta)}$ by varying Δ. Being the 2018 the last year of available data, we can explore a range of Δs from 1 to 10. All details about the training procedure of the machine learning algorithms are given in the "Methods" section.

The quality of the predictions as a function of the forecast horizon Δ are summarized in Fig. 8, where we normalized the indicators in such a way that they are all equal to 1 at $\Delta = 1$. In the left figure we have the plot for the activations prediction task: both *precision* and *precision@10* increase with Δ, while the *negative predictive value* decreases and accuracy shows an erratic behavior. This means that our ability to guess positive values improves or, in other words, the greater the time you wait the higher the probability that a country sooner or later does activate the products we predict. This improvement on positive values, however, corresponds to a worsening on negative values that can be interpreted as the fact that countries during time develop new capabilities and start to export products we cannot predict with a Δ interval too large.

If we look to a score that includes both performances on positive values and performance on negative values, like accuracy, we have a (noisy) worsening with the increase of Δ.

In the figure on the right we show instead the full matrix prediction task. In this case all the scores decrease with Δ because the algorithm can not leverage anymore on the strong auto-correlation of the RCA matrix.

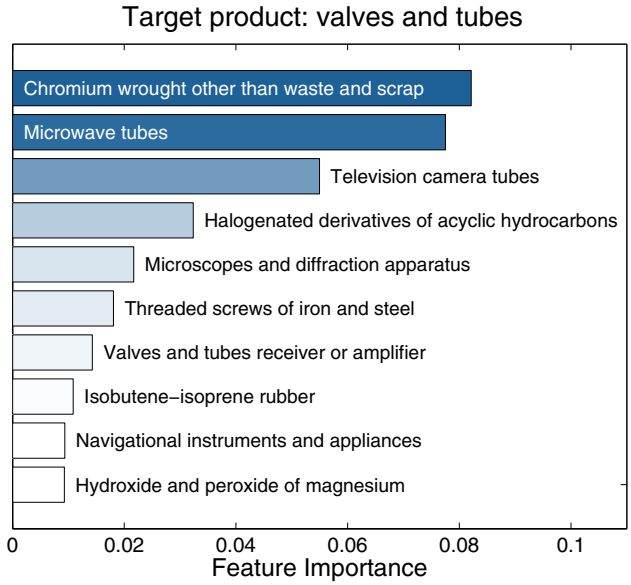

Figure 7. Feature importance is a measure of how much a product is useful to predict the activation of the target product. Here we use the average reduction of the Gini impurity at each split. All important products are reasonably connected with the target.

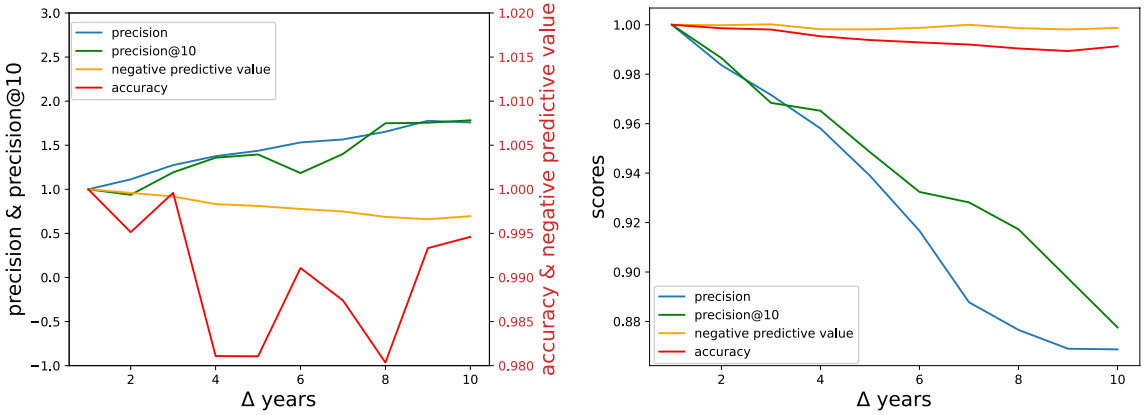

Figure 8. In the plot on the left we show the performance indicators in the case of the activations prediction task. The performance on positive values improves, while the one on negative values gets worse. On the right we show the same performance indicators in the case of the full matrix prediction task. All the scores get worse due to the vanishing auto-correlation of the matrices.

Note that the steepness of the decreasing curves is higher when we look at precision scores, the reason being the high class imbalance and the large number of true negatives with respect to true positives as shown in Table 1.

Discussion

One of the key issues in economic complexity and, more in general, in complexity science is the lack of systematic procedures to test, validate, and falsify theoretical models and empirical, data-driven methodologies. In this paper we focus on export data, and in particular on the country-product bipartite network, which is the basis of most literature on economic complexity, and the likewise widespread concept of *relatedness*, that is usually associated to an assessment of the proximity between two products or the density or closeness of a country with respect to a target product. As detailed in the Introduction, many competing approaches exist to quantify these concepts, however, a systematic framework to evaluate which approach works better is lacking, and the result is the flourishing of different methodologies, each one tested in a different way and with different purposes. We believe that this situation can be discussed in a quantitative and scientifically sound way by defining a concrete framework to compare the different approaches in a systematic way; the framework we propose is out-of-sample forecast, and in particular the prediction of the presence or the appearance of products in the future export baskets of countries. This approach has the immediate advantage to avoid a number of recognized issues[65] such as the mathiness of microfounded models[66] and the p-hacking in causal inference and regression analyses[1,67].

In this paper we systematically compare different machine learning algorithms in the framework of a supervised classification task. We find that the statistical properties of the export data, namely the strong auto-correlation and the class imbalance, imply that the appearance, or activation, of new products should be investigated, and some indicators of performance, such as ROC–AUC and accuracy, should be considered with extreme care. On the contrary, indicators such as the mean Precision@k have an immediate policy interpretation. In the prediction tasks tree-based models, such as Random Forest and Boosted Trees, clearly outperform the other algorithms and the quite strong benchmark provided by the simple RCA measure. The prediction performance of Boosted Trees can be further improved by training them in a cross validation setting, at the cost of a higher computational effort. The calibration curves, which show a high positive correlation between the machine learning scores and the actual probability of the activation of a new product, provide further support to the correctness of these approaches. A first step towards opening this black box is provided by the visual inspection of a sample decision tree and the feature importance analysis, which shows that the hierarchical organization of the decision tree is a key element to provide correct predictions but also insights about which products are more useful in this forecasting task.

From a theoretical perspective, this exercise points out the relevance of context for the appearance of new products, in the spirit of the New Structural Economics[68], but it has also immediate policy implications: each country comes with its own endowments and should follow a personalized path, and machine learning approaches are able to efficiently extract this information. In particular, the output of the Random Forest or the Boosted Trees algorithm, provides scores, or *progression probabilities,* that a product will be soon activated by the given country. This represents a quantitative and scientifically tested measure of the *feasibility* of a product in a country. This measure can be used in very practical contexts of investment design and industrial planning, a key issue after the covid-related economic crisis[69,70].

Conclusion

Measuring the relatedness between countries and products is one of the main topics in the economic complexity literature[71], given its importance to assess the feasibility of investments and strategic policies[72,73]. Starting from 2007 with the Product Space[26], many different approaches to measure the relatedness have been proposed[11,32,35–37,39,43]. With all these models in the literature, a big issue is the absence of a scientifically sound procedure to compare them and quantifying how good they are in measuring the relatedness.

The first contribution of this work is the proposal of out-of-sample forecasts of new exported products as a method to compare different relatedness models. In this way, the problem of measuring the relatedness can be casted as a binary classification exercise and, by using standard performance indicators, one can assess the goodness of a measure and compare them quantitatively. The second contribution of the present paper is the use of machine learning algorithms to measure the relatedness. We show that decision trees-based algorithms like Random Forest[51] and XGBoost[48] provide the best assessment and represent the benchmark for possible new measures of relatedness.

This paper opens up a number of research lines in various directions. One critical issue of the machine learning algorithms with respect to traditional network-based approaches is the explainability ot the results, so an important direction of research is the construction of a model that is fully explainable and do not lose quality with respect to the measures provided by machine learning algorithms. Another possible direction for future research is the application of this framework to different bipartite networks using different databases. Finally, one could use statistically validated projections[31] to build density-based predictions and compare them within our testing framework. All these studies will be presented in future works.

Methods

Data description. The data we use in this analysis are obtained from the UN-COMTRADE database, Harmonized System 1992 classification (HS 1992) and include the volumes of the export flows between countries. The raw database, however, presents internal inconsistencies: for instance, the import declaration of the buying country might not coincide with the corresponding export declaration of the selling country. The correct exchanged volumes may be inferred using a Bayesian approach[10]. The data used in this work are obtained from this cleaning procedure. The time range covered is 1996–2018 and for each year we have a matrix $\mathbf{V}$ whose element V_{cp} is the amount, expressed in US dollars, of product p exported by country c. The total number of countries is 169 and the total number of products is 5040.

To binarize the data we determine if a country competitively exports a product by computing the Revealed Comparative Advantage (RCA) introduced by Balassa[38]. The RCA of a country c in product p in year y is given by:

$$R_{cp}^{(y)} = \frac{V_{cp}^{(y)}}{\sum_{p'} V_{cp'}^{(y)}} \Big/ \frac{\sum_{c'} V_{c'p}^{(y)}}{\sum_{c'p'} V_{c'p'}^{(y)}} \tag{1}$$

$R_{cp}^{(y)}$ is a continuous value and represents the ratio between the weight of product p in the export basket of country c and the total weight of that product in the international trade. Alternatively, the RCA can be seen as the ratio between the market share of country c relatively to product p and the weight of country c with respect to the total international trade. This is the standard way, in the economic complexity literature, to remove trivial effects due to the size of the country and the size of the total market of the product. In this way, a natural threshold equal to 1 can be used to establish whether country c exports product p in a competitive way or not. As a consequence, we define the matrix $\mathbf{M}$ whose binary element M_{cp} tells us if country c is competitive in the export of product p or not:

$$M_{cp}^{(y)} = \begin{cases} 1 & if \quad R_{cp}^{(y)} \geq 1 \\ 0 & if \quad R_{cp}^{(y)} < 1 \end{cases} \tag{2}$$

In this work we will try to predict future values of M_{cp} using past values of RCA. In Table 2 we report the main features of the country-export bipartite network described by the biadjacency matrix $\boldsymbol{M}$ (in different years). The minimum country degree is zero from 1996 to 2011 due to South Sudan since it gained its independence on 2011. The minimum degree of the products is always zero because there are some products in which on some years none of the countries has a RCA value greater than 1.

A detailed description of the dataset we used is available at[74].

Supervised machine learning and relatedness. Before describing our approach to measure the relatedness, here we want to give a quick and intuitive description of how supervised machine learning works. A simple example consists in the construction of a binary classifier that predicts if a patient is healthy or it has contracted COVID-19 starting from its symptoms (called features). A simple approach consists into drawing an hyperspace with dimension equal to the number of features (N). Here a patient identifies a specific point in this space. A binary classifier could be a simple hyperplane with dimension N−1 splitting the space in two distinct areas. A patient is then classified as healthy or sick depending on which of the two areas he belongs to. The learning part consists in the definition of the hyperplane. During the training phase we provide to the model some patients with their symptoms and the information whether they contracted COVID-19 or not. By minimizing a suitable loss function the model finds the optimal hyperplane that separates the healthy from the sick.

This is a very simple example of the functioning of a supervised machine learning binary classifier (that usually does not perform well, except in trivial cases where the positive and negative classes can be linearly separated). The functioning of more complex architectures like the ones we present in this paper is not so different: what we have is always a classifier that learns its task looking at a set of training samples and their correct output. In our case, we first fix a target product. Thus a sample is a country and its exported products are the features. Looking to past data we show to the algorithm if a country after 5 years will export the target product, and, once the training phase is ended, the algorithm can be used to predict whether a country will export that product after 5 years or not given its present exports. Then this procedure is repeated for all products, each of which thus

Year	Number of countries	Number of products	Number of links	Min country degree	Max country degree	Avg country degree	Min product degree	Max product degree	Avg product degree
1996	169	5040	83,754	0	2082	496	0	64	16.6
1997	169	5040	83,666	0	2059	495	0	61	16.6
1998	169	5040	84,976	0	2023	503	0	64	16.9
1999	169	5040	86,071	0	2089	509	0	66	17.1
2000	169	5040	90,327	0	2171	534	0	67	17.9
2001	169	5040	89,242	0	2138	528	0	71	17.7
2002	169	5040	88,849	0	2114	526	0	73	17.7
2003	169	5040	88,153	0	2089	522	0	73	17.5
2004	169	5040	88,662	0	2148	525	0	69	17.6
2005	169	5040	90,807	0	2171	537	0	74	18.0
2006	169	5040	90,429	0	2162	535	0	69	17.9
2007	169	5040	90,152	0	2155	533	0	72	17.9
2008	169	5040	90,505	0	2230	536	0	69	18.0
2009	169	5040	89,388	0	2157	529	0	72	17.7
2010	169	5040	88,742	0	2195	525	0	71	17.6
2011	169	5040	87,801	0	2286	520	0	68	17.4
2012	169	5040	88,368	8	2253	523	0	73	17.5
2013	169	5040	87,482	5	2222	518	0	79	17.4
2014	169	5040	85,724	7	2236	507	0	80	17.0
2015	169	5040	83,151	10	2236	492	0	81	16.5
2016	169	5040	83,012	11	2260	491	0	78	16.5
2017	169	5040	82,992	13	2202	491	0	81	16.5
2018	169	5040	81,059	12	2256	480	0	91	16.0

Table 2. Main properties of the country-export bipartite network over the years between 1996 and 2018.

needs a different training. In Fig. 9 we show a schematic diagram with the general functioning of the machine learning algorithms discussed here. As a first step, the algorithm is trained receiving the matrix of the RCAs of countries (X_{train}) and the information whether these countries will export a product or not (Y_{train}). Once the algorithm is trained, it receives in input the exports of countries in a year y (not used during the training stage) and its output is the relatedness of countries with a product.

Training and testing procedure. We want to guess which products will be exported by a country after Δ years. To do this, we exploit machine learning algorithms with the goal of (implicitly) understanding the capabilities needed to export a product from the analysis of the export basket of countries. Since each product requires a different set of capabilities, we need to train different models: in this work, we train 5040 different Random Forests, one for each product.

The training procedure is analogous for all the models: they have to connect the RCA values of the products exported by a country in year y with the element $M_{cp}^{(y+\Delta)}$, which tells us if country c in year $y + \Delta$ is competitive in the export of product p.

In the general case we have export data that covers a range of years $[y_0, y_{last}]$. The last year is used for the test of the model and so the training set is built using only the years $[y_0, y_{last} - \Delta]$. In this way, no information about the Δ years preceding y_{last} is given.

The input of the training set, that we call $\mathbf{X}_{train}$, is vertical stack of the $\mathbf{R}^{(y)}$ matrices from y_0 to $y_{last} - 2\Delta$ (see Fig. 10). In such a way we can consider all countries and all years of the training set, and these export baskets will be compared with the corresponding presence or absence of the target product p after Δ years; this is because our machine learning procedure is supervised, that is, during the training we provide a set of answers $\mathbf{Y}_{train}$ corresponding to each export basket in $\mathbf{X}_{train}$. While $\mathbf{X}_{train}$ is the same for all the models (even if they refer to different products), the output of the training set $\mathbf{Y}_{train}$ changes on the basis of the product we want to predict. If we consider the model associated to product p, to build $\mathbf{Y}_{train}$ we aggregate the columns corresponding to the target product, $C_p^{(y)}$, of the $\mathbf{M}$ matrices from $y_0 + \Delta$ to $y_{last} - \Delta$ (so we use the same number of years, all shifted by Δ years with respect to $\mathbf{X}_{train}$). This is graphically represented on the extreme left side of Fig. 10.

Once the model is trained, in order to perform the test we give as input $\mathbf{X}_{test}$ the matrix $\mathbf{R}^{(y_{last}-\Delta)}$. Each model will give us its prediction for the column p of the matrix $\mathbf{M}^{(y_{last})}$ and, putting all the results relative to the single products together, we reconstruct the whole matrix of scores $\mathbf{M}_{pred}^{(y_{last})}$, which we compare with the empirical one. There are various ways to compare the predictions with the actual outcomes, and these performance metrics are discussed in the following section.

As already mentioned, the same models can be tested against two different prediction tasks: either we can look to the full matrix $\mathbf{M}^{(y_{last})}$, either we can concentrate only on the possible *activations*, that is products that were not present in an export basket and countries possibly start exporting. The set of possible activations is defined as follows:

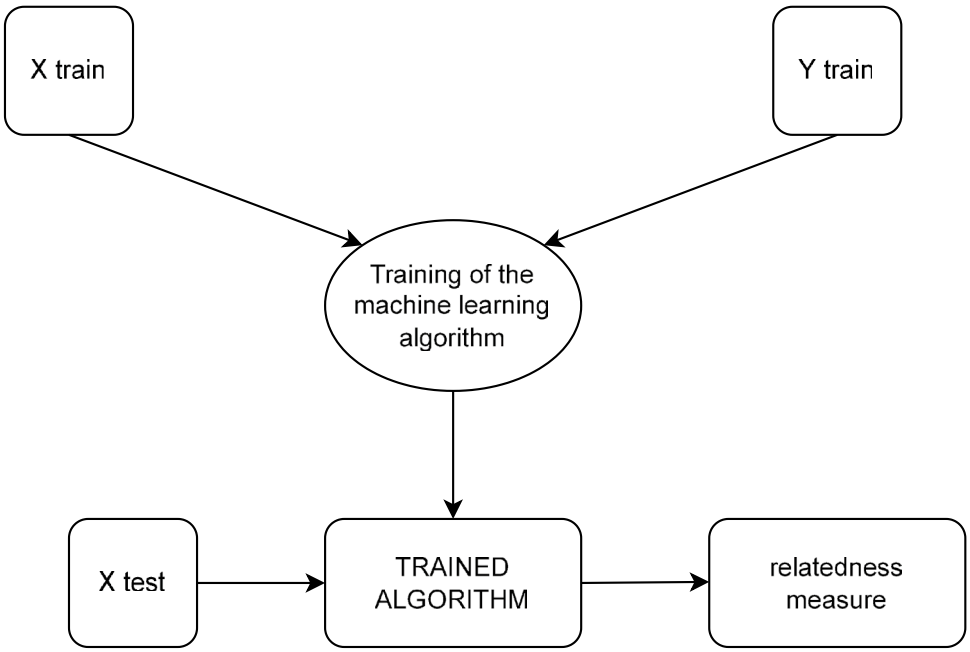

Figure 9. Schematic diagram with the functioning of machine learning algorithms to assess the relatedness between countries and a target product. During the training phase the model receives an X_{train} matrix with the training samples (countries) and their features (products) for the years from 1996 to 2008; they are compared with the Y_{train} vector that contains the corresponding possible exports the target product in 2001–2013 (that is, a binary label for each sample). Once the model is trained, it can receive in input new data (that is, an export basket) and will provide the probability that the label (the possible export of the target product)) is 1. This progression score is our assessment of the relatedness.

$$(c,p) \in activations \iff R_{cp}^{(y)} < 0.25 \quad \forall\, y \in [y_0, y_{last} - \Delta] \tag{3}$$

In other words, a pair (c,p) is a possible activation if country c has never been competitive in the export of product p until year $y_{last} - \Delta$, that is its RCA values never exceeded 0.25. This selection of the test set may look too strict, however it is key to test our algorithms against situations in which countries really start exporting new products. Because of the RCA binarization, there are numerous cases in which a country noisily oscillates around RCA = 1 and, de facto, that country is already competitive in that product; in these cases the RCA benchmark is more than enough for a correct prediction.

The way to train the models we just described performs better on the full matrix than in the activations. The reason is probably that the machine learning algorithms recognize the countries because the ones in the training set and the ones in the test set are the same. When the algorithms receive as input the export basket of a country they have already seen in the training data, they tend to reproduce the strong autocorrelation of the export matrices. To avoid this problem we used a k-fold cross validation, which means that we split the countries into k groups. Since the number of countries is 169, the natural choice is to use k = 13, so we randomly extract a group α of 13 countries from the training set, which is then composed by the remaining 156 countries, and we use only the countries contained in α for the test. In this way each model is meant to make predictions only on the countries of the group α, so to cover all the 169 countries we need to repeat the procedure 13 times, every time changing the countries in the group α. This different training procedure is depicted on the right part of Fig. 10. So there will be 13 models associated to a single product and, for this reason, the time required to make the training is 13 times longer. Like in the previous case, in the training set we aggregate the years in the range [y_0, $y_{last} - \Delta$]. $\mathbf{X}_{train}$ is the aggregation of the RCA matrices from y_0 to $y_{last} - 2\Delta$ and $\mathbf{Y}_{train}$ is the aggregation of the column p of the M matrices from $y_0 + \Delta$ to $y_{last} - \Delta$. In both cases, the countries in the group α are removed.

When we perform the test, each models takes as $\mathbf{X}_{test}$ the matrix $\mathbf{RCA}^{(y_{last}-\Delta)}$ with only the rows corresponding to the 13 countries in group α and gives as output scores the elements of the matrix $\mathbf{M}_{pred}^{(y_{last})}$. All the 5040×13 models together give as output the whole matrix of scores $\mathbf{M}_{pred}^{(y_{last})}$ that will be compared to the actual $\mathbf{Y}_{test} = \mathbf{M}^{(y_{last})}$.

Since the output of the machine learning algorithms is a probability, and most of the performance indicators require a binary prediction, in order to establish if we predict a value of 0 or 1 we have to introduce a threshold. The value of this threshold we use is the one that maximizes the F1-score. We note that the only performance measures that do not require a threshold are the ones that consider areas under the curves, since these curves are built precisely by varying the threshold value.

Figure 10 schematically shows the training procedures with and without cross validation.

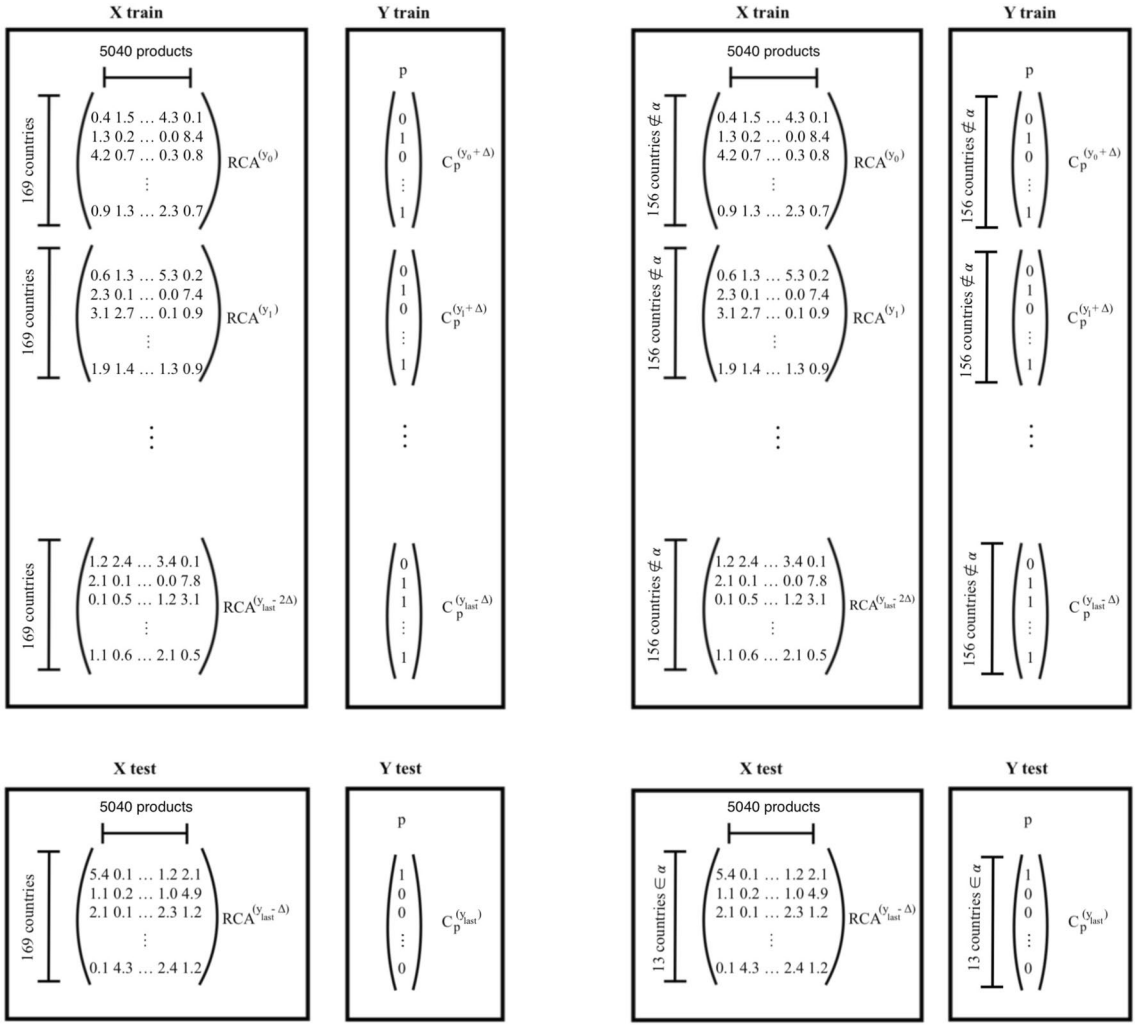

Figure 10. The training and testing procedure with (right) and without (left) cross validation. See the text for a detailed explanation.

Performance indicators. The choice of the performance indicators is a key issue of supervised learning[61,75] and, in general, strongly depends on the specific problem under investigation. Here we discuss the practical meaning of the performance indicators we used to compare the ML algorithms. For all the scores but the areas under curves, we need to define a threshold above which the output scores of the ML algorithms are associated with a positive prediction. For this purpose we choose the threshold that maximizes the F1 score[76].

- *Precision* Precision is defined as the ratio between true positives and positives[61]. In our case, we predict that a number of products will be competitively exported by some countries; these are the *positives*. The precision is the fraction that counts how many of these predicted products are actually exported by the respective countries after Δ years. A high value of precision is associated to a low number of false positives, that is if products that are predicted to appear they usually do so.
- *mean Precision@k (mP@k)* This indicator usually corresponds to the fraction of the top *k* positives that are correctly predicted. We considered only the first *k* predicted products *separately for each country*, and then we average on the countries. This is of practical relevance from a policy perspective, because many new products appear in already highly diversified countries, while we would like to be precise also in low and medium income countries. By using mP@k we quantify the correctness of our possible recommendations of k products, on average, for a country.
- *Recall* Recall is defined as the ratio between true positives and the sum of true positives and false negatives or, in other words, the total number of products that a country will export after Δ years[61]. So a high recall is associated with a low number of false negatives, that is, if we predict that a country will not start exporting a product, that country will usually not export that product. A negative recommendation is somehow less usual in strategic policy choices.
- *F1 Score* The F1 score or F-measure[59,60] is defined as the harmonic mean of precision and recall. As such, it is possible to obtain a high value of F1 only if both precision and recall are relatively high, so it is a very frequent

choice to assess the general behavior of the classificator. As mentioned before, both precision and recall can be trivially varied by changing the scores' binarization threshold; however, the threshold that maximizes the F1 score is far from trivial, since precision and recall quantify different properties and are linked here in a nonlinear way. The *Best F1 Score* is computed by finding the threshold that maximizes the F1 score.

- *Area under the PR curve* It is possible to build a curve in the plane defined by precision and recall by varying the threshold that identifies the value above which the scores are associated to positive predictions. This value is not misled by the class imbalance[46].
- *ROC–AUC* The Area Under the Receiving Operating Characteristic Curve[77,78] is a widespread indicator that aims at measuring the *overall* predictive power, in the sense that the user does not need to specify a threshold, like for Precision and Recall. On the contrary, all the scores are considered and ranked, and for each possible threshold both the True and the False Positive Rate (TPR and FPR, respectively) are computed. This procedure allows to define a curve in the TPR/FPR plane, and the area under this curve represents the probability that a randomly selected positive instance will receive a higher score than a randomly selected negative instance[45]. For a random classifier, AUC = 0.5 . It is well known[46,79] that in the case of highly imbalanced data the AUC may give too optimistic results. This is essentially due to its focus on the overall ranking of the scores: in our case, misordering even a large number of not exported products does not affect the prediction performance; one makes correct true negative predictions only because there are a lot of negative predictions to make.
- *Matthews coefficient* Matthews' correlation coefficient[80] takes into account all the four classes of the confusion matrix and the class imbalance issue[81,82].
- *Accuracy* Accuracy is the ratio between correct predictions (true positives and true negatives) and the total number of predictions (true positives, false positives, false negatives and true negatives)[61]. In our prediction exercise we find relatively high values of accuracy essentially because of the overwhelming number of (trivially) true negatives (see Table 1).
- *Negative predictive value* Negative predictive value is defined as the ratio between true negatives and negatives, that are the products we predict will not be exported by a country[61]. Also in this case, a major role is played by the very large number of true negatives, that are however less significant from a policy perspective.

Libraries for the ML models. Most of the models are implemented with scikit-learn 0.24.0 and, as described in the Supplementary Information, we performed a carefully hyperparameter optimization; in particular we used (the unspecified hyperparameters values are the default ones):

- sklearn.ensemble.RandomForestClassifier(n_estimators = 100, min_samples_leaf = 7)
- sklearn.svm.SVC(kernel = "rbf")
- sklearn.linear_model.LogisticRegression(solver = "newton-cg")
- sklearn.tree.DecisionTreeClassifier()
- sklearn.tree.ExtraTreesClassifier(n_estimators = 100, min_samples_leaf = 8)
- sklearn.ensemble.AdaBoostClassifier(n_estimators = 3)
- sklearn.naive_bayes.GaussianNB()
- xgboost.XGBClassifier(n_estimators = 15, min_child_weight = 45, reg_lambda = 1.5)

XGBoost is implemented using the library xgboost 1.3.1.

Finally, the neural network is implemented using keras 2.4.3. It consists on two layers with 64 neurons and activation function RELU and a final layer with a single neuron and sigmoid activation. We used rmsprop as optimizer, binary_crossentropy as loss function, accuracy as loss metric and we stopped the training at 10 epochs.

For a detailed explanation about the choice of the hyperparameters the reader is referred to the supplementary information. Note that in our case tree-based models perform better and it is known in the literature that the random forest default values already provide very good results[79,83,84]. In our case, the hyperparameters optimization increased our prediction performances of about 10%; in particular, it decreased the number of false positives.

Comparison with other works. Here we compare our Random Forest model with the other approaches presented in literature that we cited in the introduction section, using a consistent testing framework (4-digits classification, comparison between the relatedness computed in 2013 and the actual new exported products in 2018 that had RCA<0.25 from 1996 to 2013).

- Hidalgo et al. in 2007 define the Product Space[26] that is still widely used to measure relatedness[37]. It is a projection of the country-product bipartite network into the layer of the products (thus defining a proximity network of the products). The relatedness between a country and a product is defined as the density of the former around the latter in the Product Space;
- O'Clery et al. in 2021 introduce a new approach to define the proximity network of the products called EcoSpace[32]. From this network they define the Ecosystem density—that is the likelihood of the appearance of a product in a country—as a relatedness measure;
- Medo et al. compare different approaches to perform a link prediction on bipartite nested networks finding that the two most performing techniques are the Number of violations of the nestedness property (NViol)[85] and the preferential attachment (prefA), where the relatedness is the product of the diversification of the country with the ubiquity of the product[36] .

Algorithm	AUC-PR	Best F1	AUC-ROC	mean Precision@5
Random Forest	**0.015**	**0.042**	**0.689**	**0.049**
Product Space	0.010	0.022	0.637	0.032
EcoSpace	0.013	0.035	0.663	0.042
prefA	0.011	0.024	0.645	0.046
NViol	0.011	0.025	0.607	0.046

Table 3. Comparison between our Random Forest model and other approaches proposed in literature. The Random Forest provides a better assessment of the relatedness with all the performance indicators. The highest values of each indicator are in bold.

In Table 3 we show the AUC-PR, AUC-ROC, Best F1 and mean precision@5 of the different models. We find that the Random Forest outperforms the other approaches independently from the specific performance metric used in the comparison.

Data availibility

The data that support the findings of this study are available from UN-COMTRADE but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request to the corresponding author and with permission of UN-COMTRADE. An anonymized and processed version of the data is available at https://github.com/giamba95/SaplingSimilarity/tree/main/data/RCA to permit the full replicability of our study.

Received: 21 June 2022; Accepted: 13 January 2023
Published online: 27 January 2023

References

1. Athey, S. The impact of machine learning on economics. in *The Economics of Artificial Intelligence: An Agenda*. 507–547 (University of Chicago Press, 2018).
2. Rodrik, D. Diagnostics before prescription. *J. Econ. Perspect.* **24**, 33–44 (2010).
3. Hausmann, R., Rodrik, D. & Velasco, A. Growth diagnostics. in *The Washington Consensus Reconsidered: Towards a New Global Governance*. 324–355 (2008).
4. Baldovin, M., Cecconi, F., Cencini, M., Puglisi, A. & Vulpiani, A. The role of data in model building and prediction: A survey through examples. *Entropy* **20**, 807 (2018).
5. Hosni, H. & Vulpiani, A. Forecasting in light of big data. *Philos. Technol.* **31**, 557–569 (2018).
6. Rodrik, D. *Economics Rules: The Rights and Wrongs of the Dismal Science* (WW Norton & Company, 2015).
7. Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A. & Pietronero, L. A new metrics for countries' fitness and products' complexity. *Sci. Rep.* **2**, 723 (2012).
8. Cristelli, M., Gabrielli, A., Tacchella, A., Caldarelli, G. & Pietronero, L. Measuring the intangibles: A metrics for the economic complexity of countries and products. *PloS one* **8**, e70726 (2013).
9. Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A. & Pietronero, L. Economic complexity: Conceptual grounding of a new metrics for global competitiveness. *J. Econ. Dyn. Control* **37**, 1683–1691 (2013).
10. Tacchella, A., Mazzilli, D. & Pietronero, L. A dynamical systems approach to gross domestic product forecasting. *Nat. Phys.* **14**, 861–865 (2018).
11. Zaccaria, A., Cristelli, M., Tacchella, A. & Pietronero, L. How the taxonomy of products drives the economic development of countries. *PloS one* **9**, e113770 (2014).
12. Zaccaria, A., Cristelli, M., Kupers, R., Tacchella, A. & Pietronero, L. A case study for a new metrics for economic complexity: The Netherlands. *J. Econ. Interact. Coord.* **11**, 151–169 (2016).
13. Gaulier, G. & Zignago, S. Baci: International trade database at the product-level (the 1994–2007 version). in*CEPII Working Paper 2010–2023* (2010).
14. Hidalgo, C. A. & Hausmann, R. The building blocks of economic complexity. *Proc. Natl. Acad. Sci.* **106**, 10570–10575 (2009).
15. Albeaik, S., Kaltenberg, M., Alsaleh, M. & Hidalgo, C. Improving the Economic Complexity Index. arXiv preprint arXiv:1707.05826 (2017).
16. Gabrielli, A. *et al.* Why we like the eci+ algorithm. arXiv preprint arXiv:1708.01161 (2017).
17. Albeaik, S., Kaltenberg, M., Alsaleh, M. & Hidalgo, C. 729 new measures of economic complexity (addendum to improving the economic complexity index). arXiv preprint arXiv:1708.04107 (2017).
18. Pietronero, L. *et al.* Economic complexity:" Buttarla in caciara" vs a constructive approach. arXiv preprint arXiv:1709.05272 (2017).
19. Cristelli, M., Tacchella, A. & Pietronero, L. The heterogeneous dynamics of economic complexity. *PloS one* **10**, e0117174 (2015).
20. Cristelli, M., Tacchella, A., Cader, M., Roster, K. & Pietronero, L. *On the Predictability of Growth* (The World Bank, 2017).
21. Liao, H. & Vidmer, A. A comparative analysis of the predictive abilities of economic complexity metrics using international trade network. *Complexity* (2018).
22. Sciarra, C., Chiarotti, G., Ridolfi, L. & Laio, F. Reconciling contrasting views on economic complexity. *Nat. Commun.* **11**, 1–10 (2020).
23. Frenken, K., Van Oort, F. & Verburg, T. Related variety, unrelated variety and regional economic growth. *Region. Stud.* **41**, 685–697 (2007).
24. Hidalgo, C. A. *et al.* The principle of relatedness. in *International Conference on Complex Systems*. 451–457 (Springer, 2018).
25. Teece, D. J., Rumelt, R., Dosi, G. & Winter, S. Understanding corporate coherence: Theory and evidence. *J. Econ. Behav. Organ.* **23**, 1–30 (1994).
26. Hidalgo, C. A., Klinger, B., Barabási, A.-L. & Hausmann, R. The product space conditions the development of nations. *Science* **317**, 482–487 (2007).
27. Breschi, S., Lissoni, F. & Malerba, F. Knowledge-relatedness in firm technological diversification. *Res. Policy* **32**, 69–87 (2003).

28. Pugliese, E., Napolitano, L., Zaccaria, A. & Pietronero, L. Coherent diversification in corporate technological portfolios. *PloS one* **14** (2019).
29. Neffke, F., Henning, M. & Boschma, R. How do regions diversify over time? Industry relatedness and the development of new growth paths in regions. *Econ. Geogr.* **87**, 237–265 (2011).
30. Boschma, R. *et al.* Technological relatedness and regional branching. in *Beyond Territory. Dynamic Geographies of Knowledge Creation, Diffusion and Innovation*. 64–68 (2012).
31. Pugliese, E. *et al.* Unfolding the innovation system for the development of countries: Coevolution of science, technology and production. *Sci. Rep.* **9**, 1–12 (2019).
32. O'Clery, N., Yıldırım, M. A. & Hausmann, R. Productive ecosystems and the arrow of development. *Nat. Commun.* **12**, 1–14 (2021).
33. Gnecco, G., Nutarelli, F. & Riccaboni, M. A machine learning approach to economic complexity based on matrix completion. *Sci. Rep.* **12**, 1–10 (2022).
34. Hausmann, R., Hwang, J. & Rodrik, D. What you export matters. *J. Econ. Growth* **12**, 1–25 (2007).
35. Bustos, S., Gomez, C., Hausmann, R. & Hidalgo, C. A. The dynamics of nestedness predicts the evolution of industrial ecosystems. *PloS one* **7**, e49393 (2012).
36. Medo, M., Mariani, M. S. & Lü, L. Link prediction in bipartite nested networks. *Entropy* **20**, 777 (2018).
37. Zhang, W.-Y., Chen, B.-L., Kong, Y.-X., Shi, G.-Y. & Zhang, Y.-C. Industry upgrading: Recommendations of new products based on world trade network. *Entropy* **21**, 39 (2019).
38. Balassa, B. Trade liberalisation and "revealed" comparative advantage 1. *Manchester Sch.* **33**, 99–123 (1965).
39. Tacchella, A., Zaccaria, A., Miccheli, M. & Pietronero, L. Relatedness in the era of machine learning. arXiv preprint arXiv:2103.06017 (2021).
40. Hausmann, R. *et al. A roadmap for investment promotion and export diversification: The case of Jordan* (Technical Report. Center for International Development at Harvard University, 2019).
41. Saracco, F., Di Clemente, R., Gabrielli, A. & Pietronero, L. From innovation to diversification: A simple competitive model. *PloS one* **10**, e0140420 (2015).
42. Tacchella, A., Di Clemente, R., Gabrielli, A. & Pietronero, L. The build-up of diversity in complex ecosystems. arXiv preprint arXiv:1609.03617 (2016).
43. Che, N. X. *Intelligent export diversification: An export recommendation system with machine learning* (Technical Report. International Monetary Fund, 2020).
44. Angelini, O. & Di Matteo, T. Complexity of products: The effect of data regularisation. *Entropy* **20**, 814 (2018).
45. Fawcett, T. An introduction to roc analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).
46. Saito, T. & Rehmsmeier, M. The precision–recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one* **10**, e0118432 (2015).
47. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 1189–1232 (2001).
48. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. in *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*. 785–794 (2016).
49. Gulli, A. & Pal, S. *Deep Learning with Keras* (Packt Publishing Ltd, 2017).
50. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
51. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
52. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
53. Hosmer Jr, D.W., Lemeshow, S. & Sturdivant, R.X. *Applied Logistic Regression*. Vol. 398 (Wiley, 2013).
54. Quinlan, J. R. Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986).
55. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
56. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
57. John, G. H. & Langley, P. Estimating continuous distributions in Bayesian classifiers. arXiv preprint arXiv:1302.4964 (2013).
58. Shalev-Shwartz, S. & Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms* (Cambridge University Press, 2014).
59. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302 (1945).
60. Van Rijsbergen, C. J. Foundation of evaluation. *J. Docum.* (1974).
61. Powers, D. M. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *J. Mach. Learn. Technol.* (2011).
62. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
63. Géron, A. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (O'Reilly Media, Inc., 2019).
64. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
65. Romer, P. The trouble with macroeconomics. *Am. Econ.* (2016).
66. Romer, P. M. Mathiness in the theory of economic growth. *Am. Econ. Rev.* **105**, 89–93 (2015).
67. Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. & Jennions, M. D. The extent and consequences of p-hacking in science. *PLoS Biol.* **13**, e1002106 (2015).
68. Lin, J. Y. *New Structural Economics: A Framework for Rethinking Development and Policy* (The World Bank, 2012).
69. Fernandes, N. Economic effects of coronavirus outbreak (COVID-19) on the world economy. in *Available at SSRN 3557504* (2020).
70. Nana, I. & Starnes, S. *When trade falls-effects of covid-19 and outlook* (Technical Report. International Finance Corporation-World Bank Group, 2020).
71. Hidalgo, C. A. Economic complexity theory and applications. *Nat. Rev. Phys.* **3**, 92–113 (2021).
72. Lin, J., Cader, M. & Pietronero, L. What African industrial development can learn from east Asian successes. in *EM COmpass 88* (2020).
73. Pugliese, E. & Tacchella, A. *Economic complexity for competitiveness and innovation: A novel bottom-up strategy linking global and regional capacities* (Technical Report. Joint Research Centre (Seville site), 2020).
74. Patelli, A., Pietronero, L. & Zaccaria, A. Integrated database for economic complexity. *Sci. Data* **9**, 1–13 (2022).
75. Caruana, R. & Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. in *Proceedings of the 23rd International Conference on Machine Learning*. 161–168 (2006).
76. Lipton, Z. C., Elkan, C. & Naryanaswamy, B. Optimal thresholding of classifiers to maximize f1 measure. in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 225–239 (Springer, 2014).
77. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
78. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
79. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).

80. Matthews, B. W. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta (BBA)-Protein Struct.* **405**, 442–451 (1975).
81. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
82. Boughorbel, S., Jarray, F. & El-Anbari, M. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. *PloS one* **12**, e0177678 (2017).
83. Genuer, R., Poggi, J.-M. & Tuleau, C. Random forests: Some methodological insights. arXiv preprint arXiv:0811.3619 (2008).
84. Probst, P., Wright, M. N. & Boulesteix, A.-L. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip. Rev. Data Mining Knowl. Discov.* **9**, e1301 (2019).
85. Grimm, A. & Tessone, C. J. Analysing the sensitivity of nestedness detection methods. *Appl. Netw. Sci.* **2**, 1–19 (2017).

Author contributions

Conceptualization: A.Z., A.T.; Methodology: all; Investigation, Software: G.A.; Validation: A.Z., A.T., G.A.; Writing, Review and editing: all; Supervision: L.P.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-28179-x.

Correspondence and requests for materials should be addressed to A.Z.

Reprints and permissions information is available at www.nature.com/reprints.

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