Journal of Physics: Complexity

PAPER • OPEN ACCESS

Which will be your firm’s next technology? Comparison between machine learning and network-based algorithms

To cite this article: Matteo Straccamore *et al* 2022 *J. Phys. Complex.* **3** 035002

View the article online for updates and enhancements.

You may also like

- Network-based ranking in social systems: three challenges
  Manuel S Mariani and Linyuan Lü
- Model-free control of dynamical systems with deep reservoir computing
  Daniel Canaday, Andrew Pomerance and Daniel J Gauthier
- Topological analysis of traffic pace via persistent homology
  Daniel R Carmody and Richard B Sowers

This content was downloaded from IP address 130.192.10.10 on 06/09/2022 at 08:36

IOP Publishing J.Phys.Complex. **3** (2022) 035002 (17pp) https://doi.org/10.1088/2632-072X/ac768d

Journal of Physics: Complexity

OPEN ACCESS

RECEIVED
9 November 2021

REVISED
27 May 2022

ACCEPTED FOR PUBLICATION
7 June 2022

PUBLISHED
2 September 2022

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

PAPER

Which will be your firm's next technology? Comparison between machine learning and network-based algorithms

Matteo Straccamore[1,3,*], Luciano Pietronero[3] and Andrea Zaccaria[2,3]

[1] Dipartimento di Fisica Università 'Sapienza' P.le A. Moro, 2, 00185 Rome, Italy
[2] Istituto dei Sistemi Complessi (ISC) - CNR UoS Sapienza, P.le A. Moro, 2, 00185 Rome, Italy
[3] Centro Ricerche Enrico Fermi Piazza del Viminale, 1, 00184 Rome, Italy
[*] Author to whom any correspondence should be addressed.

E-mail: matteo.straccamore@cref.it

Keywords: economic complexity, technological innovation, predictions, patenting firms

Supplementary material for this article is available online

Abstract

We reconstruct the innovation dynamics of about two hundred thousand companies by following their patenting activity for about ten years. We define the technology portfolios of these companies as the set of the technological sectors present in the patents they submit. By assuming that companies move more frequently towards related sectors, we leverage their past activity to build network-based and machine learning algorithms to forecast the future submissions of patents in new sectors. We compare different prediction methodologies using suitable evaluation metrics, showing that tree-based machine learning algorithms outperform the standard methods based on networks of co-occurrences. This methodology can be applied by firms and policymakers to disentangle, given the present innovation activity, the feasible technological sectors from those that are out of reach.

1. Introduction

In this work, we quantify the relatedness between a firm and a technology sector in different ways, namely using standard methods based on co-occurrences networks and supervised machine learning algorithms (Tacchella *et al* (2021), Albora *et al* (2021)). In order to compare such assessments, we develop an out-of-sample prediction framework based on the assumption that, on average, the next technology sector in which a firm will patent will be among the ones that are more related to its present patenting portfolio. In this way, we can build and study the *technological adjacent possible* of innovative firms, this concept being originally introduced by Kauffman (1996) and subsequently mathematically formalized in Tria *et al* (2014) and Loreto *et al* (2016). We find that machine learning algorithms not only show better prediction performances but allow for a two-dimensional representation of technology sectors that we call Continuous Technology Space (CTS). The CTS can be used to visualize the patenting portfolio of companies and to design strategic investments and acquisitions.

The question regarding the nature of the link between the performance of firms and their internal allocation of resources (Penrose 1959) and capabilities (Teece *et al* 1994) has fueled the interest of economics and management scholars for a long time, since opening the black box of corporate strategy would be key to gain insight into the determinants of corporate heterogeneity and hence a better understanding of markets and their evolution. To the best of our knowledge, these analyses are all aimed at finding explanatory variables for the present performance and not at forecasting future activity. On the contrary, the approach known as Economic Fitness and Complexity (Tacchella *et al* (2012), Sbardella *et al* (2018)), widely applied at both country and regional level, naturally focuses on forecasting, which represent a natural and scientifically sound framework to validate and falsify the different approaches (Tacchella *et al* 2018, Albora *et al* 2021, Tacchella *et al* 2021).

© 2022 The Author(s). Published by IOP Publishing Ltd

The aim of the present paper is to apply the EFC forecasting methods at firm level, and in particular to the bipartite network of firms and the technology sectors in which they show patenting activity.

One of the main problems for the economic literature is to empirically track the capabilities and the strategic choices of companies. Unfortunately, these elements are generally intangible, so the empirical literature often struggles to find instruments to keep up with the theoretical richness of the debate. One of the more easily measurable footprints left behind by the strategic decision-making of firms is *diversification*, i.e. the scope of activities (both at technological and productive level) to which internal resources are devoted. This has been recognized early by scholars, who have often focused their efforts in this direction to reconcile theory with empirical evidence (Penrose 1960, Gort 1962, Berry 1971). Though diversification is interesting in and of itself, perhaps the more interesting question regards the degree of complementarity (or relatedness) between the various elements included in the portfolio of activities in which businesses engage. Notable early efforts to address this aspect have been proposed by Rumelt (1974) and Rumelt (1982). Both studies examine diversified manufacturing firms and focus on the link between profitability and the degree of correlation between the business units of the same firms. From this, they test the hypothesis that greater profitability correlates with expansion mainly in areas that share a competence or basic resource. Teece *et al* (1994) have built on the above intuition by employing plant-level data classifying establishments according to the standard four-digit SIC industrial codes relative to the industrial sectors in which they operate and measuring the relatedness between sectors through the frequency of their co-occurrence within the same productive plant, that is two sectors are related if many plants produce both. The hypothesis underlying this approach is the so-called *survivor principle* (Teece *et al* 1994), i.e. the assumption that economic competition eventually drives inefficient organizational forms out of the market, thus promoting the co-occurrence of activities that are well integrated with one another because of complementarities in *technological capabilities* they require. In virtue of the survivor principle, efficient combinations of activities should occur with a significantly higher frequency than one would expect if activities were paired randomly. Indeed, the authors find that internal coherence matters, as firms that diversify tend to add activities that are related to at least a part of their existing portfolio. More recent analyses confirmed this hypothesis (Rahmati *et al* 2020, Buccellato 2016, Lo Turco and Maggioni 2016).

Production is not the only aspect of corporate strategy in which building coherent portfolios of related activities has been shown to matter (for example in Gort (1962), Rumelt (1974), Berry (1971) the manufacturing sector is considered). Indeed, in the last twenty years, the empirical analysis of the innovative output of firms as measured by patents has gained increasing popularity (Rycroft and Kash 1999). It is worth noting that patent data have become in general a workhorse for the literature on technical change over the past few decades due to the growing availability of machine-readable patent documents and widespread access to sufficient computing power (Youn *et al* 2015). All the above have played a pivotal role in fueling this trend spurring scholarly (e.g. Hall *et al* (2001)), institutional (e.g. PATSTAT, REGPAT) and corporate (e.g. Google Patents) efforts aimed at constructing comprehensive collections of patent-related documents. Increasing data availability has in turn allowed researchers to inquire into the nature of patented inventions, their role in explaining the technical change, their reciprocal connections, and their link to the inventor-and applicant-specific characteristics (Strumsky *et al* 2011, 2012, Youn *et al* 2015). One of the characteristics of patent documents, which historically has lent itself more to economic analysis, is the presence of codes associated with the claims contained in the patent applications. These are used to mark the boundary of the commercial exclusion rights demanded by inventors. To allow evaluation by patent office examiners, claims are classified based on the technological areas they impact according to classifications (e.g. the IPC classification (Fall *et al* 2003)), which consist of a hierarchy of six-digit codes that associate progressively finer-grained definitions of technological areas to codes lower in the hierarchy. Mapping claims to classification codes allows for localize patents and patent applications within the technology space. Taking advantage of the increasing availability of patent data, several studies (Jaffe *et al* 2000, Leten *et al* 2007, Joo and Kim 2010, Rigby 2015) have found significant empirical evidence suggesting that evidence that relatedness in the composition of R & D activities has implications for the ability of firms to innovate successfully.

Within this stream of literature, a well-known study (Breschi *et al* 2003) has recovered the methodology proposed by Teece *et al* (1994) and built upon it to investigate whether firms tend to diversify their innovative efforts in a coherent fashion by patenting in technological fields that share a common knowledge base with the technological fields in which they innovated in the past. In particular, the authors have analyzed the technological diversification of firms through the co-occurrences between technology codes.

In another well-known paper, Nesta and Saviotti (2006) have studied corporate knowledge coherence in the US pharmaceutical industry showing that both the scope and the coherence of the knowledge base 'contribute positively and significantly to the firm's innovative performance', as measured by the number of patents it produces weighted by the number of citations received.

Some authors of the present paper introduced the concept of ‘coherent diversification’ (Pugliese *et al* 2019b), showing that firms that diversify (i.e., expand their technological portfolios by patenting in a relatively large number of technological sectors) in a coherent way (i.e., by preferring related sectors to unrelated ones) on average show a higher performance in terms of labor productivity.

In Yan and Luo (2017), the authors present a method for choosing an optimal compromise between the explanatory power of the diversification and the removal of the weak links in a network of technology codes.

Finally, we mention the work by Kim *et al* (2021), whom have studied the relatedness between technology codes in Korean firms, finding that ‘firms are more likely to develop a new technology when they already have related technologies’.

2. Results

The data we will use in this study is the matrix representation of the temporal bipartite company-technology network. In particular, we will consider 643 technology sectors embedded in the patents submitted by 197944 firms in 12 years. In practice, we will use 12 $\mathbf{V}^y$ matrices that link the layer of firms with that of technology codes, where y ranges from 2000 to 2011. In the following, we will interchangeably use the terms technological code, sector, or simply technology to express the same concept, since the codes written in the patents do represent technology sectors and so, in this sense, technologies.

The matrix element $V^y_{f,t}$ quantifies the patenting activity of firm f in the technology field t during year y. In particular, it is the number of patents submitted by the given firm in that sector. Note that this number can be fractional, since (usually) more than one code is present in each patent and (rarely) a single patent could be submitted by more than one applicant firm. In these cases, the unitary weight corresponding to one patent is split among the sectors and/or the applicants. Note that it may also happen that the same invention is linked to multiple patent application documents. In this case, each group of documents in the PATSTAT database is called ‘Patent Family’ according to primary citations among them (OECD 2001). Referring to the same inventions, these families are associated with the same technology codes, and they are counted as single patents. In summary, each matrix element V^y_{ft} is obtained as follows: we assign to each patent (or family of patents), in a given year y, one unit of weight. This is then divided into equal shares between all the observed (firm f-technology t) pairs and, finally, the matrix $\mathbf{V}$ is built by summing element-wise these contributions. The construction process is explained in more detail and a numerical example in the supplementary information (https://stacks.iop.org/JPCOMPLEX/3/035002/mmedia). The starting data is obtained by matching the *AMADEUS* database (https://amadeus.bvdinfo.com), that covers over 20 million firms with European registered offices, with the *Patstat* (www.epo.org/searching-for-patents/business/patstat) database about patent submissions. More details can be found in the Methods section and in Pugliese *et al* (2019b).

The matrix element $V^y_{f,t}$ gives a quantification of the patenting activity of firm f in the technology t. However, in the EFC framework, one usually deals with binary matrices; our choice is to use different thresholds T to define the 12 binary matrices $\mathbf{M}^y$, one for each year from 2000 to 2011, and to compare *a posteriori* the effect of using different values of T. In formulas, the binarizing procedure reads

$$\text{If } T = 0 \rightarrow M^y_{ft} = \begin{cases} 1 & \text{if } V^y_{ft} > T \\ 0 & \text{if } V^y_{ft} = T \end{cases}$$

$$\text{If } T > 0 \rightarrow M^y_{ft} = \begin{cases} 1 & \text{if } V^y_{ft} \geqslant T \\ 0 & \text{if } V^y_{ft} < T. \end{cases}$$

So the element $M^y_{f,t}$ is equal to 1 if a firm f submits more than T patents with technological code t in the year y, and 0 otherwise. We point out that in the economic complexity framework one usually binarizes the export (or, if patents are considered, the innovation activity) matrix using Balassa’s Revealed Comparative Advantage (Balassa 1965, Hidalgo *et al* 2018, Pugliese *et al* 2019a). Since in this case the $\mathbf{V}$ matrix is very sparse, the effect of RCA is practically negligible so we preferred to use the V^y_{ft} elements for clearer interpretability.

These $\mathbf{M}$ matrices can be used to train different algorithms to calculate our predictions about their temporal evolution. In order to have an out-of-sample forecast, we use data from 2000 to 2009 for the training phase and to obtain a score matrix $\mathbf{S}^{2011}$ which will represent the relatedness between companies and technologies; in other words, we expect that a higher value of the matrix elements $S^{2011}_{f,t}$ is connected to a higher probability for firm f to patent in technology code t in the year 2011.

We point out that both the matrices $\mathbf{V}$ and $\mathbf{M}$ are highly *autocorrelated* in time: if a firm does submit patents with a given technological code in a year y, it is likely that it will also be in the year $y + \delta$, and vice versa. As a

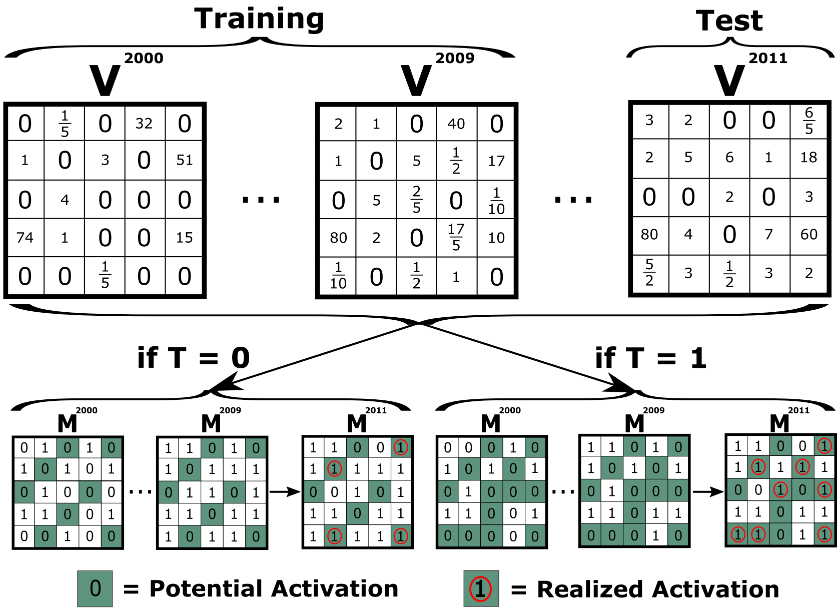

Figure 1. Schematic representation of the data processing. The **V** matrices represent the yearly bipartite networks that link firms and technologies; each element $V_{f,t}$ represents how much a technology t is present in the patenting activity of firm f. The element $M_{f,t}$ of **M** matrices instead, gives us the binary information about *whether* a technology t is made by a firm f or not, that is if $V_{f,t}$ exceeds or not a threshold T. The dark green elements are the potential activations and the red circles in 2011 matrices are the realized activations (i.e. those elements that were 0 in all training years and then, in 2011, become 1: a new technology for that firm). Also, the distinction between training and test set is shown.

consequence, we focus our attention on those matrix elements that we call *potential activations*: the elements of **M** that are 0 in all training years (from 2000 to 2009). Then, we will check whether in the test year (2011) this element remains equal to 0 or becomes 1. We will call this last case a *realized activation*: a firm enters (that is, starts patenting) in a technological sector which is new to this firm. In figure 1 we represent a numerical example clarifying how we managed the **V** and the **M** matrices, the division of the data in training and test set, and the definitions of both potential and realized activations.

Our forecast exercise permits to compare different prediction algorithms using the test year 2011. So we will compute one score matrix $\mathbf{S}^{2011}$ for each algorithm and we will compare it with $\mathbf{M}^{2011}$ (obtained by binarizing the empirical $\mathbf{V}^{2011}$), and quantify the prediction performance as in usual supervised classification tasks (Kotsiantis *et al* 2007).

In order to obtain the prediction scores, we use different algorithms to evaluate the relatedness (Hidalgo *et al* 2018) between a firm and a technology. In the case of co-occurrences based networks, an intermediate step is to assess the *similarity* between technology codes. Here we list the tested algorithms by category, leaving a more detailed discussion for the methods section.

- **Benchmarks**: we use a quasi-trivial random and autocorrelation-based predictions as benchmarks. The first is a random model where we fix the diversification of the firm $d_f = \sum_t M_{f,t}$, i.e. the number of the technology codes in its patents. The second is a benchmark model that takes into account the temporal autocorrelation of the **M** matrices: the scores **S** are equal to the mean of V^y_{ft} in all the training years (i.e. with $y \in [2000, 2009]$).
- **Networks**: the standard economic complexity approach usually starts from the evaluation of normalized co-occurrences; in the simplest case

$$B^y_{t,t'} = \sum_f M^y_{f,t} M^y_{f,t'}$$

that is, t and t' are similar if many firms patent in both sectors. Different normalizations lead to the product space, or in this case, the Technology Space (Hidalgo *et al* 2007), the Taxonomy Network (Zaccaria *et al* 2014), and the micro-partial network, based on the paper of Teece *et al* (1994). In all these cases, network **B** represents a projection of the bipartite network **M** into the space of technology codes, and each element $B_{t,t'}$ represents the proximity/similarity between the two technology codes. In order to obtain a measure of the relatedness between a firm f and a target technology t, to be used as a prediction score, one then computes the coherence (Pugliese *et al* 2019b) using equation (2). Other approaches, such as the density normalization introduced by Hidalgo *et al* (2007), perform sensibly worse. More details are provided in the methods section.

- **Machine learning**: since our prediction exercise can be expressed as a supervised classification exercise, we can use the random forest (RF) algorithm (Breiman 2001, Albora *et al* 2021), and what we call the CTS. The first is a popular machine learning algorithm based on decision trees, while the CTS is based on the studies of Tacchella *et al* (2021), and it is a projection on the space of the technology codes obtained from the scores obtained with the RF, which can be seen as a high dimensional representation of the codes themselves. This is done by using a variational auto encoder (Kingma and Welling 2013) followed by the t-SNE dimensionality reduction algorithm (Van der Maaten and Hinton 2008). In this way, we are able to make the results of the RF, in a sense, more interpretable. As better specified in the methods, the RF scores can be seen as a high dimensional representation of the technology codes (one dimension for each firm). In order to visualize this space, the t-SNE dimensionality reduction algorithm is applied, which results in the CTS. Note that in order to produce prediction scores from the CTS one has to compute a coherence or density measure as in the network-based approaches. The use of RF somehow hides the reason why a company is close to a technology code, in other words, where the relatedness results come from. However, by applying t-SNE to the prediction scores one can obtain a visual representation of the relative position of the codes in this new space we define. Now the motivation behind our relatedness assessment, and the consequent forecast, becomes (hopefully) more interpretable: the company is close to a given technology if it is already patenting in close technologies. This is visible as a diffusion process only if a low-dimensional representation is adopted. We point out that this is not an explanation of how the RF works, which is beyond the scope of the paper, but *a posteriori* justification of our results that, being represented in a two-dimensional plane, can provide insights into companies' innovation strategy.

  Two types of RF are used, the non-cross validated (RF) and the cross validated one (RF_CV). With the cross validation, we remove a portion of firms at a time from the training, and then we use them in the test. The starting rationale is that the algorithm produces its predictions by using two pieces of information: the similarity between technologies and its ability to recognize a firm. By cross-validating the RF we try to force the algorithm to use the former, and not the latter (Albora *et al* 2021).

2.1. Prediction results

Here we compare the relatedness assessments of the co-occurrences based networks with the machine learning algorithms, showing how the latter are able to give better prediction results. The results are shown in figure 2.

In order to compare the various prediction methods from various viewpoints, we adopted different metrics to quantify the goodness of a prediction (these metrics are discussed in detail in Methods section):

- **Area under the PR curve**: the area under the curve in the precision-recall plane. The latter is obtained by varying the threshold that identifies the value above which the scores are associated with positive predictions;
- **Precision@100**: the fraction of the largest 100 elements of the score matrix $\mathbf{S}^{2011}$ that are actually activated;
- **mPrecision@10**: for each firm, we consider the largest 10 scores and we compute the fraction of realized activations; then we average over the firms.

In figure 2 we report the scores of the previous metrics for different values of the threshold parameter T; the results are consistent even if one varies such threshold (or uses the RCA to binarize).

We start noticing that the random benchmark is surpassed by all the different approaches, showing that all are able to compute a measure of similarity that is able to grasp links between the technology codes.

Also the autocorrelation benchmark is outperformed but in the mPrecision@10 case. In particular, it performs better when T increases, because the number of zeros in both the training and the test matrices increases (that is, the number of potential activations, the green elements in figure 1, that are not realized).

In the area under the PR curve and Precision@100, the only network-based algorithm that manages to overcome the CTS is the Taxonomy. In particular, it is interesting to observe how this network exceeds the Technology Space. We can argue that for the technology codes, a network based on the taxonomy principle, i.e. how firms move from low-complexity to high-complexity technologies only after developing the necessary skills (Zaccaria *et al* (2014)) shows a better prediction performance that a proximity-based one, i.e. a network where two technologies have a high link if they need the same capabilities (Hidalgo *et al* 2007).

The micro-partial approach does not show a competitive performance despite being quite popular in both academic and corporate applications (Smith and Linden 2017).

In any case, the superiority of the RF_CV and of the RF with respect to both benchmarks and density-based approaches (networks and CTS) is evident. Although the other algorithms are able to give prediction scores able to overcome the benchmark models (especially for $T = 0$), clearly these are not able to fully highlight the

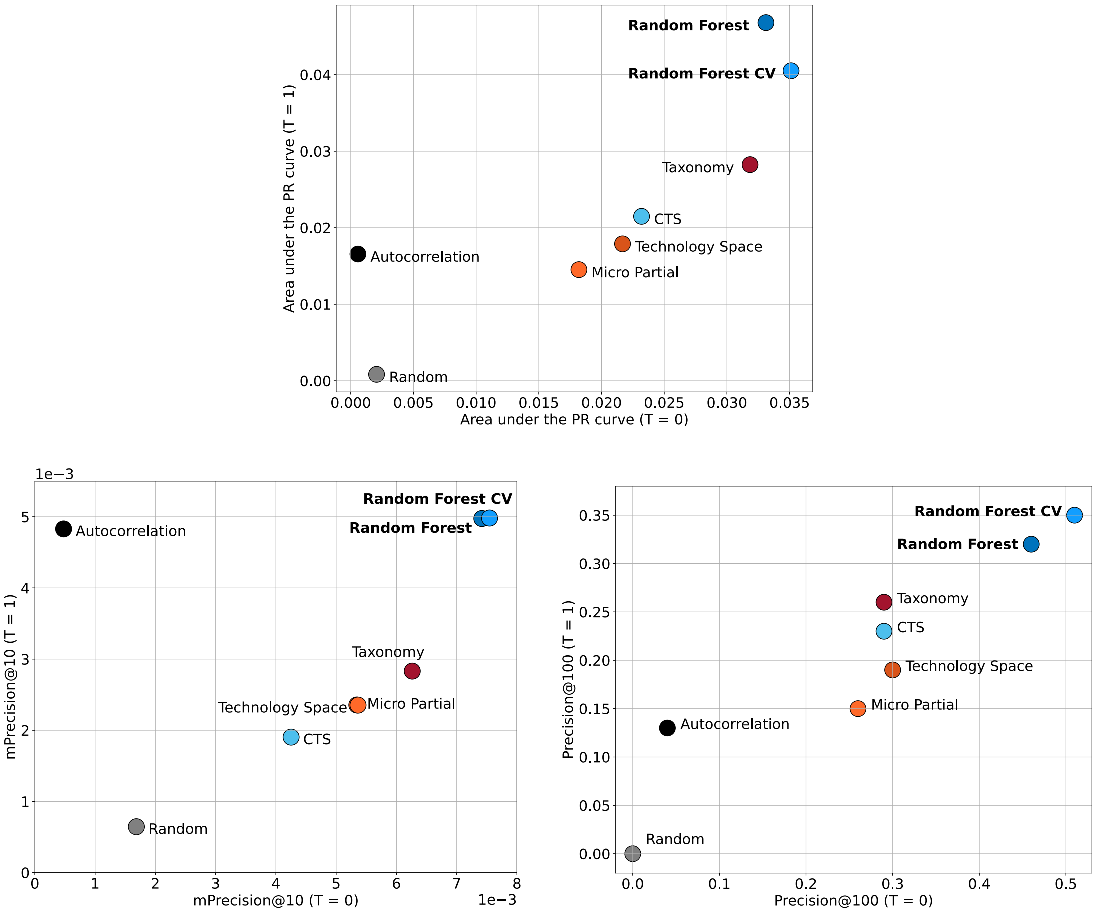

Figure 2. Comparison of the prediction results obtained by different approaches, using three different evaluation metrics and two binarization thresholds. RF outperforms all other approaches in all perspectives. Regarding the area under the PR curve, the second bests are the Taxonomy Network and the Continuous Technological Space.

non-linear relationships among the technological portfolios of firms and the technological sectors they will move to.

In figure 3 we compare the frequency distributions of the scores of both the realized and the not realized activations for the RF (y-axis) and Technology Space (x-axis). In order to make them comparable, both scores are rescaled using the respective maxima and minima. The red line is the bisector, shown for further reference. From the left figure, it emerges that the RF assigns, on average, higher scores to those potential activations which will be actually realized in two years. On the contrary, the possible but not realized activations show similar distributions; this is due to the much greater number of true negatives which is present in both approaches. Note that, as expected, the scores given to the not realized activations are lower than the realized ones.

Finally, it is also important to point out that in the present study, true positives are more significant than true negatives. This has a twofold rationale:

- The high class imbalance implies that a performance measure such as accuracy is not adequate for the problem. The majority of the elements of our matrices are equal to zero and therefore an accuracy measure would consider only the overwhelming number of true negatives. Even if we made a prediction in which we assume that all the elements of the matrix will be zeros we would get an accuracy higher than 99%. More in detail, the percentage ratio between elements equal to 1 and 0 is about 0.2% for $T = 0$ and 0.08% for $T = 1$.
- For a firm, it is more interesting to know which technologies are close to it than those it is already active in (i.e. which technologies it can successfully activate in the future), rather than knowing which ones it will not do in the future (which is often trivial information, due to a totally different scope, for instance).

2.2. Continuous Technology Space

Even if the prediction performance of the RF algorithm vastly outperforms the other approaches, its practical feasibility in policymaking could be limited by its low interpretability. From a policy perspective, indeed, it is

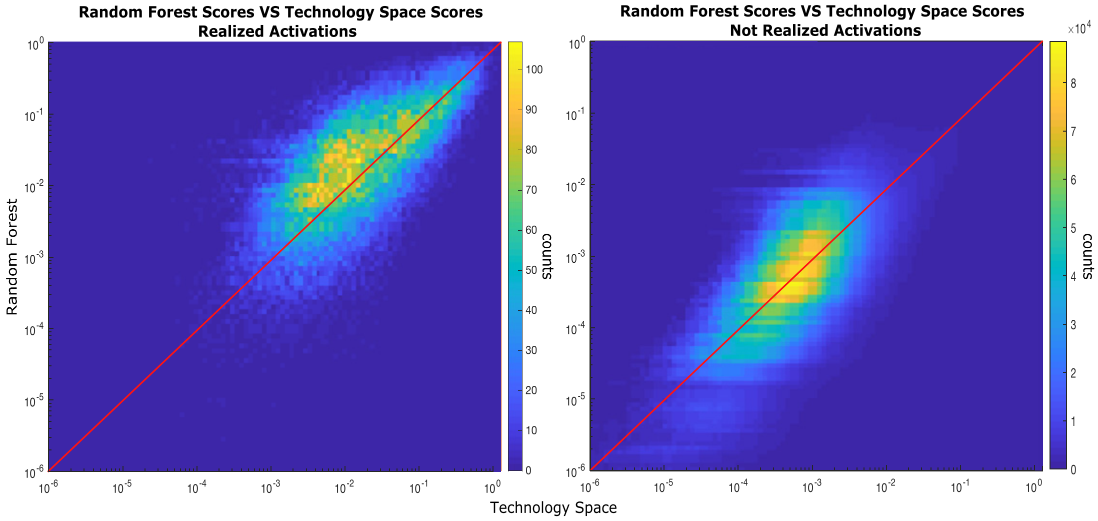

Figure 3. Comparison of the RF and Technology Space scores with a 2D histogram plot. In each figure, the bins' color represents the number of data points within each bin. The left figure is referred to the realized activations, i.e. those elements that are always 0 during the training years (from 2000 to 2009) and then become 1 in the test year (2011). The red line is the bisector. The realized activations' scores obtained with the RF are, on average, higher than those obtained with the Technology Space. The right figure is referred to the not realized activations, i.e. elements that also in 2011 are zero: here, the distributions are roughly similar.

not easy to justify a strategic decision such as investing or not in a technological sector on the basis of a quasi-black box algorithm. In order to provide a visual tool to inform and justify strategic decisions, we introduced the continuous Projection Space (Tacchella *et al* 2021), that uses the scores obtained from the machine learning algorithms to build a two-dimensional and, as such, easy interpretable space to visualize and describe the temporal evolution of bipartite networks. The key idea is to interpret the scores matrix obtained with the RF as a matrix of coordinates of technology codes in a high-dimensional space. These embeddings are then made representable by applying suitable dimensionality reduction techniques; in this case, t-SNE (Van der Maaten and Hinton (2008)). Note that here we are using the term 'interpretable' in a policy perspective: in this sense, machine learning algorithms are a black box in the sense that, for a policymaker, the origin of our results is not clear *a priori* (for instance, a company being close to a technology code). However, by applying t-SNE to the RF scores one can obtain a visual representation of the relative position of the codes in this new space we define. Now the motivation behind our relatedness assessment, and the consequent forecast, become (hopefully) clear: the company is close to a given technology if it is patenting in the close technologies. This is visible as a diffusion process only in a low-dimensional representation. We point out that this is not an explanation of how the RF works, which is beyond the scope of the paper, it is just an *a posteriori* justification of our results that, being representable in a two-dimensional plane, can provide insights about companies' innovation strategy (e.g., exploration vs exploitation).

Here we apply this methodology—which is fully described in the Methods section—to the firm-technology network; the result is a plane in which each technology sector is a point, and similar sectors are close. In figure 4(a) the different colors correspond to IPC macro-categories, i.e. the first of the 4 digits that define the classification codes. We point out that, differently in network-based representations, here the similarities are simply represented by the spatial proximity between technology codes. The use of Euclidean distances instead of topological ones permits to use of a wider range of tools, for instance, clustering and anomaly detection algorithms. A visual inspection of the CTS permits to obtain a number of insights: in figure 4(a) one can observe that technology codes tend to cluster following the macro-categories; this is a first hint that the positions in the plane present a certain degree of significance. However, also the departures from the classification reveal meaningful relationships. In particular, on the upper left one can observe the presence of a dense area where it is possible to find veterinary medicine close to farm. In the motor vehicles area on the left, we find motor vehicle technology codes; in particular, there is a red color technology code (A47C) that corresponds to chairs and seats specially adapted for vehicles, black technology codes colors, corresponding to B60 (considering the first 3 digits), that represent vehicles, light blue technology codes corresponding to the first 3 digits F01 and F02, i.e. machines and engines, and combustion engineering, and one brown technology code color (G05G), physics of command systems. Weapons area is associated with weapons technologies: we find principally (considering the first 3 digits) codes B63 and B64, i.e. ships and aircrafts, C06 associated with explosive chemistry, and F41 and F42, i.e. weapons and ammunition.

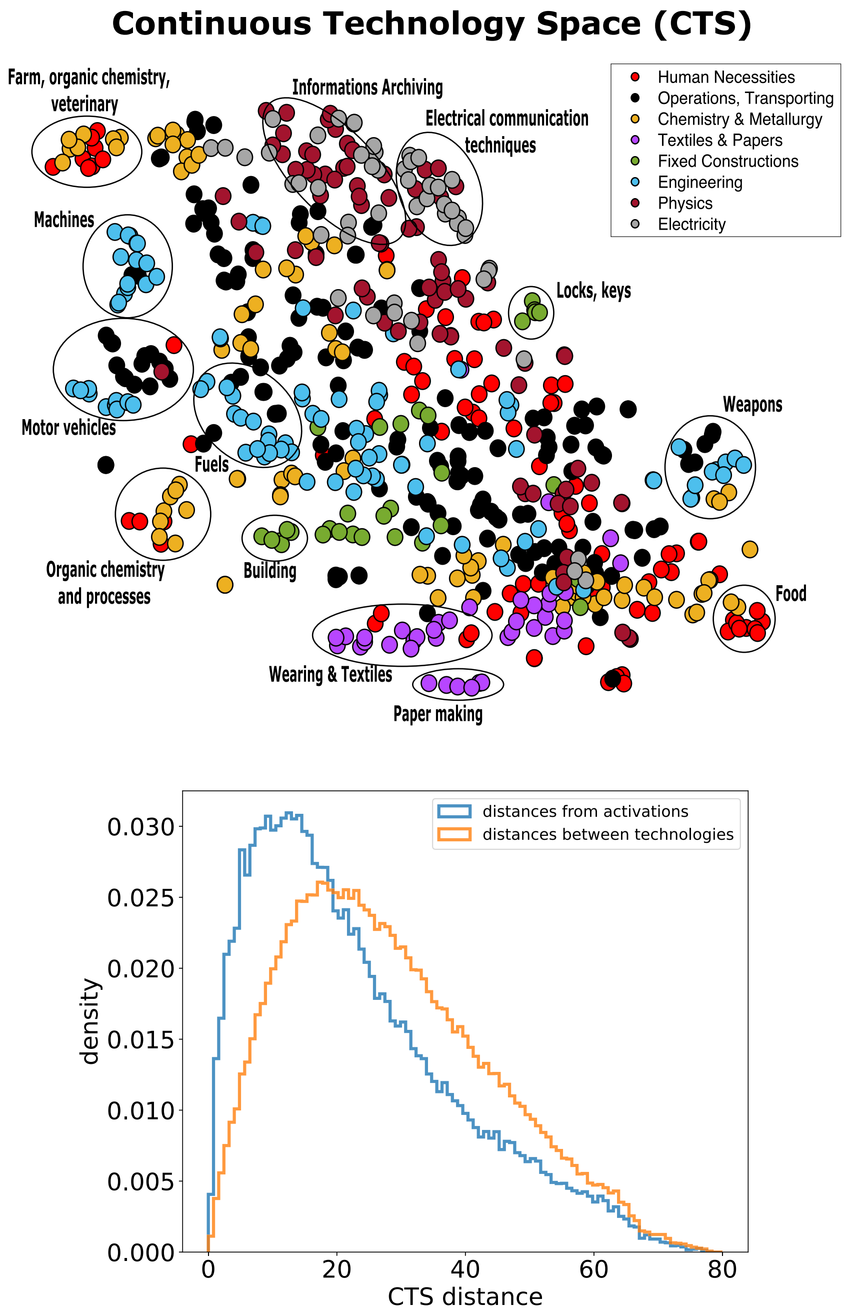

Figure 4. Above: representation of the CTS. Each point is a technology code and each color is associated with the respective macro category, i.e. the first of the 4 digits. Clusters related to macro categories are evident. Below: comparison between two density distributions of distances. The first (cyan line) represents the distances between activation codes patented by a company from the closest code the company used; the second (orange line) is the density distribution of distances between points in the CTS.

The example discussed above can be generalized by comparing two frequency distributions of distances (see figure 4(b)). The first distribution (cyan line) is relative to the distances between the newly activated codes and the patenting company (that is, the distance from the closest sector the company patented in); the second (orange line) is the density distribution of all distances between the points in the CTS. We can see how the first distribution has a lower mean, evidencing that, on average, firms tend to patent in codes that are relatively close to what they already do. Obviously, also high distances are present, indicating strategic choices which lead the firm far from its usual scope.

In order to show a concrete application of the CTS, we show in figure 5 the portion of this space relative to an American nanotechnology company, Nanotek Instruments Inc., as an example. In 2002 Nanotek patented three inventions, two based on batteries (https://patentimages.storage.googleapis.com/f4/d8/3d/d663e43fe48e2b/US6773842.pdf and https://patentimages.storage.googleapis.com/66/b3/7f/6fa873ae402fbf/US6864018.pdf) and the third is the *nano-scaled graphene plates* (https://patentimages.storage.googleapis.com/e5/3d/0d/1c25e5f68a77ab/US7071258.pdf). The first two are associated with

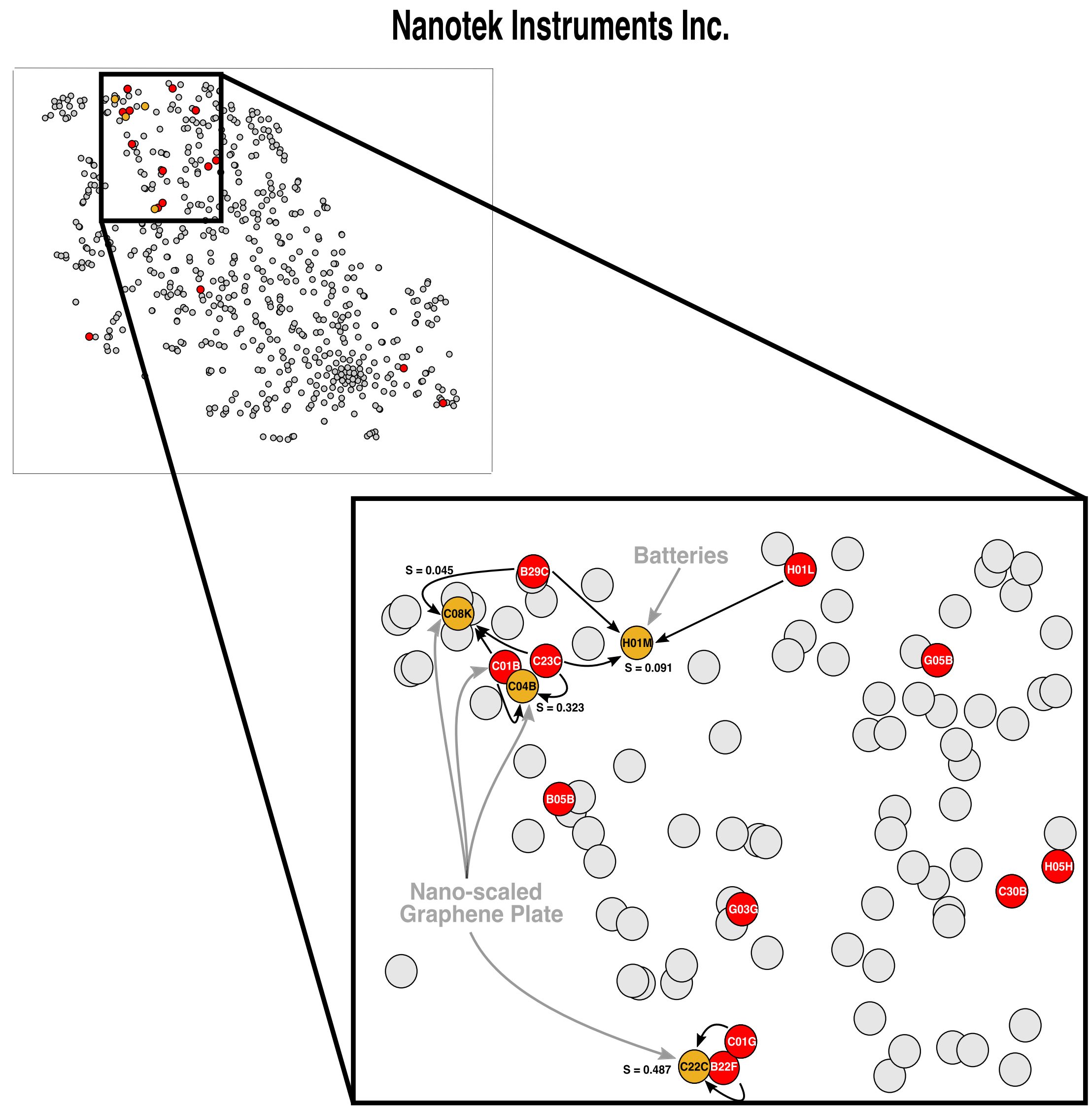

Figure 5. The exploration of the continuous Technological Space by an American nanotechnology company. Starting from the red sectors, which Nanotek patented in the past, this company moved nearby, patenting in the gold sectors in the next years.

the code *H01M*, while the third with the codes *C08K*, *C04B*, *C01B* and *C22C*, which correspond to the gold points. The red points are the technology codes which Nanotek patented in 2000 and 2001, while, as mentioned, the gold ones are those activated in 2002. The black arrows underline the non-random position of the new technologies, that are close to the ones already present in the patenting activity of the company. This is because we find that technology codes that have a high similarity are represented close to each other, and therefore a sort of 'technological diffusion' is expected starting from the codes that firms already have in their portfolios (as shown in figure 4(b)).

3. Discussion

In this work we compare machine learning and network-based approaches to forecasting which will be the future patenting activity of firms; in particular, their next technological sector of innovation. To the best of our knowledge, this is the first attempt to assess the relatedness between a firm and a technology sector using machine learning. In order to compare the various possible measures of relatedness, we analyze a very large database consisting of about two hundred thousand firms and 643 technology sectors and we develop a forecasting exercise using the assumption that, on average, firms will patent in sectors related to their present technological activity. We find that supervised machine learning techniques RF clearly outperform the standard methodologies usually adopted in economic complexity, that is, networks of co-occurrences. Our results are robust with respect to different definitions of what a 'new' technological sector is, and if different metrics to evaluate the prediction performance are adopted. Indeed, RF assigns on average higher activation scores to those technologies which will be explored by firms with respect to all network-based approaches. Finally, we introduce the CTS, which permits to visualize the dynamics of firms during their innovation activity. The introduction of this approach opens up a number of possible applications and developments. First of all, our activation scores represent an assessment of the achievability of a given jump to a new technology sector, a measure of how easy will be to produce innovations in that sector given the present activity of the firm. Moreover, the CTS allows a compact visualization of the past, the present and the possible patenting activity of a firm. Using these tools, and in the spirit of the 'adjacent possible' approach of Kauffman (1996), Tria *et al* (2014), it is now possible to quantify how much a firm is *exploring* the space of technologies or *exploiting* what it already does. One can then compare different strategic choices with various measures of performance, both in terms of profitability and further innovation activity. Furthermore, these measures can be applied to investigate mergers and acquisitions, and in particular to study whether acquirers prefer to target companies that are 'close' or 'far' from their present patenting activity. Finally, following the work of Brummit *et al* (2020) and Pugliese *et al* (2019b), a future research project could be the prediction of some performance-related monetary variables of firms, such as revenue or labor productivity, from knowledge of firms' patent activity.

4. Methods

In this section, we describe in more detail the database, algorithms, and evaluation metrics used in the analysis.

4.1. Data

The bipartite firm-technology network is obtained by matching two databases: AMADEUS for firms and PATSTAT for the technology codes.

4.1.1. Firms

AMADEUS (https://amadeus.bvdinfo.com) contains information about over 20 million companies, mainly concentrated in the European continent. This database is managed by Bureau van Dijk Electronic Publishing (BvD) which specializes in providing financial, administrative and budget information relating to companies. It is compatible with the PATSTAT database for patents as BvD includes the same patent identifiers as the European Patent Office (Pugliese *et al* 2019b). We mention here one of the well-known problems with AMADEUS, namely that large companies are fully covered while those with less than 20 employees are under-represented (Ribeiro *et al* 2010); however, this is not a severe issue for the present analysis.

4.1.2. Technology codes

The dataset from which we take information about the patent and the technology codes is PATSTAT (www.epo.org/searching-for-patents/business/patstat). Globally, PATSTAT considers approximately 100 million patents registered in about 100 Patent Offices. This information spans from the mid-19th century to three-four years before the release of the database; this is evident from the quickly decreasing number of patents in the last available years. As a consequence, we decided to restrict our analysis to a conservative time interval (2000–2011). A key element is the presence of a set of alphanumeric codes in each patent submission; these codes can be assigned by the inventors or by the reviewers and represent the technology sector the patent belongs to. The WIPO (World International Patent Office) uses the IPC (International Patent Classification) (Fall *et al* 2003) to assign these technology codes to each patent in such a way as to classify, and better manage, the inventions presented. The IPC codes define a hierarchical classification consisting of six levels: sections (that we call macro category), sub-sections, classes, sub-classes, groups and sub-groups. For example, code Axxxxx corresponds to the 'Human necessities' macro category and Hxxxxx to the 'Electricity' macro category; considering the following digits we have, for example, with A01xxx the sector 'Agriculture; Hunting', and with A43xxx the 'Footwear' sector. It is important to note that we discard classes '99' and sub-classes 'Z', as they represent other technologies not classified in other classes or sub-classes, and they are therefore not well defined.

It may happen that the same invention may be referred to for multiple patent application documents. In this case, each group of documents in PATSTAT is called 'Patent Family' according to primary citations among them (Publishing *et al* 2001), which is nothing more than the set of patents presented in different countries to protect the single invention. Patent families can be built with different criteria (Martínez 2011), but among these, we choose the one related to the 'Extended family', also called IN-PADOC. This corresponds to the category considered in such a way as to associate the inventions with the widest possible technological spectrum. Once patents are assigned to firms, we can assign them the corresponding technology codes and build the firm-technology bipartite network, and its adjacency matrix $\mathbf{V}^y$, one for each year y. The matrix element represents the number of the patent submitted by the firm in the technology sector. Note that this number may be fractional since more than one code is usually present in the same patent: for instance, if a firm submits only one patent with three technology codes, the three nonzero elements of the corresponding row of $\mathbf{V}^y$ will be equal to one third. The interested reader can find more details about this data in the results section and in Pugliese *et al* (2019b).

4.2. Data processing

The starting database can be represented using the following structure: 12 matrices, one for each year from 2000 to 2011, that link 426983 firms f (rows) to 7456 (six-digits) technology codes t (columns). We chose to work at a higher aggregation level, and so to compress the technology codes from 6 to 4 digits, summing the columns corresponding to the 6-digit codes with the same first 4-digits. From the 6 to the 4 digit level the number of technologies goes from 7456 to 643. This operation leads to both better quantitative results and shorter computation times (from a qualitative point of view, instead, the results are unchanged).

A key element of both the machine learning and the network-based approaches is to provide an assessment of the similarity between technology codes; this information can be extracted from the co-occurrences of technological sectors in the same firms. So we consider only firms that, in the years from 2000 to 2009, make at least 2 technology codes; these firms are 197944.

This leads to the data mentioned in the main text: 12 $\mathbf{V}$ yearly matrices that link 197944 firms and 643 technology codes.

In order to compute the relatedness measures, in the economic complexity literature (Hidalgo *et al* 2007, Zaccaria *et al* 2014, Pugliese *et al* 2019a) one usually computes the revealed comparative advantage or RCA (Balassa 1965), and then these matrices are binarized using a threshold equal to 1. As far as exports are concerned this choice of threshold has a natural economic meaning, traceable to the works of Ricardo and Balassa himself: considering the bipartite country-product network, $\mathrm{RCA}_{c,p} \geqslant 1$ means that country c is significantly competitive in the export of the product p (Hidalgo *et al* 2007). So the country's share of that product in its market is equal to or greater than the product's share on the world market. However, the economic meaning of patent submission is different, so the choice of RCA is not straightforward. In this work, we binarize the matrices $\mathbf{V}$ with different values of threshold T, without computing the RCA; in this way, the matrices $\mathbf{V}$ are better interpretable as how much a technology code t is present in the patenting activity of a firm f. We have in any case check the robustness of our results for different threshold values and the use of RCA.

4.3. Network-based approaches

In this and in the next sections we discuss how to obtain a prediction score matrix $\mathbf{S}$ for 2011 from each method starting from the same training data $\mathbf{V}$ and $\mathbf{M}$, relative to the years 2000–2009. The score matrix gives the model's estimation of the likelihood that a firm will patent in the given technology sector, and the comparison between the scores and the actual $\mathbf{M}^{2011}$ using the performance metrics will give an assessment of the models' performance.

The basic idea of network-based approaches is to compute similarity of technology codes from their co-occurrences in companies. Introduced by Teece *et al* (1994), and popularized in the network/complexity community by Hidalgo *et al* (2007), the basic quantity is the number of firms that have patented inventions relating to both codes:

$$B^{\mathrm{CO}}_{t,t'} = \sum_f M_{f,t} M_{f,t'}. \tag{1}$$

The idea is that if many firms are active in two technology sectors t and t' at the same time, this means that the capabilities, the techniques and, in general, the necessary means to patent in these sectors, are roughly the same, and so these sectors are, in this sense, similar, or related.

Different scholars presented various ways to normalize the co-occurrences, on the basis of different theoretical frameworks or interpretations. In general, we can write:

$$B_{t,t'} = \frac{1}{A} \sum_f \frac{M_{f,t} M_{f,t'}}{C}$$

and discuss the various options for the quantities A and C:

- Simple co-occurrences (Teece *et al* 1994): for $A = 1$ and $C = 1$ one simply counts the number of companies that are active in both sectors. This case corresponds to $B^{\mathrm{CO}}_{t,t'}$ of equation (1);
- Technology Space (same normalization of the Product Space (Hidalgo *et al* 2007)): $A = \max(u_t, u_{t'})$ and $C = 1$, where $u_t = \sum_f M_{f,t}$ is the ubiquity of technology code t, that is, the number of firms active in that technology sector. Using this type of normalization we give a lower connection weight to those technology codes done by many firms, which we can consider basic.
- Taxonomy (Zaccaria *et al* 2014): $A = \max(u_t, u_{t'})$ and $C = d_f$, where $d_f = \sum_t M_{f,t}$ is the *diversification* of firm f. The Technology Space, for how it is built, gives a higher score for high complexity technology

codes (i.e. codes done by few firms) and, as a result, bias towards them. Consequently, it is not possible to justify the evolution of low-complexity technology codes toward high-complexity ones. Normalizing also for the diversification we avoid this problem as we penalize low ubiquity scores and low complexity technology codes are weighted more.

- Micro partial (Teece *et al* 1994): we compute

$$B^{\mathrm{MP}}_{t,t'} = \frac{B^{\mathrm{CO}}_{tt'} - \mu_{tt'}}{\sigma_{tt'}}$$

with

$$\mu_{tt'} = \frac{u_t u_{t'}}{N},$$

and

$$\sigma^2_{tt'} = \mu_{tt'} \frac{(N - u_t)(N - u_{t'})}{N(N-1)},$$

where N is the number of companies. Here we use a null model in which the ubiquities of the technologies are kept fixed and everything else is randomized. This case can be analytically solved: the resulting distribution for the co-occurrences is hypergeometric with mean $\mu_{tt'}$ and variance $\sigma^2_{tt'}$. We call this network micro partial following the notation used by Cimini *et al* (2022): this null model is microcanonical in the sense that the degree sequence is exactly fixed and partial because only one layer is constrained. So the idea is that, if the weight of the link between two technology codes t and t' exceeds the expected value $\mu_{tt'}$, this means that t and t' are highly related with respect to this random case. Furthermore, as a t-statistic, $B^{\mathrm{MP}}_{t,t'}$ measures how much the observed link between the two technology codes exceeds what would be expected if the companies were randomly assigned.

For the latest formulas, we obtain one matrix $\mathbf{B}^{\mathrm{Net}}$ for each network. In order to consider all years available in the training data, we used as **M** matrix in the previous formulas a total matrix obtained by summing the **V** matrices from the years 2000 to 2009, using all the 197944 firms, and then binarizing this sum.

Based on the network used, we get a $\mathbf{B}^{\mathrm{Net}}$ which we use in the coherence equation from Pugliese *et al* (2019b):

$$S^{2011}_{f,t} = \sum_{t'} M^{2009}_{f,t'} B^{\mathrm{Net}}_{t't}, \tag{2}$$

where $M^{2009}_{f,t'}$ is the **M** matrix obtained by binarizing the $\mathbf{V}^{2009}$ matrix. In practice, t is highly coherent with the patenting activity of firm f if f is active in many sectors highly connected with t. On the contrary, if a sector is far from what a firm actually does, we will assign it a lower activation score. Note that this equation differs from the density equation of the Product Space (Hidalgo *et al* 2007); we use coherence instead of density since we have found a better predictive performance.

4.4. Random forest

RF (Breiman 2001) is a tree-based machine learning algorithm that we use to better capture the non-linear links between technology codes. In particular, we use this binary classification algorithm to determine whether or not a technology code will appear in the patenting portfolio of a particular company in the future starting from the knowledge of the technology codes in which the firms patented in the last training year.

In general, during the training of a supervised machine learning algorithm, an input data **X** matrix is passed. Because our problem is a supervised one, each vector (row) of the matrix is associated with a label presented in a different input **y**. To give an example, **X** can be the matrix where each row is a flattened handwritten digit, and each element of the row is the intensity of a pixel; in this case, **y** will be the label corresponding to the digit, and that must be associated, in order to be recognized, to all those present in **X**. Once the model is trained, one gives new samples $\mathbf{X}_{\mathrm{test}}$ and the model is able to make associate a prediction $\mathbf{y}_{\mathrm{test}}$ (in this case, a digit), to each sample.

In our case, we train one RF for each technology code: we want the RF to learn which typologies of portfolios are associated with each code after two years. So, as samples matrix **X** we use the matrix obtained by concatenating, or stacking vertically, the **V** matrices from the year 2000 to 2007, and as **y** we use one column at a time (and therefore one technology code at a time) of the matrix obtained by concatenating the matrices **M** from the year 2002 to 2009. In this way, each row is a firm in a year from 2000 to 2007, that has 643 features. We associate this sample to the respective label in **y**, that is, if after 2 years the technology code associated with the element in **y** is active, or not. In such a way, we associate the codes of each portfolio with the possible presence of the target code in the future.

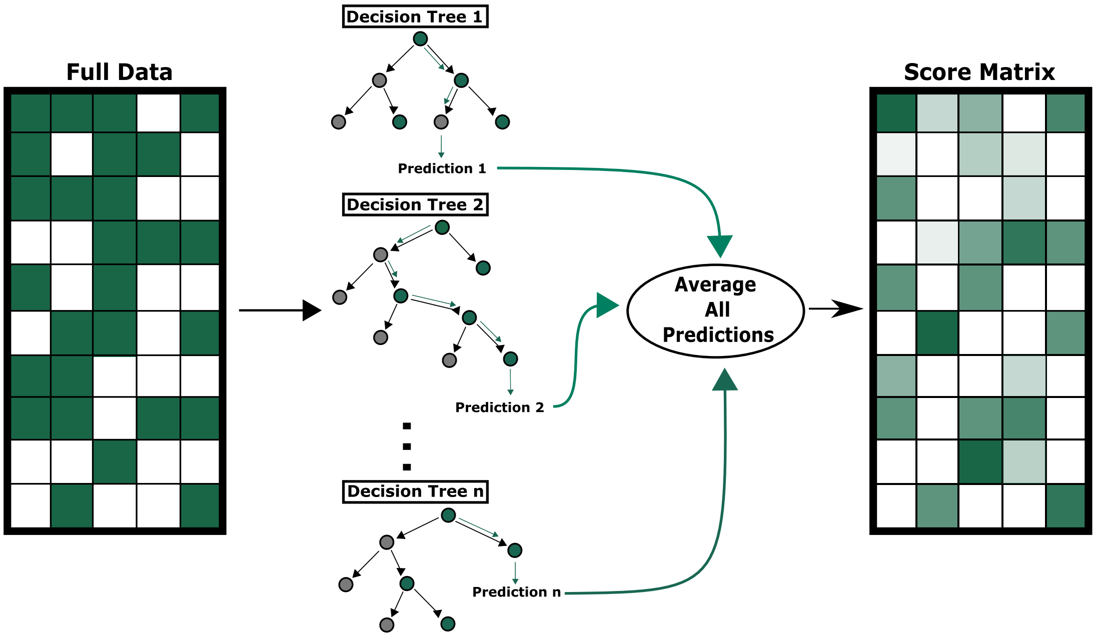

Figure 6. We show how the score matrix $\mathbf{S}^{2011}$ is obtained by the RF by combining the predictions of different decision trees.

From a practical viewpoint, we use the 'RandomForestClassifier' from the 'sklearn.ensemble' python library (Pedregosa *et al* 2011), called in this way:

$$\text{RandomForestClassifier.fit}(\mathbf{V}_{2000}^{2007}, \overrightarrow{\mathbf{M}}_{2002}^{2009}),$$

where $\mathbf{V}_{2000}^{2007}$ are the vertically stacked matrices and with the vector symbol over $\mathbf{M}$ we indicate that one column is used at a time, that is, we train one RF for each technology code. The delay of 2 years is used to insert a dependence on time, as we want to produce forecasts about the innovative development of firms. We optimized the RF parameters as described in the supplementary information; the results shown here refer to: number of trees $= 50$, min_samples_leaf $= 4$; max_depth $= 40$ and method $=$ 'entropy'. The use of all available companies in the training is computationally demanding, so we used only the top 10 000 most diversified firms (10KHD firms). If we use more firms for the training we get a saturation of the forecast performances (see the supplementary information). The fact that firms with higher diversification should be used is due to the fact that these provide better coverage of the possible technologies and the possible combinations among them.

After fitting the data, that is, training the machine learning model, we obtain the $\mathbf{S}^{2011}$ scores by using the $\mathbf{V}^{2009}$ matrix as $\mathbf{X}_{\text{test}}$ in a predicting phase. The command line reads

$$\mathbf{S}^{2011} = \text{RandomForestClassifier.predict\_proba}(\mathbf{V}^{2009})$$

and this associates a probability to activate the target technology to each firm in 2009. In figure 6 we schematically represent how $\mathbf{S}^{2011}$ is obtained from the RF, the latter being represented by a set of differently trained decision trees.

In this paper, we also compare the approach described above with a cross-validated RF, for which we use the same optimized parameters and, in order to provide a consistent comparison of the results, the same training and test sets. In this respect, note that the $\mathbf{X}_{\text{test}}$ should always produce a prediction for all the 197944 firms (including the 10KHD firms used in the training). In the cross-validation framework, we train $k = 4$ different RFs, using the technique called *k-fold cross validation*, that is we separate into $k = 4$ groups both the 10KHDs used for the training and the 197944-10KHDs used for the test. Then, in training each RF we remove one of the 4 groups of 10KHD companies. The test is instead performed on the removed group from the 10KHD companies along with 1/4 of the 197944-10KHD firms used for the test. This is performed 4 times, each time removing a different set of firms. In the end, we will have prediction scores for all companies by merging the scores produced by the four RFs.

The idea behind the use of cross validation is the following. During the training, the RF basically learns two pieces of information: to recognize the portfolio of a company and the similarity among technologies. Even if we are more interested in the latter, the learning of the two cannot be avoided. However, we can try to force the algorithm to use the similarities in the test phase: if we give a new company in the $\mathbf{X}_{\text{test}}$, the RF cannot recognize it and so it is forced to use the similarities to produce its predictions. This procedure, even if computationally more demanding, leads to better results, as shown in figure 2.

4.5. Continuous Technology Space

RF shares with most the machine learning algorithms an intrinsic difficulty of interpretation, i.e. the rationale behind how the input is connected to the output is not evident. In this respect, network approaches (note: if made sparse by a suitable filter) are more clear, since the coherence or density-based approach are clearly visualizable: a technology is coherent with a firm's portfolio if has a lot of heavy connections with what the firm already does. In order to restore the interpretability of networks and keep the predictive performance of machine learning, Tacchella *et al* (2021) propose the Continuous Projection Space, which here we reformulate, with suitable modifications, as the CTS.

To compute the CTS we start from the RF CV method but, in $\mathbf{X}$, only the first 2K HD firms are used, because we have a saturation of the scores: using more firms does not change the scores and increases the computational time.

Another difference with the RF CV is that the predictions are obtained using as $\mathbf{X}_{\text{test}}$ the same 2K HD firms used for the training (i.e., in the CTS $\mathbf{X} = \mathbf{X}_{\text{test}}$; however, we use k-fold CV to avoid overfitting problems). At the end we obtain a scores matrix of shape $[N \times \text{years}] \times [\#t]$, where N is the number of companies ($N = 2000$), years $= 10$ and $\#t =$ number of technology codes $= 643$; in total, this scores matrix has shape $20\,000 \times 643$.

Each column of the score matrix represents the likelihood that each company (rows) will patent each technology code (columns). We can then argue that two sectors are *similar* if the RF predicts that the same companies will (or will not) produce patents in these sectors. In this sense, the columns of the score matrix can be seen as the coordinates in a high-dimensional space for each technology code, where the number of dimensions is given by the number of companies multiplied by the number of training years (in this case, 20 000). In order to provide better interpretability to the relatedness assessment, one should find a low-dimensional visualization of this high-dimensional representation. Obviously, it is impossible to visualize this continuous space of technologies in such a high dimensionality; so we project these points in a lower dimension by combining a variational—autoencoder neural network (Kingma and Welling 2013), to reduce the dimension from 20 000 → 150, and then t-SNE (Van der Maaten and Hinton 2008), to reduce the embedding space from 150 → 2 dimensions, finally obtaining the CTS, that we show in figure 4(a). Now the similarity between technology codes is simply given by the relative distance in this $2 - D$ space, and it is easy to understand and visualize how firms move from the codes already present in their portfolios to the ones that are immediately close, as shown in figure 5.

Now we want to use the idea of a coherent diffusion in this low dimensional space to produce forecasts; in practice, to obtain a score matrix $\mathbf{S}^{2011}$ to compare with the possible activations of 2011. We start by computing a similarity matrix for the CTS, which for the sake of simplicity we keep calling $\mathbf{B}$. We use the distances between technology codes on the CTS and Gaussian kernels:

$$B_{i,j} = \frac{e^{-\|\mathbf{y}_i - \mathbf{y}_j\|^2/2\sigma_i^2}}{\sum_k e^{-\|\mathbf{y}_i - \mathbf{y}_k\|^2/2\sigma_i^2}},$$

where $\mathbf{y}_i$ is the coordinate of the ith technology code in the CTS (i.e. in the $2 - D$ space). The σ_i is the standard deviation of the Gaussian kernel related to the technology code ith; this parameter can be set differently for each i code, through a binary search process in which the number of first neighbors is fixed. As we can see in figure 4(a), there are codes in dense areas and codes in less dense areas, so the idea is to assign a high sigma value to the codes in less dense areas and low sigma values in more dense areas in order to keep the interaction with the number of first neighbors constant. The binary search process is described in the supplementary information where we also show that the best optimal value of nearest neighbors is 75.

After the similarity matrix $\mathbf{B}$ is obtained, one can compute the score matrix $\mathbf{S}^{2011}$ from the coherence equation equation (2):

$$S^{2011}_{f,t} = \sum_{t'} M^{2009}_{f,t'} B_{t',t}.$$

The number of nearest neighbors is calculated out of sample using the four-fold cross validation as in the case of the RF CV: we use 3/4 of the companies to determine the number of nearest neighbors that maximize the area under the PR curve and then we calculate the scores using the equation (2) for the remaining companies.

Note that the CTS is, as the network approaches, density-based: the more a firm surrounds a technology sector, the more likely it will be part of its patenting activity in the near future.

4.6. Benchmark models

In order to understand the effective goodness of our forecast results, a comparison with some relatively trivial benchmark models is required. We used two benchmark models:

- The first consists of simple randomization of the technology codes. In practice, we shuffle the columns of the $\mathbf{M}^{2009}$ matrix in the calculation of equation (2). The B used is that calculated with the Technology Space network starting from the not randomized $\mathbf{M}^{2009}$ (using the other networks there is no significant change in the metric scores). In this way, the diversification of firms is preserved.
- The second benchmark model checks the hypothesis that the simple temporal autocorrelation of the bipartite networks can explain the observed dynamics. In this case, we use the mean of the **V**s matrices from 2000 to 2009, $\overline{\mathbf{V}}$ (all the years used in the training) of the test firms as score matrix $\mathbf{S}^{2011}$, that is, element-wise:
$$\mathbf{S}^{2011} = \overline{\mathbf{V}}.$$
In this way, we check if the number of patents done in the past can forecast the number of patents done in the future by the same company in the same technology sector. As shown in figure 2, this benchmark model can outperform some of the density-based approaches.

4.7. Prediction performance metrics

In order to compare the goodness of the predictions of the different approaches, we use standard evaluation metrics, widely used for classification tasks in supervised machine learning (Hossin and Sulaiman 2015). As different metrics capture different aspects of the prediction problem, only the comparison between various measures of performance can provide a global view of the effectiveness of a forecasting approach.

The elements that we want to predict are the possible activations, that is, those elements of $\mathbf{M}^{2011}$ that were always zero in 2000–2009. The 0s are called negatives, and the 1s are called positives. The elements equal to 1 that are correctly predicted are called true positives (TP); and similarly one can define the false positives (FP), the true negatives (TN) and the false negatives (FN) as, respectively, the 0s predicted as 1s, the correctly predicted 0s, and the 1s predicted as 0s.

To evaluate the predictions done with the different approaches, we have used three performance metrics:

- **Area under the PR curve**: this indicator is equal to the area in the precision-recall plane. Precision is defined as $\frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau)+\mathrm{FP}(\tau)}$, while recall is equal to $\frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau)+\mathrm{FN}(\tau)}$. These quantities are close to 1 if FP and FN are respectively minimized. Note that in order to compute precision and recall one has to specify the scores' binarization threshold τ, that is, the number above which the score is associated with a positive prediction (1). The PR curve is defined by varying the τ parameter because for different values of τ we obtain different precision and recall values. The last step is the computation of the area under this curve, which is independent of the threshold τ.
- **Precision@100**: to compute this indicator we focus on the top 100 score elements in $\mathbf{S}^{2011}$: if the model is correct, many of these possible activations should become realized activations. The Precision@100 is the ratio between the number of these 100 that are true positives (that is, correctly predicted realized activations), and 100, i.e. the number of elements that we are considering. This represents a global assessment, that considers the score matrix as a whole.
- **mPrecision@10**: while the Precision@100 provides a global measure of the precision of the approach, we would like to have a measure of our average predictive performance for each firm. To do this, we evaluate the mPrecision@10. We consider the 10 highest scores for each row, i.e. for each firm, and compute the fraction of true positives. Then we average over the firms. Since most of the firms do not show at least 10 realized activations, the global number is far from 1. We have computed the mPrecision also restricting ourselves only to the firms with 10 or more realized activations, finding similar qualitative results.

Acknowledgments

We thank Lorenzo Napolitano for providing the patent data used for this work and Andrea Tacchella for insightful discussions. All the authors acknowledge the CREF project 'Complessità in Economia'.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Matteo Straccamore https://orcid.org/0000-0002-3351-2323
Andrea Zaccaria https://orcid.org/0000-0002-4478-3292

References

Albora G, Pietronero L, Tacchella A and Zaccaria A 2021 Product progression: a machine learning approach to forecasting industrial upgrading (arXiv:2105.15018)

Balassa B 1965 Trade liberalisation and 'revealed' comparative advantage *Manch. Sch.* **33** 99–123

Berry C H 1971 Corporate growth and diversification *J. Law Econ.* **14** 371–83

Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32

Breschi S, Lissoni F and Malerba F 2003 Knowledge-relatedness in firm technological diversification *Res. Pol.* **32** 69–87

Buccellato T 2016 The competences of firms are the backbone of economic complexity available at SSRN 2827468

Brummitt C D, Gómez-Liévano A, Hausmann R and Bonds M H 2020 Machine-learned patterns suggest that diversification drives economic development *J. R. Soc. Interface.* **17** 20190283

Cimini G, Carra A, Luca D and Zaccaria A 2022 Meta-validation of bipartite network projections *Commun. Phys.* **5** 76

Fall C J, Törcsvári A, Benzineb K and Karetka G 2003 Automated categorization in the international patent classification *Acm Sigir Forum* vol 37 (New York: ACM) pp 10–25

Gort M 1962 *Diversification and Integration in American Industry: A Study by the National Bureau of Economic Research* (Princeton, NJ: Princeton University Press)

Hall B H, Jaffe A B and Trajtenberg M 2001 The NBER patent citation data file: lessons, insights and methodological tools *Technical Report* National Bureau of Economic Research

Hidalgo C A, Klinger B, Barabási A-L and Hausmann R 2007 The product space conditions the development of nations *Science* **317** 482–7

Hidalgo C A *et al* 2018 The principle of relatedness *Int. Conf. Complex Systems* (Springer) pp 451–7

Hossin M and Sulaiman M N 2015 A review on evaluation metrics for data classification evaluations *Int. J. Data Mining Knowl. Process Manag.* **5** 1

Jaffe A B, Trajtenberg M and Fogarty M S 2000 Knowledge spillovers and patent citations: evidence from a survey of inventors *Am. Econ. Rev.* **90** 215–8

Joo S H and Kim Y 2010 Measuring relatedness between technological fields *Scientometrics* **83** 435–54

Kauffman S A 1996 *Investigations: The Nature of Autonomous Agents and the Worlds They Mutually Create* (Santa Fe Institute)

Kim S H, Jun B and Lee J-D 2021 *Technological Relatedness: How Do Firms Diversify Their Technology?*

Kingma D P and Welling M 2013 Auto-encoding variational Bayes (arXiv:1312.6114)

Kotsiantis S B *et al* 2007 Supervised machine learning: a review of classification techniques *Emerg. Artif. Intell. Appl. Comput. Eng.* **160** 3–24

Leten B, Belderbos R and Van Looy B 2007 Technological diversification, coherence, and performance of firms *J. Prod. Innovat. Manag.* **24** 567–79

Lo Turco A and Maggioni D 2016 On firms' product space evolution: the role of firm and local product relatedness *J. Econ. Geogr.* **16** 975–1006

Loreto V, Servedio V D P, Strogatz S H and Tria F 2016 Dynamics on expanding spaces: modeling the emergence of novelties *Creativity and Universality in Language* (Berlin: Springer) pp 59–83

Martínez C 2011 Patent families: when do different definitions really matter? *Scientometrics* **86** 39–63

Nesta L and Saviotti P-P 2006 Firm knowledge and market value in biotechnology *Ind. Corp. Change* **15** 625–52

OECD 2001 Publishing, organisation for economic co-operation, and development staff *OECD Science, Technology and Industry Scoreboard 2001: Towards a Knowledge-Based Economy* (Organisation for Economic Co-Operation and Development)

Pedregosa F *et al* 2011 Scikit-learn: machine learning in Python *J. Mach. Learn. Res.* **12** 2825–30

Penrose E 1959 *The Theory of the Growth of the Firm* (New York: Wiley)

Penrose E T 1960 The growth of the firm-A case study: the hercules powder company *Bus. Hist. Rev.* **34** 1–23

Pugliese E, Cimini G, Patelli A, Zaccaria A, Pietronero L and Gabrielli A 2019a Unfolding the innovation system for the development of countries: coevolution of science, technology and production *Sci. Rep.* **9** 16440

Pugliese E, Napolitano L, Zaccaria A and Pietronero L 2019b Coherent diversification in corporate technological portfolios *PLoS One* **14** e0223403

Rahmati P, Tafti A R, Westland J C and Hidalgo C 2020 When all products are digital: complexity and intangible value in the ecosystem of digitizing firms *Forthcom. MIS Q.*

Ribeiro S P, Menghinello S and De Backer K 2010 The OECD ORBIS database: responding to the need for firm-level micro-data in the OECD *OECD Statistics Working Papers*

Rigby D L 2015 Technological relatedness and knowledge space: entry and exit of US cities from patent classes *Reg. Stud.* **49** 1922–37

Rumelt R P 1974 *Strategy, Structure, and Economic Performance*

Rumelt R P 1982 Diversification strategy and profitability *Strat. Mgmt. J.* **3** 359–69

Rycroft R W and Kash D E 1999 *The Complexity Challenge: Technological Innovation for the 21st Century* (Burns & Oates)

Sbardella A, Pugliese E, Zaccaria A and Scaramozzino P 2018 The role of complex analysis in modelling economic growth *Entropy* **20** 883

Smith B and Linden G 2017 Two decades of recommender systems at Amazon.com *IEEE Int. Comput.* **21** 12–8

Strumsky D, Lobo J and Van der Leeuw S 2011 Measuring the relative importance of reusing, recombining and creating technologies in the process of invention *Technical Report, Working Paper*

Strumsky D, Lobo J and Van der Leeuw S 2012 Using patent technology codes to study technological change *Econ. Innovat. N. Technol.* **21** 267–86

Tacchella A, Cristelli M, Caldarelli G, Gabrielli A and Pietronero L 2012 A new metrics for countries' fitness and products' complexity *Sci. Rep.* **2** 723

Tacchella A, Mazzilli D and Pietronero L 2018 A dynamical systems approach to gross domestic product forecasting *Nat. Phys.* **14** 861–5

Tacchella A, Zaccaria A, Miccheli M and Pietronero L 2021 Relatedness in the era of machine learning (arXiv:2103.06017)

Teece D J, Rumelt R, Dosi G and Winter S 1994 Understanding corporate coherence *J. Econ. Behav. Organ.* **23** 1–30

Tria F, Loreto V, Servedio V D P and Strogatz S H 2014 The dynamics of correlated novelties *Sci. Rep.* **4** 5890

Van der Maaten L and Hinton G 2008 Visualizing data using t-SNE *J. Mach. Learn. Res.* **9**

Yan B and Luo J 2017 Filtering patent maps for visualization of diversification paths of inventors and organizations *J. Assoc. Inf. Sci. Technol.* **68** 1551–63

Youn H, Strumsky D, Bettencourt L M A and Lobo J 2015 Invention as a combinatorial process: evidence from us patents *J. R. Soc. Interface* **12** 20150272

Zaccaria A, Cristelli M, Tacchella A and Pietronero L 2014 How the taxonomy of products drives the economic development of countries *PLoS One* **9** e113770