Journal of Economic Behavior and Organization 234 (2025) 107011

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Predicting the technological complexity of global cities based on unsupervised and supervised machine learning methods

Federico Nutarelli [a], Samuel Edet [b], Giorgio Gnecco [a], Massimo Riccaboni [a,c,*]

[a] IMT School for Advanced Studies, Lucca, Italy
[b] International Finance Corporation, Washington D.C., USA
[c] University School for Advanced Studies IUSS, Pavia, Italy

ARTICLE INFO

JEL classification:
C53
C45
O33
O34
R58
O18

Keywords:
Innovation
Urban studies
Technological change
Artificial intelligence
Global cities

ABSTRACT

Analyzing and predicting innovation in global cities, i.e. cities with a high degree of economic integration into the world economy, can help identify emerging technologies and inform investment decisions that facilitate talent attraction and urban planning. In this context, the contribution of this paper is to analyze the technological complexity of global cities. We show how the combination of state-of-the-art network community detection and supervised machine learning can support local innovation and development policies by predicting the future competitiveness of global cities based on an up-to-date patent dataset. Network community detection with the Poisson stochastic block model is used as an unsupervised pre-processing step to find cities with similar innovation profiles and create homogeneous training sets that improve predictive power, interpretability and computational efficiency in a subsequent supervised learning task. The paper then compares the use of different supervised machine learning methods to predict the future competitiveness of global cities. Tree-based methods turn out to achieve better prediction performance than other supervised machine learning methods on various metrics based on the ground truth derived from historical patent production. The analytical method used in this paper can help policy makers identify technology sectors where global cities could focus their future investments and provide information on the temporal evolution of geographical patterns related to innovation.

1. Introduction

Since the publication of Michael Porter's "The Competitive Advantage of Nations" (Porter, 1990), interest in what cities need to be competitive has increased. With increasing competition between cities around the world for investment and talent, a large literature shows that their competitiveness is important for the cross-fertilization of ideas (Florida et al., 2018; Verginer and Riccaboni, 2021; Belderbos et al., 2022). Cities have long been recognized as the key drivers of innovation and prosperity (Bettencourt et al., 2007; Balland et al., 2020): according to Li et al. (2017), cities are home to over 50% of the world's population, more than 80% of the world's wealth and at least 90% of innovation. This is no surprise, as scientists and inventors live and work mainly in cities (Verginer and Riccaboni, 2020, 2021; Belderbos et al., 2022). This phenomenon is particularly evident in global cities (i.e. cities highly connected to the world economy), which attract a disproportionate amount of talent and investment (Bettencourt et al., 2007; Chakravarty et al., 2021; Du et al., 2022).

* IMT School for Advanced Studies, Lucca, Italy.
E-mail addresses: federico.nutarelli@imtlucca.it (F. Nutarelli), sedet@ifc.org (S. Edet), giorgio.gnecco@imtlucca.it (G. Gnecco), massimo.riccaboni@imtlucca.it (M. Riccaboni).

https://doi.org/10.1016/j.jebo.2025.107011
Received 15 February 2024; Received in revised form 4 March 2025; Accepted 1 April 2025
Available online 17 April 2025
0167-2681/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Innovation can be defined as the development of inventions that have gained broad social acceptance, generate increasing returns and influence territorial development (Arthur, 1994). As emphasized in the literature on innovation economics (see Dosi (2023) for a recent overview), technological progress has an evolutionary character in the sense that different efforts are made at each point in time to advance technology, competing with each other and with prevailing practices. While inventions do not necessarily originate in cities, the processes of innovation emergence and diffusion are closely intertwined with urban development (Raimbault and Pumain, 2023). Innovations tend to be geographically concentrated in their early stages before they are imitated and diffused on a larger scale (Audretsch and Feldman, 1996a). Innovation clusters indeed play an important role in local urban systems (Moreno et al., 2006), as spatial proximity facilitates synergies, knowledge spillovers and the sustainability of innovations (Frenken et al., 2007; Boschma and Iammarino, 2009). In recent decades, the scale of innovation processes has become more internationalized due to the activities of multinational companies (Rozenblat and Pumain, 2007) and a more global reach of networks (Edet et al., 2021).

Various definitions of urban competitiveness have been proposed in the literature, although no single definition has yet prevailed (see Kachniewska et al. (2018) for an overview and Martin and Simmie (2008) for a discussion of some approaches proposed in the literature to provide a theoretical basis for urban competitiveness). This lack of common understanding is due to several factors. First and foremost is the fact that modern urban landscapes challenge the antiquated concept of homogeneous, monolithic structures; instead, they present themselves as diverse, multi-centered metropolitan regions. Recently, there has been a significant paradigm shift from traditional, industrial chain systems to more dynamic, hardly predictable, and innovation-driven economic structures, as described by Kamiya et al. (2020). Second, the competitiveness of a city results from a complex network of social interactions that contribute to the economic vitality of the city. This includes a wide range of factors and innovations that not only shape but also drive the socio-economic fabric of an urban center (Harris, 2007; Kachniewska et al., 2018; Belderbos et al., 2022). There is also a lack of consensus on a global definition of cities and comparable statistics to measure the performance of global cities across borders (OECD, 2020).

In this paper, we contribute to the growing literature on regional and urban complexity (Balland et al., 2019, 2020; Fritz and Manduca, 2021; Mewes and Broekel, 2022; Straccamore et al., 2023; Buyukyazici et al., 2024) by developing a framework for analyzing the innovative capacities of cities. Our approach is based on the classical Ricardian model of comparative advantage (Ricardo, 1817), which is applied to a patent dataset to compare cities in terms of their innovative capacity. Such a model provides a theoretical basis for quantifying the competitiveness of a location (i.e. a city) in a specific field of activity (e.g. industry, technology or scientific domain) (Hidalgo, 2021). However, its typical application offers little guidance for predicting future competitiveness. To address this gap, we combine a Ricardian measure of technological competitiveness with Machine Learning (ML) to develop an approach for predicting the future competitiveness of global cities in specific technology sectors that takes into account the complex higher-order relationships underlying technological change.

In the Ricardian tradition, a city is considered competitive in a given sector if it produces more than the global share in that sector. This can be modeled using the Revealed Technological Advantage (**RTA**) matrix (Soete, 1987), which expresses the relative technological capabilities of cities in specific technological fields derived from an analysis of their overall patent activities.[1] More precisely, the RTA value $RTA_{ij} \in [0, +\infty)$ of a city i in technology j is the city's share of patents in that technology relative to the global share. If RTA_{ij} is greater than or equal to one, then the city is considered competitive in this technology. This means that the signal for the city's capabilities exceeds the baseline level and is significant enough for the city to be considered competitive.

Unraveling the linkage between the competitiveness of cities and their resource allocations and capabilities (Zhang et al., 2020) is crucial to understanding why some cities are more competitive, productive, or resilient than others. In the past literature, most analyses about the economic competitiveness of cities were geared towards finding explanatory variables for how countries, cities, or firms acquire and increase capabilities in different products, technologies, and scientific domains, in such a way that their production output exceeds their fair share (Sassen, 2001; Rigby, 2015; Balland et al., 2019). However, tracking or measuring such capabilities or strategic choices can be hard as these are typically intangible (Straccamore et al., 2022). In this context, the literature on economic complexity (Hausmann et al., 2014) provides useful insights on how to measure the current capabilities of cities or the capabilities they need to become competitive in a particular sector, scientific or technological field.[2]

Economic complexity models rely on a fine-grained representation of economic activities and learn how to find a combination of latent factors that explain economic activities in a location. This is based on the application of appropriate dimensionality reduction techniques (e.g. to export, patenting and scientific activities and capabilities observed in different locations) (Hidalgo, 2021; Balland et al., 2022; Buyukyazici et al., 2024). The dimensionality reduction techniques used in this context are usually based on matrix factorization (Hidalgo, 2021; Gnecco et al., 2022). Such techniques provide a concise way to summarize the economic complexity of different locations. Indices of economic complexity based on an analog of the **RTA** matrix have been developed in the literature (Hidalgo and Hausmann, 2009; Sciarra et al., 2020; Gnecco et al., 2022). They can be also applied to the case of the **RTA** matrix,[3] e.g. with the aim of investigating the evolution of technological complexity of cities over time.[4] Patent data have been

[1] In the context of competition between cities, the **RTA** matrix is the analog of the classical Revealed Comparative Advantage (**RCA**) matrix (Balassa, 1965), which expresses the relative advantages of countries in specific classes of goods or services derived from an analysis of their total trade flows.

[2] As is common in the literature (Tacchella et al., 2013), we assume here that more complex global cities are more competitive.

[3] Or, more precisely, to its binarized version, or incidence matrix.

[4] See Section 4 for related results. It is worth remarking that Lapatinas et al. (2022) recently applied one of these indices to study the economic complexity of cities. However, their dataset has a different nature than the one used in the present work (as the data matrices considered therein relate every city with its number of firms with global presence operating in each economic sector). Moreover, differently from the present work, that article is not focused on the application of state-of-the-art ML methods to the prediction of elements of the **RTA** matrix, and more in general of the technological complexity of cities.

used repeatedly in the economic complexity literature to analyze the competitiveness of different locations in the fields of science and innovation (Cimini et al., 2014; Pugliese et al., 2016; Morrison et al., 2017; Pugliese et al., 2019).

Models of economic complexity usually capture the current capabilities of cities. However, it is also important to predict the future technological competitiveness of cities. This can be achieved by using as a ground truth an **RTA** matrix for each time period in which a prediction is made.

Various approaches in network science attempt to address the problem of competitiveness prediction based on the degree of correlation between technologies (see, e.g. the so-called relatedness principle, as detailed by Hidalgo and Hausmann (2009) and Zaccaria et al. (2014)). In this context, our paper enhances the concept of "counterfactual (or implied) comparative advantage" by leveraging technological relatedness and advanced ML methods to predict and analyze cities' technological competitiveness. In the economic complexity framework, a city's patent production in a technological area is inferred from its productivity in related technologies and from the behavior of similar cities. More precisely, a standard Ricardian model, which assumes that the productivity parameter of a city in a technological field is a realization of a random variable associated with a suitable probability distribution, cannot explain the patterns of co-location of technological skills in cities or of the same innovations across cities. Indeed, a Ricardian model is consistent with actual observations only if it assumes that innovations differ in their technological relatedness and that cities tend to have similar productivity in technologically related products. Under these assumptions, a city will innovate with a similar intensity as cities with comparable patterns of comparative advantage. In our work, we build on these assumptions by leveraging the concept of technological relatedness using the **RTA** matrix, which reflects the technological competencies of cities. The incidence matrix derived from the **RTA** matrix captures the relatedness between technologies. We further develop this concept by utilizing a community detection algorithm to cluster cities in communities based on their technological competencies, ensuring that cities within the same community have similar productivity patterns in related technological fields. Adopting this clustering method is motivated by the assumption that cities with similar technological profiles will have similar productivity in related technologies. Finally, our approach extends the economic complexity approach by incorporating advanced supervised ML techniques to predict the future competitiveness of global cities in various technologies, by using models trained on historical incidence matrices, which are then validated against actual data from five years later. This temporal validation ensures the robustness of our predictions over time, making them reliable for future forecasts. Additionally, by computing various indices of the technological complexity of global cities, we provide a clear and actionable ranking of global cities' technological competitiveness. This ranking can inform policy makers about strategic investment areas, thus extending the theoretical insights into practical applications.

It is worth remarking that, in the present work, we do not rely merely on the amount of correlation between any two different technologies to estimate the density of competitive structures of cities in different technology sectors. Such correlation can be estimated following, e.g. the product space approach (Hidalgo and Hausmann, 2009) or the taxonomy network approach (Zaccaria et al., 2014). The idea is that correlations within the ecosystem of technologies provide information on which technology serves as a gateway to patent in other technologies.[5] One point of criticism of this method of predicting competitiveness is that the assumed correlation structure of the technologies does not consider the higher-order relationships between the technologies.[6] ML methods, as part of Artificial Intelligence (AI) techniques, can be used to take more complex relationships into account (Tacchella et al., 2023). By using ML, the predicted future RTA value of a city in a given technology can be modeled, for example, as a non-linear function of appropriate features (e.g. the observed current RTA values of comparable cities in the same or analogous technologies). To this end, we apply ML to economic complexity by using a combination of unsupervised and supervised ML methods to predict the future technological capabilities of cities or the technological capabilities required to compete in a given technological field. Policy implications of such predictions include identifying emerging technologies to make investment decisions that facilitate talent attraction and retention by activating the technological edge in specific sectors; improving urban planning and infrastructure; fostering collaboration with other selected cities to increase the competitiveness of a specific city; directing migration flows to cities identified as having untapped potential, as this could accelerate their progress towards innovation and diversification; gaining meaningful information on the evolution over time of geographical patterns related to innovation (e.g. focusing on trajectories of suitable points representing barycenters of groups of global cities, weighted by their different innovation capabilities).

To summarize, the most important contributions of this work can be described as follows. The first contribution is the analysis of the technological complexity of global cities. The analysis of the technological complexity of cities has so far been limited to specific countries (Straccamore et al., 2022) due to the lack of comparable statistics at the global level. In this paper, we provide a first insight into the evolution of international competitiveness of global cities in different technology areas. In this sense, the study also contributes to the evolving literature on the geography of innovation by analyzing the spatial trajectories of innovation centers over time using a centroid-based approach. Similar methods were explored by Maggioni and Uberti (2009), where the authors examined shifts in innovation centers across Europe, providing a basis for the spatial econometric analysis of innovation clusters,

[5] Combining the correlation matrix of technologies alongside the current competitive structure of cities can be used to estimate the density of cities' competitive structure around different technological areas. The lower the density of a city's capabilities around a technology, the less likely the city will upgrade to or activate that technology. In this case, the technologies the city is currently competitive in are likely distant from the given technological area. Then, one can evaluate the accuracy of the estimated density against what one observes to be the actual density of the competitive structure by using different binary score measures.

[6] For example, patenting an electric car might require capacity in electronics and mechanics, which implies the existence of a relationship among these three entities (jointly). Evaluating this relationship simply by means of pairwise correlations means that the former would be represented, e.g. by the sum of the correlation between production of electric cars and amount of expertise in electronics, and the correlation between production of electric cars and amount of expertise in mechanics. Hence, the presence of a large enough correlation of either pair alone could lead to acknowledge the presence of a relationship that in reality does not exist. This implies that describing the competitiveness development path of a city through a sum of correlations is an oversimplification.

and by Sciarra et al. (2020), where trajectories of centroids of countries weighted by a generalized index of economic complexity were analyzed.

A second contribution is the prediction of the future competitiveness of global cities in different technologies based on ML methods. Two important policy implications of our methodology are the possible identification of specific technology sectors on which each global city could effectively focus its future investments and the identification of other global cities with innovation patterns that they should emulate to increase their own competitiveness.

The third contribution is the enhancement of the employed supervised ML algorithms' performance through a suitable pre-processing based on clustering (or network community detection). This unsupervised pre-processing aims to facilitate the application of such algorithms by pre-selecting similar cities in the training set for the city under analysis. By training supervised ML models only on relevant clusters, we aim to focus the predictions on comparable entities, thus avoiding irrelevant noise and improving the predictive performance of these models. It is worth remarking that, while data clustering is commonly used in various fields to study diversification,[7] as already mentioned, our focus is specifically on enhancing the predictive power of supervised ML methods.[8] Finally, a majority voting matrix is introduced in our pre-processing to maintain the temporal consistency of community structures over several years and thus make our methodology more robust.

The article is structured as follows. Section 2 defines revealed technological advantage as a measure of the competitiveness of global cities and describes the data we use. Section 3 presents our methodological approach, starting with the pre-processing of network community detection (Section 3.1), followed by the solution of a sequential supervised ML task (Section 3.2), for which a description of the supervised ML method that turns out to be the best performer for different metrics is provided (Section 3.3). Section 4 describes the results of our analysis, focusing first on the results of the supervised ML methods (Section 4.1), then on the technological complexity results (Section 4.2), on their policy implications (Section 4.3), and on their interpretability (Section 4.4). Section 5 concludes the paper with a discussion of possible future developments.

Details on the construction of the dataset used for our analysis of innovation in global cities are provided in Appendix A of the Online Supplementary Material, whereas Appendices B and C of the Online Supplementary Material describe, respectively, the network community detection method and the other supervised ML methods used in this work. Appendix D of the Online Supplementary Material presents the results of a robustness check with two alternative pre-processing methods and no pre-processing. Appendix E of the Online Supplementary Material deals with the Economic Complexity Index (ECI) and the GENeralized Economic comPlexity (GENEPY) index. A comparison of our results with the Global City Power Index (GCPI) is provided in Appendix F of the Online Supplementary Material. Appendix G of the Online Supplementary Material presents additional results on the trajectories of centroids of sets of global cities. Finally, Appendix H of the Online Supplementary Material provides a feasibility and attractiveness analysis of technological investments.

2. Data

In our analysis, we look at the technological competitiveness of 150 global cities selected from the results of the Globalization and World Cities (GaWC) research project by Beaverstock et al. (1999). It is worth remarking that the term "global" here refers to the degree of economic integration of a city with the global economy. This does not necessarily imply that the selected cities cover uniformly the globe. In more detail, these cities were classified by the GaWC project as either *alpha* or *beta* cities based on their economic connectivity to the global economy. More specifically, *alpha* cities are connected to several major economic regions, while *beta* cities connect only moderate economic regions to the global economy. The selected cities are spread across Africa (6), Asia (45), Europe (48), North America (34), Oceania (7) and South America (10). The analysis is conducted at the level of 638 classes of the 4-digit "International Patent Classification" (IPC), which classifies patents filed by inventors in such cities according to the technological fields to which they relate.[9] A total of 6,242,439 patents filed by inventors in these cities across these 638 technological fields in the period 2000–2014 are considered in this study. Details on the construction of the patent dataset containing the association of each global city with every IPC are provided in the subsection "Construction of the Dataset" of Appendix A of the Online Supplementary Material, with that subsection based on Edet (2022). The names of all the global cities considered in the present study are reported in Appendix A of the Online Supplementary Material.

Our analysis of the technological competitiveness of global cities is based on the measurement of the Revealed Technological Advantage (RTA), which we derive from patent production. In the following, we provide the expression of the Revealed Technological Advantage RTA_{ij}^{t} of a city i with respect to an IPC j in a given year t. It measures the share of patenting activity of city i in the IPC j compared to the global average patenting activity in the same IPC.[10] To account for partial ownership of patents attributed to

[7] For instance, regional studies often apply clustering procedures before econometric analyses to define technological relatedness and diversification, as seen in works like Balland et al. (2019).

[8] More specifically, although diversification processes in economics can be either technological, regional, or network-based, our emphasis is on the performance improvement achieved through our pre-processing approach rather than on the diversification processes themselves. The existing literature can be categorized into several strands: studies that utilize community detection outputs alongside other similarity metrics to boost accuracy (Kerkache et al., 2022), those employing supervised ML to enhance community detection algorithms (Aljalbout et al., 2018; Filippone et al., 2008), and research leveraging community detection for link prediction, vertex classification (Tu et al., 2018), and network reconstruction (Guimerà and Sales-Pardo, 2009). Nevertheless, there is a lack of work explicitly addressing network community detection to improve the predictive power of supervised ML in forecasting competitiveness through feature engineering.

[9] The list of all the IPC codes considered in the present study is reported in Appendix A of the Online Supplementary Material. Their meanings are available at the following hyperlink: https://ipcpub.wipo.int/.

[10] To simplify the notation, we omit the dependence on t in this paragraph.

cities for a number of IPCs, we use a fractional count of patents (at the patent family level),[11] i.e. given a patent $p \in P$ (where P is the set of all patents) instantiated with the n_p IPCs from the set $J_p = \{j_{p_1}, j_{p_2}, \dots, j_{p_{n_p}}\}$ and done by inventors from cities in the set $I_p = \{i_{p_1}, i_{p_2}, \dots, i_{p_{m_p}}\}$, we consider the following quantities. The fractional count of patents to any given city i is $\sum_{p \in P: i \in I_p} 1/m_p$, the fractional count of patents to any given IPC j is $\sum_{p \in P: j \in J_p} 1/n_p$, and the fractional count of patents produced by city i in a given IPC j is $N_{ij} = \sum_{p \in P: i \in I_p, j \in J_p} \frac{1}{m_p n_p}$. Then, the RTA of city i in the IPC j is computed as:

$$RTA_{ij} = \frac{\frac{N_{ij}}{\sum_{j'} N_{ij'}}}{\frac{\sum_{i'} N_{i'j}}{\sum_{i'} \sum_{j'} N_{i'j'}}}. \tag{1}$$

The matrix $\mathbf{RTA} \in \mathbb{R}^{150 \times 638}$, whose entries are the RTA_{ij}, is used to construct an incidence (or competitiveness) matrix $\mathbf{M} \in \{0,1\}^{150 \times 638}$, whose elements are binarized RTA values. Specifically, if RTA_{ij} is greater than or equal to 1, the city i has specialized positively in the technology j. In this case, one sets $M_{ij} = 1$. Otherwise, one sets $M_{ij} = 0$. Descriptive statistics of the incidence matrix $\mathbf{M}$ derived from the patent dataset by Edet (2022) are reported in Appendix A of the Online Supplementary Material.

3. Methodology

In this section, we present a methodological approach that allows us to predict the evolution of revealed technological advantage and economic complexity of global cities based on network analysis and ML methods. In our approach, we combine a pre-processing step, in which we identify city communities using a state-of-the-art network community detection method (the Poisson Stochastic Block Model, or PSBM, see Karrer and Newman (2011)) with the subsequent application of supervised ML. In the following, first, we detail our pre-processing of the dataset, describing the procedure we apply for identifying a set of cities similar to the cities for which we want to predict future binarized RTA values (i.e. values of the incidence matrix, which provide information about the comparison of the RTA values with the threshold 1). We then provide a formal description of our specific supervised ML task and of the supervised ML method that turns out to be the best for our application according to various metrics, namely, Random Forest (RF). Finally, in Appendix B of the Online Supplementary Material, we provide some technical details about the specific network community detection method employed for the pre-processing step, whereas Appendix C of the Online Supplementary Material provides a brief overview of the other supervised ML methods we apply to solve the supervised ML task.

3.1. Pre-processing

A pre-processing procedure is adopted to identify cities that are similar to any given city i. The idea is to employ similar cities in the training set of a successive supervised ML task in order to boost the predictive power of supervised ML methods by avoiding exploiting information coming from less similar cities, which is expected to be less important from a predictive perspective. In this way, it is also possible to improve interpretability (as only information coming from a subset of cities is used to generate supervised ML predictions) and reduce the computational effort needed for training each supervised learning machine.

The core part of the adopted pre-processing step consists of the formation of a collection of K communities of cities according to the so-called Poisson Stochastic Block Model or PBSM (Karrer and Newman, 2011), which is a variant of the Stochastic Block Model or SBM (Holland et al., 1983; Nowicki and Snijders, 2001). The choice of this clustering method over more traditional alternatives, such as Louvain clustering (Blondel et al., 2008; Traag et al., 2019), is due to the latter's possible lack of statistical robustness. Louvain clustering, for instance, has some limitations (Peixoto, 2018, 2023).[12] Indeed, as discussed by Peixoto (2018, 2023), such a method typically finds high-scoring partitions in networks sampled from its own null model. This issue occurs because the method does not consider the deviation from the null model in a statistically consistent manner.

Both the SBM and the PSBM are forms of Bayesian clustering whose aim is to cluster nodes based on the posterior probabilities of the cluster assignments. The latter are updated at every step of the respective clustering algorithm until convergence. The key probability involved is the posterior probability of each data point belonging to each possible cluster, given the observed data and the model parameters. Details about the SBM and the PSBM are reported in Appendix B of the Online Supplementary Material. In the following, we describe the application of the PSBM to our scenario and motivate its use (in place of the SBM).

In our application, we first construct a graph of cities by multiplying the incidence matrix $\mathbf{M}^t$ by its transpose $(\mathbf{M}^t)^\top$. The generic element in position (i,j) of the resulting (weighted) adjacency matrix, denoted by $\mathbf{N}^t := \mathbf{M}^t(\mathbf{M}^t)^\top$, represents the sum of the number of IPC technological areas in which city i and city j have a competitive technology in common (giving more weight to the IPC technological areas in which they both patent more intensively). The co-occurrence of IPCs in which a pair of cities have a comparative advantage is assumed to be distributed as a Binomial. However, given the large number of possible cases, equal to the number of IPCs (638), we can approximate this distribution as a Poisson with a success rate of $\frac{50}{638}$, where 50 is the average co-occurrence of IPCs with a comparative advantage for all cities and all years.[13] For this reason, we choose to apply the PSBM as a more suitable variant of the SBM.

[11] A patent family is a set of patent applications related to the same or similar technical content.

[12] See also the following hyperlink for a discussion about this issue: https://skewed.de/tiago/posts/modularity-harmful/.

[13] The results of a simulation of the Binomial distribution and the Poisson distribution with a success rate of 50/638 are available upon request.

For every $t \in \{2000, \dots, 2008\}$, the PSBM is applied to the observed adjacency matrix for that specific year t, i.e. $\mathbf{N}^t$,[14] to cluster global cities. The reason we focus on the years up to 2008 is that these years are then considered to construct the predictors in the training set for our successive prediction task detailed in Section 3.2, in which input data from each year $t \in \{2000, \dots, 2008\}$ are used to predict data in the year $t+5$ (with the final 5-year period being 2008–2013). To avoid too small communities with less than ten cities, we merge small communities with the closest community they would join if the PSBM were configured to produce fewer clusters. This reassignment process is justified by the need to identify a subset of most similar cities to each city i. More specifically, the reassignment procedure adopted is detailed as follows[15]:

(i) For any community c_k^t with cardinality 1, isolate the city i that belongs to c_k^t (i.e. for which $Z_{ik}^t = 1$, being $\mathbf{Z}^t$ a node membership matrix representing the assignment of nodes to communities).
(ii) Identify the community c_j^t (with $j \neq k$) that city i would belong to if the number of blocks K were reduced by one. This is based on maximizing the posterior probability of community assignment given the adjacency matrix $\mathbf{N}^t$ (which represents the observed connections between nodes or cities), and a block matrix $\mathbf{C}^t$ (which represents edge probabilities between groups).
(iii) To merge small communities c_k^t having a number of nodes from 2 to 9 into the "closest" neighboring community, use the probabilities from the PSBM that express how likely each community is to connect with others. Specifically, after running the PSBM, one gets the block matrix $\mathbf{C}^t$ showing the connection strength or likelihood between every pair of communities. For each small community c_k^t, one looks at its connection strengths in this matrix to find the neighboring community c_j^t with the highest connection probability (the highest entry in the row of $\mathbf{C}^t$ associated with the small community c_k^t). This community is considered the "closest" in terms of network structure, as it has the strongest links with the small community according to the model. Then, one merges the small community c_k^t with the closest one c_j^t, thus ensuring that the merged community remains closely aligned with the PSBM's inferred network structure.
(iv) Continue this reassignment process (updating each time the dimension and the estimate of the $\mathbf{C}^t$ matrix, given that every merge reduces the number of communities) until each community satisfies a minimum size of 10 (i.e. its cardinality is at least 10).

After assigning the cities to the clusters for every $t \in \{2000, \dots, 2008\}$ using the PSBM, we perform a majority voting step by counting how often two cities belong to the same cluster.

More formally, denoting with n_c the number of cities, we define the Majority Voting matrix $\mathbf{MV}$ as an $n_c \times n_c$ matrix where each entry MV_{ij} represents the number of times city i and city j belong to the same cluster in one of the partitions $\mathbf{Z}^t$ inferred by the PSBM algorithm, independent of the specific year t. More in detail, we define the elements of the majority voting matrix $\mathbf{MV}$ as:

$$MV_{ij} = \sum_{t \in \{2000,\dots,2008\}} \sum_{k=1}^{K} Z_{ik}^t Z_{jk}^t, \tag{2}$$

where $Z_{ik}^t = 1$ if city i belongs to group k in year t, and 0 otherwise. Therefore, MV_{ij} counts the number of times that cities i and j belong to the same cluster in all the partitions.

We finally choose the first 50 most similar cities to every city i,[16] according to the majority voting matrix $\mathbf{MV}$.[17] These are used to construct the input to the successive supervised ML task (as detailed in the next subsection). For the sake of clarity, a pipeline of the main pre-processing steps described above is reported in Fig. 1.

Fig. 2 illustrates the results of a similar analysis in which the PSBM is applied to the sum of all the adjacency matrices in all the years, obtaining similar communities. This additional analysis is performed to provide the reader with a unique and more comprehensive representation of the detected communities.[18]

As a robustness check, we repeat the analysis using other pre-processing methods. The reader is referred to Appendix D of the Online Supplementary Material for further details.

[14] Following Lee and Wilkinson (2019)'s notation for the PSBM, such a matrix corresponds to the matrix $\mathbf{Y}$ in Appendix B of the Online Supplementary Material, to which the reader is referred for further information.

[15] The reader is referred to Appendix B of the Online Supplementary Material for more details about the node membership matrix $\mathbf{Z}^t$, the block matrix $\mathbf{C}^t$, and the expression of the posterior probability of membership assignment.

[16] The set of the 50 cities most similar to a given city i is replaced by the set of cities j for which $MV_{ij} > 0$ when the latter set has cardinality smaller than 50. In our case, this occurs only for a few cities i. The same modification is made for the supervised learning task reported successively in Section 3.2.

[17] By choosing (whenever possible) the top 50 cities, we generally ensure a consistent training set size, which is beneficial for model stability and simplifies the training process in our application. For instance, in this way, when applying Matrix Completion (MC), it turns out that the matrices to be completed have in most cases the same dimension. This not only makes the calculations easier, but is also fairer, as the choice of search range for the regularization parameter λ can depend on the matrix dimensions. While an alternative approach would be to use a threshold to filter similar cities (e.g. for each city i, to select only the cities j with $MV_{ij} \geq 7$ or another value), this could result in a more variable number of similar cities for each target city, complicating training (e.g. one would end up with inputs of different dimension in the various training sets).

[18] It is worth mentioning that, in our application, it turns out that the communities identified based on the first approach do not change so much from one year to the successive year, which further justifies this second approach in which network community detection is performed once for all the years. Still, the first approach has, in principle, the advantage of detecting changes in time in the communities found.

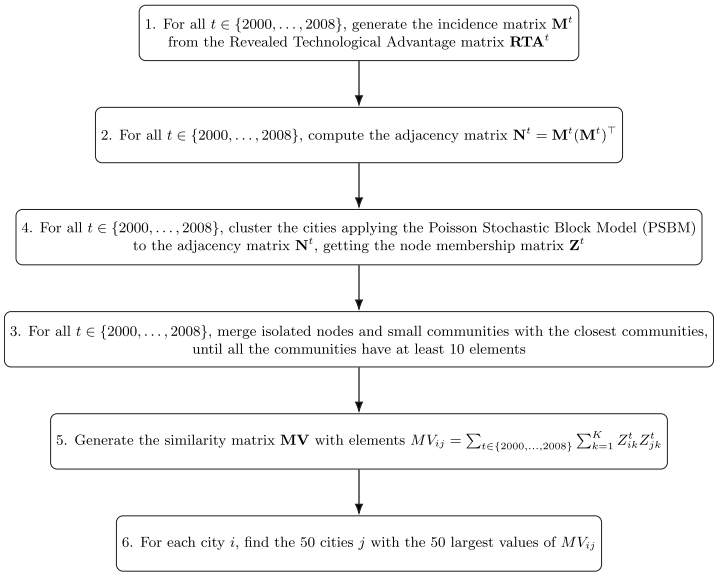


Fig. 1. Pipeline of the main pre-processing steps.

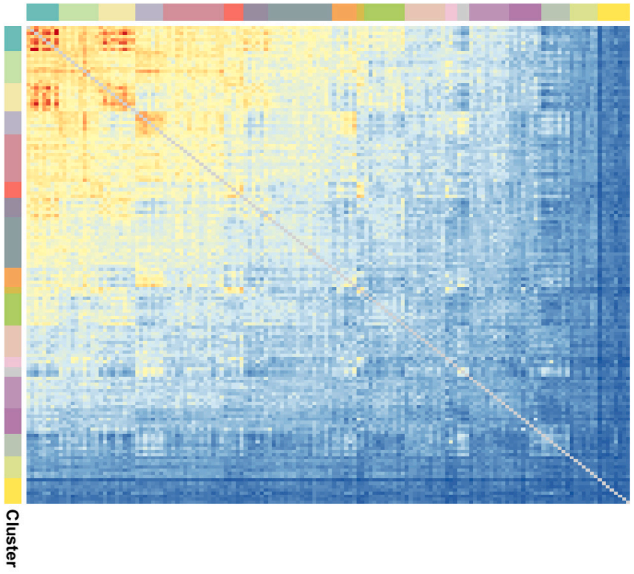


Fig. 2. Community partition obtained for a matrix that is the sum of all adjacency matrices $\mathbf{N}^t$ in all years $t = 2000, \ldots, 2008$. In the figure, the 19 clusters identified by the PSBM are represented by colored labels along the rows and columns that visually identify the cluster assigned to each city. The colors in the figure represent different connection strengths between pairs of cities (with city identifiers ordered by cluster affiliation), with lighter colors indicating weaker connections or lower similarity and darker colors indicating stronger connections or higher similarity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. Application of supervised machine learning

The pre-processing step detailed in Section 3.1 is used to generate a dataset associated with each specific city i, made by 50-dimensional input feature vectors related to its 50 most similar cities $\hat{\imath}$, and corresponding labels. By using the majority voting matrix, the set of these 50 most similar cities does not change with time.

In more detail, after the pre-processing phase, we consider, for each of the 150 global cities, a supervised machine learning problem. Its goal is to predict whether the given global city (say city i) will have a Revealed Technological Advantage (RTA) in a specific technology j five years into the future, in year $t + 5$. Specifically, the label M_{ij}^{t+5} equals 1 if $RTA_{ij}^{t+5} \geq 1$, indicating the presence of an RTA, and 0 otherwise. For each city i, the prediction is made based on an input feature vector that represents the

current RTA status of each of the 50 cities most similar to i in year t, for the same technology j. For every similar city $\hat{i}$, the feature vector includes a component that is 1 if $RTA_{\hat{i}j}^{t} \geq 1$, indicating the presence of an RTA, and 0 otherwise.[19] By varying the technology j, a total of 638 feature vectors (and corresponding labels) is obtained for each city i and pairs of years $(t, t+5)$. The length of the forecast horizon is chosen to be equal to five years, as the competence-enhancing strategies, R&D investments, and patenting activities take time to translate into higher technological competitiveness.[20]

In this work, various supervised ML methods – specifically, Logistic Regression (LR), Feedforward Neural Network (FNN), Support Vector Machine (SVM), Random Forest (RF), Bayesian Additive Regression Trees (BART), and eXtreme Gradient Boosting (XGB)[21] – are used to solve the prediction task described above. Moreover, another supervised ML method – namely, Matrix Completion (MC) –, is applied to solve a slight variation of the task (as this method requires a portion of a matrix as input, in place of a vector input). Details about the best-performing supervised ML method according to various metrics are in the next subsection, whereas the other supervised ML methods employed in our study are described in Appendix C of the Online Supplementary Material. Apart from the different sets of parameters and hyperparameters (detailed in Appendix C of the Online Supplementary Material), in this work, models for the various supervised ML methods listed above are trained/validated/tested according to the following common procedure.[22]

1. *Data Preparation*

- For every city i and each five-year period $(t, t+5)$ for $t = 2000, \ldots, 2008$, we construct the input design matrix[23] $\mathbf{X}^{t}_{\text{train,validation}} \in \mathbb{R}^{638\times 50}$ from the binarized RTA_{ij}^{t} values of the 50 most similar cities to city i (here, the dependence of the input design matrix on i is omitted from the notation, for simplicity).[24] Each row of this matrix corresponds to one observation, and refers to a specific choice of the IPC j (for $j = 1, \ldots, 638$). In other words, this row is made of a subset of elements (corresponding to the 50 most similar cities to city i) of the jth column of the incidence matrix $\mathbf{M}^{t}$.
- For every city i, IPC j, and year t, the value of the target variable $y^{t}_{j,\text{train,validation}}$ is the binarized RTA_{ij}^{t+5} value for city i and IPC j in the year $t+5$ (here, the dependence of the value of the target variable on i is omitted from the notation, for simplicity).

2. *Hyperparameter Tuning through Cross-Validation*

- For each supervised ML method, every city i, and each five-year period $(t, t+5)$, for $t = 2000, \ldots, 2008$, a grid search is conducted to find the optimal values of the hyperparameters. The F1-score is used as the scoring metric for 5-fold cross-validation (other metrics computed to assess each model's predictive performance are the Precision Recall (PR)-Area Under the Curve (AUC), Matthew's correlation coefficient, the Receiver Operating Characteristic (ROC)-AUC, and the average classification error).[25] The number of folds is selected in such a way as to get a reduced computation burden associated with cross-validation (Hastie et al., 2009).
- The best hyperparameters found by 5-fold cross-validation for each supervised ML method, city i, and period $(t, t+5)$ are recorded, ensuring the model's robustness and generalization ability. It is worth mentioning that, on the one hand, for this step and for the next step 3, to evaluate the different metrics mentioned above for each supervised ML model, we exclude from the validation/test set all pairs (city-IPC) for which there are no patents in our available dataset in the validation/test year.[26] On the other hand, such pairs are not excluded from each training set (with the associated RTA values set to 0) to avoid that the dimension of each input feature vector is equal to the (variable) number of its observed entries, which would complicate the application of the supervised ML methods used.

[19] This binarization in both the input feature vector and the target variable is done to simplify the learning task, thus facilitating the application of supervised ML.

[20] This choice of the forecast horizon is also motivated by the length of the period (2000–2014). Choosing a larger prediction horizon would reduce the size of the training sample, which would have a negative impact on the generalization ability of the trained supervised ML models. For example, increasing the prediction horizon to 10 years would reduce the size of the training sample by about one half.

[21] These methods are listed in increasing chronological order of development in the literature.

[22] The case of Matrix Completion (MC), which solves a similar supervised ML task based on a different type of input (a portion of a matrix), is described separately at the end of Appendix C of the Online Supplementary Material.

[23] For the supervised ML methods based on an input feature vector, we do not include data related to the focal city i in the input design matrix $\mathbf{X}^{t}_{\text{train,validation}} \in \mathbb{R}^{638\times 50}$ associated with its 50 closest cities. This decision is made to ensure that the predictions for each focal city are based solely on the information derived from other, similar cities. Including information related to the focal city's own historical RTA values in the training set could introduce path dependence, which could artificially improve each model's predictive accuracy for that city, as it would benefit from a direct reference to its own historical trends. Excluding the focal city ensures that the predictive framework remains unbiased. If we included the focal city's historical data, there would be the risk of overfitting each model to that city's specific characteristics, reducing its ability to generalize well to other cities. In summary, if a model were allowed to learn from the city's own historical data, it might simply memorize those trends rather than learning general patterns that could be applied to other cities. It is worth mentioning that, in this work, there are two exceptions to this rule (which provide a way to assess the impact of data related to the focal city): the case of the "Constant Estimator" (CE) introduced in Section 4.1, whose prediction is based only on data related to the city of interest i (so, by its definition, data related to other similar cities cannot be used by that method); the case of Matrix Completion (MC), for which partial information related to the matrix row associated with the focal city is necessary to estimate in a nontrivial way the remaining portion of that row, based on the elements in the rows associated with the other 50 most similar cities.

[24] The notation $\mathbf{X}^{t}_{\text{train,validation}}$ and the successive one $\mathbf{y}^{t}_{\text{train,validation}}$ refer to the fact that, for the years $t = 2000, \ldots, 2008$, the data for training and validating each model are extracted from this matrix and vector, respectively.

[25] For details on the definitions of the various performance metrics, the reader is referred to the final part of Appendix C of the Online Supplementary Material and to Athey and Imbens (2019) and Chicco and Jurman (2020).

[26] In our dataset, the percentage of such pairs depends on the city and the year and averages around 75%.

3. *Final Model*

- For each supervised ML method, the optimal hyperparameters are averaged across all cities i and periods $(t, t+5)$ to derive the average best hyperparameters.
- With these average best hyperparameters, for each supervised ML method and each city, a final ML model is trained and tested on data in the five-year period (2009, 2014), providing insights into the technological competitiveness of cities over time. In more detail, for each city i and the five-year period (2009, 2014), we construct the input design matrix $\mathbf{X}^{2009}_{\text{train,test}} \in \mathbb{R}^{638\times 50}$ from the binarized RTA^{t}_{ij} values of the 50 most similar cities to city i.[27] Similarly, for each city i and IPC j, the value of the target variable $y^{2009}_{j,\text{train,test}}$ is the binarized RTA^{2014}_{ij} value for city i and IPC j in 2014. In the final model, 5-fold cross-validation is performed only to evaluate the average performance of the model after its training (not to select its hyperparameters, which are fixed to their average best values).

Since our original dataset is imbalanced – on average 30% of the elements of the incidence matrices are 1 and the remaining 0 – we balance the classes for training the supervised ML models. This is done by employing the Synthetic Minority Over-sampling Technique for Nominal features (SMOTE-N) procedure (Chawla et al., 2002) and by cross-validating the hyperparameters by carefully managing class imbalance throughout the process. Specifically, for each fold of the cross-validation process, SMOTE-N is applied to the associated training set[28] to create a balanced 50%/50% distribution between the two classes, allowing the model to be trained on a dataset without majority-class bias. However, the validation set within each fold retains the original class imbalance, thus providing an unbiased performance evaluation that reflects the real-world class distribution. Additionally, when performing cross-validation, we use repeated stratified sampling to generate the folds. This approach ensures that each training and validation fold maintains the original class distribution of the dataset (approximately 68% of the majority class and 32% of the minority class), reflecting the true class proportions before SMOTE-N is applied to the training data. By maintaining this natural distribution in each fold, repeated stratified sampling allows us to fairly assess each model's performance on validation data that is not artificially balanced, as recommended in the literature (Fontanari et al., 2022).[29]

3.3. Best performing supervised machine learning method: Random Forest (RF)

For illustrative purposes, in this section we detail the application of a representative supervised ML method, among the ones considered in this work.

Among the various supervised ML methods we apply in combination with network community detection, the best predictive performance (according to various metrics, see Section 4.1) turns out to be obtained by the Random Forest (RF) (Breiman, 2001), which is an ensemble supervised ML method that combines predictions coming from a suitably-constructed random set of trees. A possible explanation for the superior performance of RF for our prediction problem comes from the sparsity of the feature vectors of the specific classification task (their components, being binarized RTA values, can assume only the values 0 and 1), and from the fact that the RF method is well-suited to handle sparse data. Moreover, according to the work by Athey and Imbens (2019), which focuses on the application of ML methods to economics, RF is particularly effective in settings in which a large number of input features (to be identified) are not related to the target variable. Indeed, in this case, the splits performed by RF typically ignore such predictors. As a consequence, the performance of RF remains strong even in the presence of a large number of such redundant features. Additionally, by identifying the most similar cities to a given city, the network community detection pre-processing step applied in our analysis allows RF to focus only on a subset of relevant features, by reducing the dimension of the input feature vectors. Moreover, pre-processing by network community detection makes training of RF models faster.

The following are some other possible explanations, more related to the specific application investigated in this study. Economic complexity often relies on non-linear relationships and interactions among various actors (Feng et al., 2022). RFs naturally handle these complexities without requiring extensive manual feature engineering. This makes them suitable for modeling the intricate dynamics of technological innovation and competitive advantage (Chu and Qureshi, 2023; Goulet Coulombe, 2024). Specifically, technological forecasting benefits from understanding the complex interactions within innovation networks. The incidence matrix captures these interactions, and tree-based ensemble methods (such as RF) that can manage such complexity are particularly effective. This approach aligns with the broader literature on technological forecasting from a complex systems perspective (Feng et al., 2022). Finally, economic and technological landscapes are dynamic, often undergoing structural changes. RFs can adapt to these changes better than some other supervised ML methods, ensuring consistent performance even when the underlying data-generating processes evolve in time (Goulet Coulombe, 2024).

In the following, we provide a brief overview of the RF model. For more technical details, the reader is referred to Breiman (2001). The RF model is an ensemble ML approach that uses bootstrapping and aggregation to train many decision trees simultaneously.

[27] The notation $\mathbf{X}^{2009}_{\text{train,test}}$ and the successive one $\mathbf{y}^{2009}_{\text{train,test}}$ refer to the fact that, for the year 2009, the data for training and testing each model are extracted from this matrix and vector, respectively.

[28] According to He and Garcia (2009), procedures like SMOTE should not be applied to the validation and test data, as they would artificially inflate the performance of a model by introducing synthetic data. Instead, the test and validation sets should retain the original class imbalance. This would allow the model to be evaluated realistically on data that reflect the true class distribution.

[29] Our code implementing SMOTE-N, repeated stratified sampling, and cross-validation is available in the following GitHub repository: https://github.com/federiconuta/SMOTE_RF_smart_cities.git.

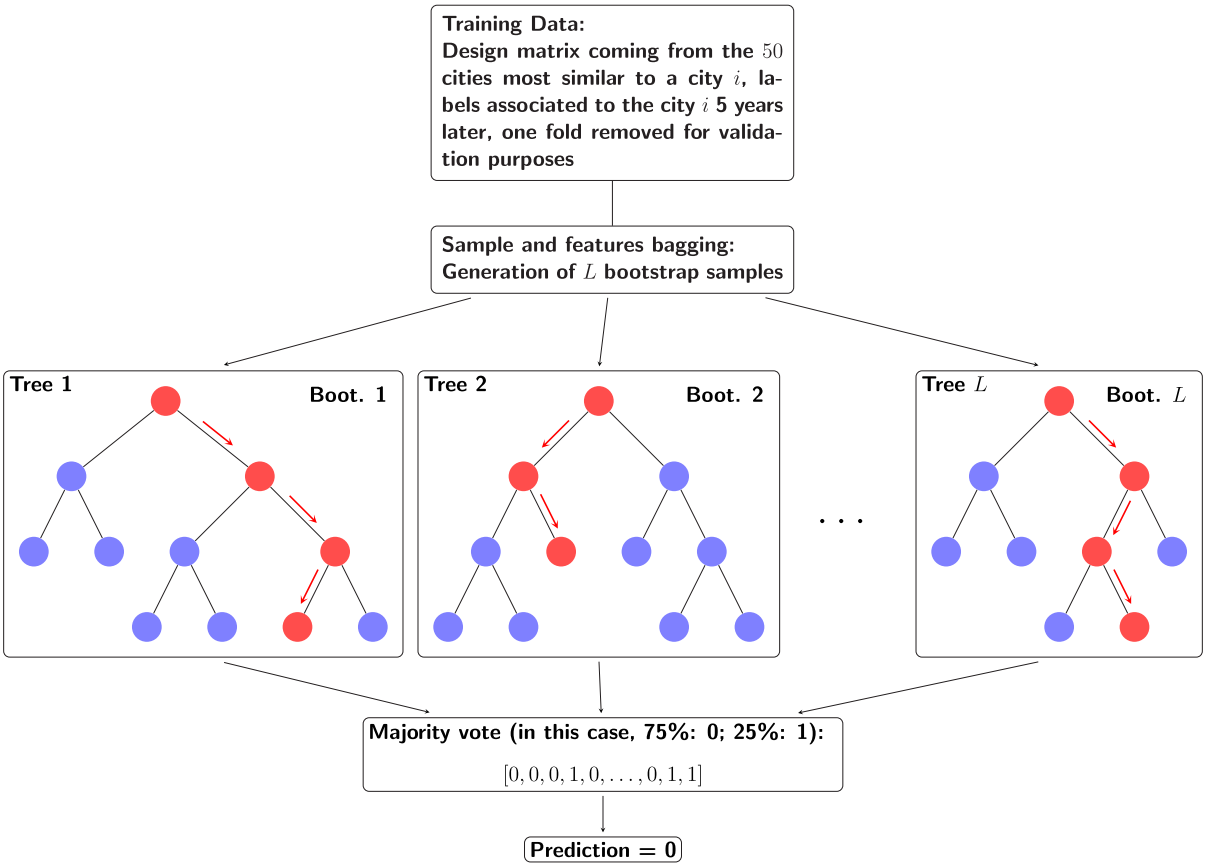


Fig. 3. Graphical representation of how we apply RF to our supervised ML task. For each city i, the RF is provided with an initial data set containing the 50 cities that are most similar to city i. After removing one of the 5 folds from the dataset, a series of bootstrap samples are drawn and used to build the RF model. Finally, a majority vote is performed. The procedure is executed for all 5 folds.

Bootstrapping refers here to the training of many independent decision trees in parallel on different subsets of the training set with varying random subsets of accessible characteristics. The RF classifier then aggregates individual tree decisions for the final decision. Aggregation is done to lower the RF classifier's total variance. Thus, it is usually good in generalizing to unseen data and is less prone to issues of overfitting with respect to other ML methods (Misra et al., 2019). In the case of RF, each decision tree corresponds with a partition of the feature space. Such a partition is constructed in a non-linear way, starting from the training set.

In our study, the hyperparameters of the RF model (to be tuned to achieve an optimal performance) include the number of trees L (n_estimators), the maximum depth of each tree (max_depth), the minimum number of samples required to split an internal node (min_samples_split), the minimum number of samples required to be at a leaf node (min_samples_leaf), and the criterion for evaluating the quality of each split in every node ("Gini" and "entropy" in our case). In our application, tuning of these hyperparameters is obtained following the procedure detailed in Section 3.2. The average best hyperparameters turn out to be min_samples_leaf = 1.4, min_samples_split = 4.5, n_estimators = 92, max_depth = 16, and criterion = "Gini".

Fig. 3 provides a graphical representation of our application of the RF model to the supervised ML task described in Section 3.2 (for the sake of clarity, we assume that the data have been already split into 5 folds, and one is currently running the RF model on the training dataset obtained by removing one of those folds for validation purposes and bootstrapping).

4. Results

This section summarizes the main results of the work. Specifically, Section 4.1 reports the outcome of the supervised ML analysis, selects the best classifier according to several criteria, and provides details on various additional indicators relevant to the study, such as confusion matrices and the performance of the trained supervised ML models on a suitable additional activation/deactivation task. Section 4.2 evaluates the predictions of three complexity indices – the Economic Complexity Index (ECI) (Hidalgo and Hausmann, 2009), the GENeralized Economic comPlexitY) (GENEPY) (Sciarra et al., 2020), and the Matrix cOmpletion iNdex of Economic complexitY (MONEY) (Gnecco et al., 2022) – achieved by applying the various trained supervised ML models. This is done for both the 2009–2014 analysis already detailed in Section 3.2 and, at a reduced time-scale, for an additional rolling-window analysis where

Table 1
Performance on the test set (according to several metrics) of the various supervised ML methods adopted for prediction, in the presence of the Poisson Stochastic Block Model (PSBM) network community detection pre-processing step. For each supervised ML method and for each metric, an average with respect to all the global cities considered in this study is indicated. For instance, the F1-score reported in the table for each specific supervised ML method is the average of the F1-scores of its final models (one for each of the 150 global cities). In bold, the performance of the best supervised ML method for each case is reported. For details on the definitions of the various performance metrics, the reader is referred to the final part of Appendix C of the Online Supplementary Material and to Athey and Imbens (2019) and Chicco and Jurman (2020).

Performance metric	Supervised ML method						
	LR	FNN	SVM	RF	BART	XGB	MC
With PSBM pre-processing							
F1-score	0.735	0.566	0.732	**0.798**	0.733	0.763	0.631
PR-AUC	0.443	0.464	0.468	**0.678**	0.578	0.535	0.502
Matthew's coefficient	0.275	0.202	0.220	**0.303**	0.247	0.250	0.201
ROC-AUC	0.731	0.689	0.742	**0.786**	0.683	0.720	0.700
Average classification error	0.235	0.322	0.240	**0.198**	0.202	0.211	0.204

predictions are made with a forecast horizon of one year (in this second analysis, the complexity indices are also recalculated yearly). Section 4.3 discusses some of the policy implications and applications of our methodology, which could provide governments and public institutions with novel insights into the innovation patterns of global cities. Finally, Section 4.4 examines the ML results obtained from the point of view of their interpretability.

4.1. Supervised machine learning results

Table 1 reports the performance achieved by the various supervised ML methods adopted for solving our prediction task detailed in Section 3.2 according to various metrics, each obtained by averaging the same metric over different final models (i.e. one final model for each of the global cities considered in our study).[30] Among all the supervised ML methods considered, RF achieves the best performance for all the metrics considered in the Table 1, ranking first for the F1-score, Matthew's coefficient, the PR-AUC, the ROC-AUC, and the average classification error. As a robustness check, Table D.1 in Appendix D of the Online Supplementary Material reports the results achieved by replacing the PSBM with two alternative pre-processing methods and with the case of no pre-processing. They confirm the high suitability of the RF method to the specific supervised ML task.[31]

Fig. 4 further details the average classification error, focusing this time on each global city and on the validation set.[32] Interestingly, the results highlight that the values of the incidence matrix associated with some cities look more difficult to classify on average than the values of the incidence matrix associated with other cities (quite independent of the specific supervised ML method adopted). We also find that 42 out of the 60 cities that turn out to be the hardest to predict (i.e. the top 60 cities from the left side of the x-axis in Fig. 4) exhibit the highest 5-year variability in their associated values in the incidence matrix, with an average such variability of 0.318.

In addition, the confusion matrices in Fig. 5 on the main diagonal represent the percentages of correct classification for each of the two classes for the different binary classifiers considered in this study. The confusion matrices refer to the collections of all test sets (by varying the reference global city i in the set of 150 global cities considered) for the supervised ML task reported in Section 3.2,[33] related to the time period 2009–2014. According to these confusion matrices, it turns out that the best trade-off between the two percentages of correct classification (each weighted equally) is obtained by the RF.

As a further analysis, Fig. 5 also shows the confusion matrix obtained for a simple baseline method, namely a "Constant Estimator" (CE), where the predicted M_{ij}^{t+5} value is identical to the true M_{ij}^{t} value. Although at first glance the CE method appears to achieve similar performance results to other supervised ML methods, by construction it is not able to predict temporal changes in M_{ij} values and in particular whether some M_{ij} values will change from 0 to 1 in 5 years, in the period from 2009 to 2014 (hence its predictions would have no real policy implications; see also the comments on the next Fig. 6).

Table 2 shows the percentage of correct classification for the subset of "activating" elements (i, j) of the collection of all test sets, i.e. those that undergo a transition from $M_{ij}^{2009} = 0$ to $M_{ij}^{2014} = 1$, for the different supervised ML methods (including the CE baseline method). Similarly, the table reports the percentage of correct classification for the subset of "deactivating" elements (i, j) of the collection of all test sets, i.e., for those where a transition from $M_{ij}^{2009} = 1$ to $M_{ij}^{2014} = 0$ occurs. For the CE method, by construction,

[30] Similar averaging procedures have been used in other works on economic complexity, e.g. Tacchella et al. (2023). We have chosen the same approach to ensure comparability of results.

[31] More precisely, for each of the two alternative methods, RF achieves the best performance for all the 5 metrics considered. In the case of no clustering, it achieves the best performance for 2 metrics, ranking second for the remaining 3 metrics. In all cases, it turns out that the best performance is always achieved by one of the tree-based supervised ML methods (i.e. RF, BART and XGB).

[32] In this case, being the performance metric computed for each global city, the validation set is considered, since it has a larger numerosity than the test set. Indeed, it covers a larger number of years (9 in place of 1).

[33] Or to a slight variation in the case of MC.

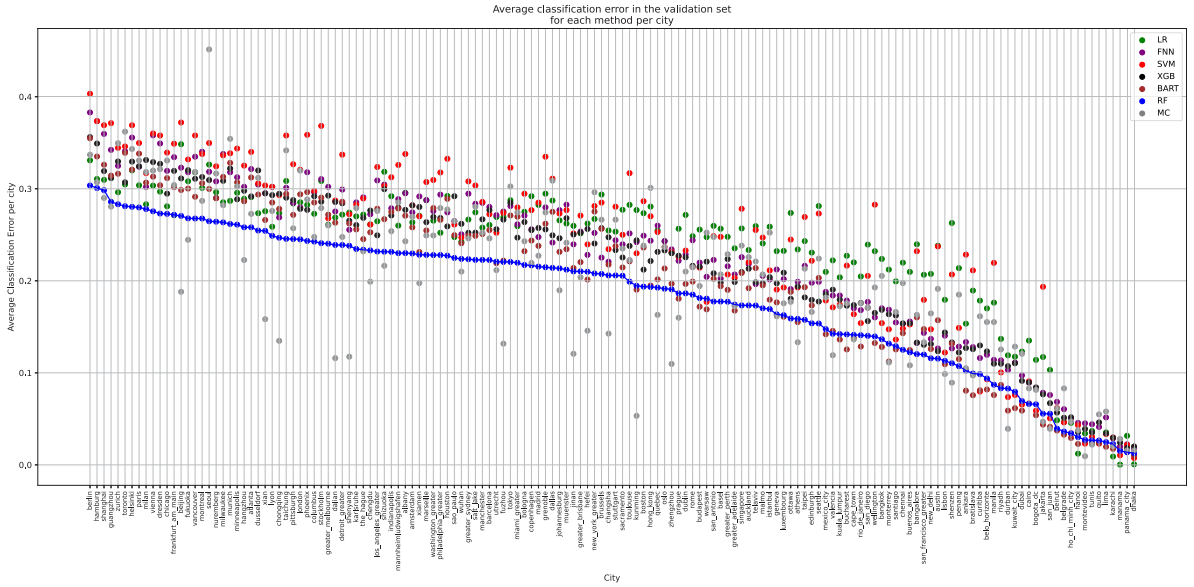


Fig. 4. Average classification error in the validation set per global city for each of the supervised ML methods used in this study. Each average classification error (per global city) is calculated as the number of misclassified labels relative to the total number of labels. The black line shows the trend of the average classification error of the best performing supervised ML method for all metrics considered in Table 1, namely RF.

Table 2
Percentage of "activating" and "deactivating" elements correctly predicted in the collection of all test sets by the different supervised ML methods (including the simple CE baseline method). The highest percentage for each case is highlighted in bold.

Supervised ML method	Percentage of transitions from 0 to 1 caught	Percentage of transitions from 1 to 0 caught
LR	50%	59%
FNN	49%	56%
SVM	45%	61%
RF	**77%**	82%
BART	62%	58%
XGB	66%	59%
MC	51%	**88%**
CE	0%	0%

both these percentages turn out to be equal to 0%.[34] For comparison, the number of zeros and ones in 2009 is 73,456 and 22,244, respectively, while the number of zeros and ones in 2014 is 74,209 and 21,491, respectively. Finally, the number of activating and deactivating elements is 10,556 and 11,309.

4.2. Technological complexity results

To further explore the implications of our results in terms of technological complexity, Table 3 shows the Spearman's correlation coefficients between the rankings of the global cities obtained using several measures of complexity drawn from the literature – i.e. the Economic Complexity Index (ECI) (Hidalgo and Hausmann, 2009), the GENeralized Economic comPlexitY (GENEPY) (Sciarra et al., 2020) and the Matrix cOmpletion iNdex of Economic complexitY (MONEY) (Gnecco et al., 2022) –,[35] when such measures are computed, respectively, based on the true incidence matrix in 2014 and the one containing, for each entry, its prediction[36] in 2014 obtained on the test set by one of the supervised ML methods employed in the present study. Given the nature of the patent dataset used in this work, all these values can be interpreted as measures of the technological complexity of global cities. The table clearly

[34] This task is quite hard to solve, as transitions are somewhat rare (in our dataset, there are about 23% transitions roughly equally distributed in the two directions from 0 to 1 and from 1 to 0).

[35] These measures were originally analyzed using the matrix of Revealed Comparative Advantage (**RCA**) as input in their respective definitions, but can be applied without modification to the case of the Revealed Technology Advantage (**RTA**) matrix (or, more precisely, to the associated incidence matrix, as they all apply a binarization of the input matrix as a pre-processing step). Details on the definitions of the ECI and the GENEPY are reported in Appendix E of the Online Supplementary Material. The reader is referred to Gnecco et al. (2022) for details on the definition of MONEY, which involves more complex computations.

[36] Average prediction, in the case of MC.

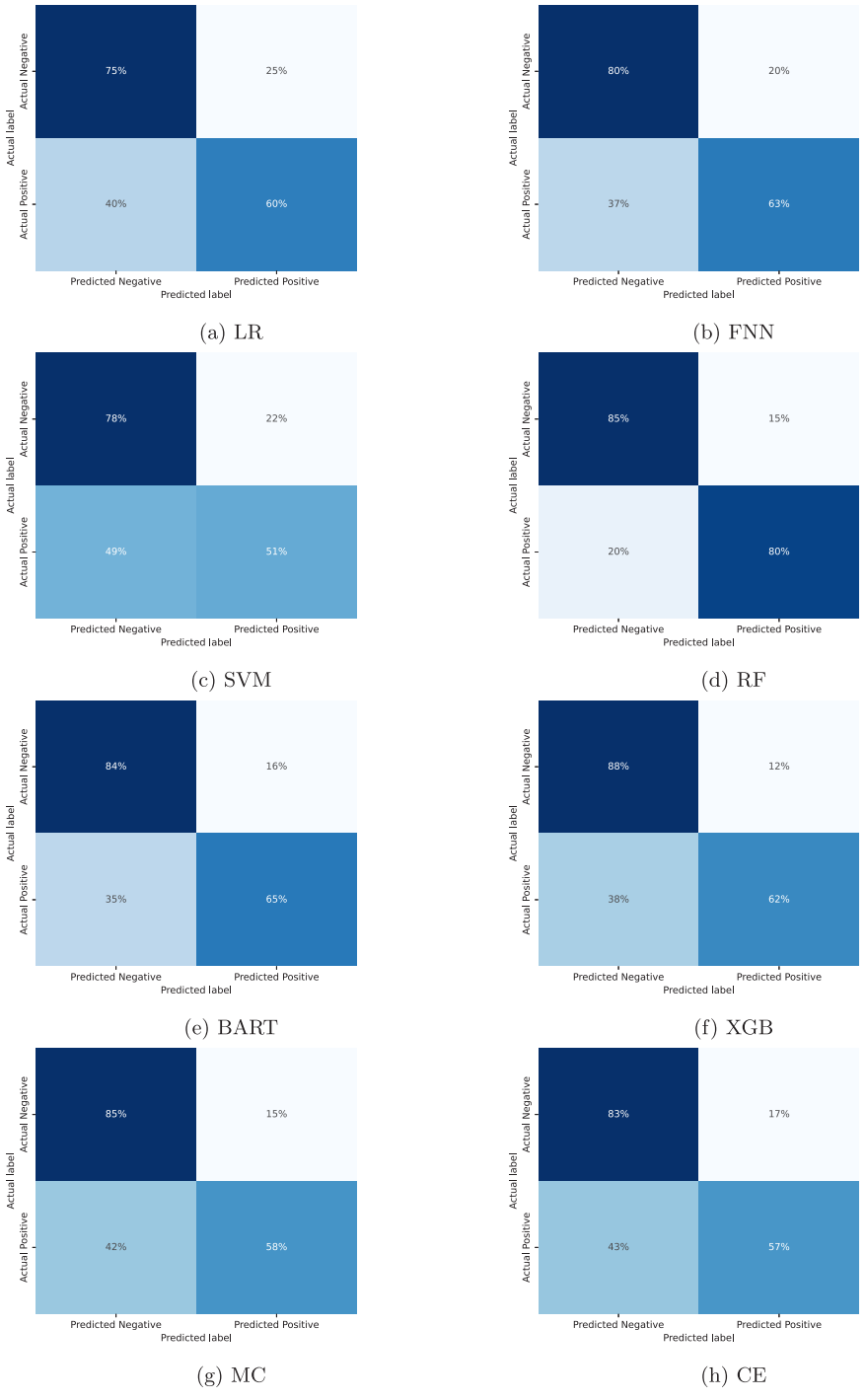


Fig. 5. Confusion matrices with the percentages of correct/incorrect classification for each of the two classes, for the seven types of binary classifiers considered in this study, and for an additional simple base method. The confusion matrices refer to the collections of all test sets (by varying the reference global city i in the set of 150 global cities considered) for the supervised ML task reported in Section 3.2 (and for the additional baseline method), with respect to the time period 2009–2014. Note that the percentages in each row were obtained by row-wise normalization. The original incidence matrices contain on average 68% of zeros and 32% of ones. Accordingly, the threshold for classifying an element as 1 in the displayed subfigures is set to 0.7. Alternative choices for the threshold (0.65 and 0.6) do not lead to significant differences in the displayed confusion matrices (the corresponding results are available on request).

shows that the highest Spearman's correlations are obtained with the tree-based ensemble methods (RF, BART and XGB), whose advantages over other methods have already been discussed in Section 3.3 and in Section 4.1 for the results of other analyses.

Table 3
For each supervised ML method and measure of technological complexity: Spearman's correlation coefficient between the ranking of global cities resulting from the true incidence matrix in 2014 and their ranking resulting from the predicted incidence matrix in the same year. The Spearman's correlation coefficient for the best ML method(s) is shown in bold.

Supervised ML method	Measure of technological complexity		
	ECI	GENEPY	MONEY
LR	0.578	0.591	0.633
FNN	0.657	0.665	0.611
SVM	0.799	0.806	0.703
RF	**0.909**	0.892	0.810
BART	0.836	**0.946**	0.775
XGB	0.878	**0.946**	**0.812**
MC	0.855	0.874	0.788

Table 4
Spearman's correlations between the complexity rankings of cities in successive years. The results are shown for different measures of technological complexity, namely ECI, GENEPY and MONEY. The table contains both the correlations obtained when these measures of technological complexity are calculated from the original incidence matrices $\mathbf{M}^t$ and the correlations obtained when their elements are replaced by the test-set predictions achieved by the RF, BART and XGB methods.

Years	Spearman's correlation based on true $\mathbf{M}^t$ matrices			Spearman's correlation based on RF predictions			Spearman's correlation based on BART predictions			Spearman's correlation based on XGB predictions		
	ECI	GENEPY	MONEY	ECI	GENEPY	MONEY	ECI	GENEPY	MONEY	ECI	GENEPY	MONEY
(2002,2003)	0.867	0.884	0.943	0.882	0.920	0.780	0.905	0.921	0.854	0.949	0.944	0.934
(2003,2004)	0.842	0.956	0.950	0.808	0.909	0.788	0.918	0.909	0.812	0.943	0.944	0.952
(2004,2005)	0.925	0.963	0.957	0.893	0.922	0.878	0.894	0.909	0.844	0.938	0.942	0.932
(2005,2006)	0.912	0.970	0.966	0.840	0.905	0.805	0.876	0.923	0.823	0.964	0.959	0.965
(2006,2007)	0.877	0.952	0.948	0.922	0.899	0.900	0.904	0.905	0.891	0.962	0.955	0.964
(2007,2008)	0.900	0.943	0.938	0.801	0.884	0.838	0.889	0.899	0.867	0.952	0.948	0.951
(2008,2009)	0.409	0.945	0.940	0.855	0.886	0.892	0.901	0.884	0.857	0.947	0.948	0.951
(2009,2010)	0.933	0.944	0.895	0.776	0.892	0.830	0.873	0.892	0.826	0.947	0.951	0.952
(2010,2011)	0.835	0.977	0.975	0.923	0.903	0.900	0.928	0.938	0.912	0.961	0.952	0.960
(2011,2012)	0.977	0.982	0.978	0.944	0.931	0.920	0.920	0.902	0.917	0.942	0.947	0.957
(2012,2013)	0.978	0.963	0.960	0.934	0.894	0.893	0.934	0.894	0.893	0.858	0.863	0.866
(2013,2014)	0.975	0.973	0.970	0.936	0.890	0.885	0.936	0.890	0.885	0.906	0.912	0.915

Table 4 reports, for each of the three measures of technological complexity introduced above (ECI, GENEPY, and MONEY), its Spearman's correlation between its induced rankings in consecutive years to further investigate changes in (true and predicted) technological complexity at a smaller time-scale.[37] The table shows both the Spearman's correlation obtained by computing each measure of complexity based on the original $\mathbf{M}^t$ matrices and the Spearman's correlation obtained by replacing the elements of the $\mathbf{M}^t$ matrices with their test-set predictions achieved by the RF, BART, and XGB methods. To get these predictions, the same methodology described in Section 3.2 is applied, with the following two differences: to refine the granularity level of the analysis with respect to time, the period $(t, t+5)$ is replaced by the rolling window $(t, t+1)$; for each period $(t, t+1)$, the average best hyperparameters are taken based on the cross-validation results generated from the analysis of all the previous pairs of years.[38] The largest Spearman's correlations turn out to be obtained for the GENEPY, consistently for both the ground truth and the various supervised ML methods. In the case of the ECI, the results appear to be unstable even in the ground truth – as the value of the Spearman's correlation obtained for the pair (2008, 2009) is quite different from the ones obtained for the other pairs of years –, which suggests the adoption of the other two measures of technological complexity.[39] Since GENEPY is currently more commonly adopted than MONEY and its evaluation is less computationally expensive, it turns out that GENEPY looks like the most suitable means to evaluate and compare the true and predicted technological complexity of global cities.

As a further step of our analysis on technological complexity, we compare the predicted 2014 GENEPY rankings of global cities (based on the test-set predictions of the RF, BART, and XGB methods obtained by solving the supervised ML task of Section 3.2) with their true 2014 GENEPY ranking, as reported in Table E.1 in Appendix E of the Online Supplementary Material.[40] Predicting the complexity of cities is a difficult task, as the GENEPY rankings of cities usually differ greatly from the corresponding GENEPY rankings of previous years: consider that only 15 of the top 20 cities in 2009 are still in the top 20 in 2014, typically in different positions in the ranking.

Table 5 shows the top 20 global cities in 2014 according to the RF, BART and XGB methods. The RF method correctly identifies 6 of the top 10 cities (according to the GENEPY index) in 2014, with 2 incorrectly predicted cities (Los Angeles and Stuttgart) still

[37] See also the end of Appendix A of the Online Supplementary Material.

[38] For this reason, data for the years 2000 and 2001 are not reported in the table.

[39] Another advantage of the GENEPY over the ECI is discussed in Appendix E of the Online Supplementary Material.

[40] The reader is referred to Figure E.3 in Appendix E of the Online Supplementary Material for a more detailed comparison of true and predicted GENEPY rankings of cities in 2014.

Table 5
Top 20 global cities in 2014 according to the GENEPY, computed starting from the $\mathbf{M}^{2014}$ incidence matrix predicted by means of the RF, BART, and XGB methods (based on information coming from the $\mathbf{M}^{2009}$ incidence matrix). For each city, we report the country, the predicted position in the ranking (Pred. #), the actual position in the ranking (True #) and the variation of its position in the ranking with respect to Table E.1 in Appendix E of the Online Supplementary Material (Δ). Distinct colors correspond with different continents: America (blue), Asia (red), and Europe (green). The color in the column corresponding to true rankings (True #) is used to distinguish the top 10 cities (blue) from the cities ranked 11–20 (purple), 21–30 (orange) or below (yellow). The color intensity in the "Deviation" column (Δ) is related monotonically to the size of the deviation (red if the actual position in the ranking is below the predicted position in the ranking, green if it is above).

Pred. #	Predicted, RF				Predicted, BART				Predicted, XGB			
	City	Country	True #	Δ	City	Country	True #	Δ	City	Country	True #	Δ
1	Shanghai	CN	4	−3	Nuremberg	DE	8	−7	Seoul	KR	1	0
2	Chicago	US	11	−9	Munich	DE	6	−4	Chicago	US	11	−9
3	Tokyo	JP	2	1	Seoul	KR	1	2	Tokyo	JP	2	1
4	Nuremberg	DE	8	−4	Stuttgart	DE	5	−1	Nuremberg	DE	8	−4
5	Seoul	KR	1	4	Tokyo	JP	2	3	Taipei	TW	3	2
6	Paris	FR	18	−12	Frankfurt	DE	32	−26	Shanghai	CN	4	2
7	Zurich	CH	7	0	Los Angeles	US	9	−2	Zurich	CH	7	0
8	Minneapolis	US	17	−9	Washington	US	24	−16	Munich	DE	6	2
9	Shizuoka	JP	19	−10	London	UK	26	−17	Shizuoka	JP	19	−10
10	Munich	DE	6	4	Shizuoka	JP	19	−9	New York	US	12	−2
11	London	UK	26	−15	Chicago	US	11	0	Boston	US	14	−3
12	San Diego	US	44	−32	Atlanta	US	28	−16	Detroit	US	15	−3
13	Dusseldorf	DE	23	−10	New York	US	12	1	Taichung	TW	16	−3
14	Los Angeles	US	9	5	Toronto	CA	29	−15	Los Angeles	US	9	5
15	Stuttgart	DE	5	10	Dusseldorf	DE	23	−8	Stuttgart	DE	5	10
16	Taichung	TW	16	0	Karlsruhe	DE	31	−15	Minneapolis	US	17	−1
17	Boston	US	14	3	Paris	FR	18	−1	Paris	FR	18	−1
18	Frankfurt	DE	32	−14	Miami	US	37	−19	Frankfurt	DE	32	−14
19	Philadelphia	US	35	−16	Boston	US	14	5	Vancouver	CA	21	−2
20	Detroit	US	15	5	Philadelphia	US	35	−15	Stockholm	SE	22	−2

ranking in the top 20. Similarly, the BART method correctly identifies, in the first 10 positions, 6 of the top 10 cities (according to the GENEPY index) in 2014, with none of the incorrectly predicted cities still ranking in the top 20. Finally, in the first 10 positions, the XGB method correctly identifies 7 of the top 10 cities (according to the GENEPY index) in 2014, with 2 of the incorrectly predicted cities (Los Angeles and Stuttgart) still ranking in the top 20. For each method, Spearman's correlations of the two rankings (true and predicted) have already been reported in Table 3.[41] Interestingly, a common pattern of both the true and predicted GENEPY rankings is that German, Japanese, and US cities tend to appear in the top positions, which is in line with the findings by Hausmann et al. (2024a).[42]

To conclude this part of the analysis on technological complexity, in Figure E.2 of Appendix E of the Online Supplementary Material, we report the results of an additional robustness check obtained by fitting (through Ordinary Least Squares, or OLS) the true GENEPY values in 2014 using their predicted values (with the predictions achieved using RF, BART, and XGB). The results show that the RF turns out to yield reliable predictions with an intercept of 0.0714, a slope of 0.736, and an R^2 of 0.587, while BART and XGB produce similar results (an intercept of 0.0727 and 0.0733 for BART and XGB, respectively, and a slope of 0.729 and 0.698 for BART and XGB, respectively).[43] Then, in Appendix F of the Online Supplementary Material, we compare the actual and predicted rankings of technological complexity with those obtained by using other measures of city competitiveness.

4.3. Policy implications

The present subsection explores the policy implications of innovation patterns, competitive dynamics, and regional technology potential. We begin by analyzing how competitive dynamics within technological sectors can guide strategic investments by identifying cities with potential technological advantages in specific IPCs (Section 4.3.1). Next, we examine the spatial evolution of innovation centers in global cities, tracking shifts in the centroids of innovation to reveal regional clusters and west–east movements over time, which offer insights into global competition and knowledge diffusion (Section 4.3.2). We also address the predictive challenges and innovation potential within key IPC sectors, particularly in biotechnology, green energy, and traditional industries, highlighting where strategic investments could unlock hidden technological advantages (Section 4.3.3). Finally, in Appendix H of the Online Supplementary Material, we assess the feasibility and attractiveness of technological investments, identifying cities with competitive advantages and providing a framework for targeted regional investment to enhance local competitiveness.

[41] It is worth noting that such correlations take into account all the positions in the rankings, without giving more importance, e.g., to the first positions. This is the reason why, for instance, the Spearman's correlation in the case of BART turns out to be higher than that in the case of RF, although RF is associated with better (on average) ranking predictions in the first positions.

[42] The latter rankings are at the country level and regard the direct study of the ubiquity of patents.

[43] For the sake of comparison, consider that an intercept of 0 and a slope of 1 constitute an ideal match.

By concentrating on competitive dynamics, regional clustering, and sectoral potential, the present subsection addresses policy dimensions that are both impactful and actionable for innovation-driven growth. As underscored by Hidalgo and Hausmann (2009) and Hausmann et al. (2024b), understanding where cities can achieve a competitive edge is essential for developing targeted, place-based innovation policies. Additionally, identifying emergent regional clusters aligns with the work by Porter (1998), which emphasized the economic significance of local knowledge networks and cluster-based growth. Lastly, our analysis sheds light on hidden opportunities within specific sectors – particularly in biotechnology, green energy, and traditional industries – as drivers of sustainable growth, reflecting the strategic importance of sectoral focus in guiding investments, as discussed by Aghion et al. (2016). Together, the analyses performed in the present subsection provide a structured approach to understanding regional and sectoral innovation potential, highlighting the competitive, spatial, and sectoral aspects most pertinent for policy makers aiming to foster resilient and dynamic technological ecosystems.

4.3.1. Policy implications of innovation patterns and competitive dynamics in technology sectors

Given the better performance of tree-based methods (RF, BART, and XGB) in terms of the percentage of "activating" elements correctly predicted in the collection of all test sets, the following analysis is focused on such methods.[44] Fig. 6 illustrates, using a world map, the results obtained by the RF, BART, and XGB classifiers, and presents their possible policy implications. In the figure, which refers to a specific IPC j taken as an example, the following notation is adopted: as usual, M_{ij}^t indicates the true binarized RTA value, in the year t, of a city i in the IPC j (i.e. $M_{ij}^t = 0$ if $0 \leq RTA_{ij}^t < 1$, $M_{ij}^t = 1$ if $RTA_{ij}^t \geq 1$); additionally, $\hat{M}_{ij}^t$ indicates its test-set prediction obtained by a supervised ML method in the year t (with $\hat{M}_{ij}^t = 1$ denoting a potential technological advantage in the year t, according to that method, of the city i in the IPC j). Each subfigure reports, for the corresponding supervised ML method analyzed: in blue, the locations of the global cities i for which $M_{ij}^{2009} = 1$; in green, the locations of the global cities i for which $M_{ij}^{2009} = 0$, $M_{ij}^{2014} = 1$, and $\hat{M}_{ij}^{2014} = 1$; in yellow, the locations of the global cities i for which $M_{ij}^{2009} = 0$, $M_{ij}^{2014} = 0$, and $\hat{M}_{ij}^{2014} = 1$; in red, the locations of the global cities i for which $M_{ij}^{2009} = 0$, $M_{ij}^{2014} = 1$, and $\hat{M}_{ij}^{2014} = 0$.

In other words, the "blue" cities are global cities i that already achieve a technological advantage for the technology j in the year 2009. Among the cities that do not achieve a technological advantage for the technology j in the year 2009, the "green" cities are global cities i that express their potential technological advantage for the technology j in the year 2014; the "yellow" cities are global cities i that do not express their potential technological advantage for the technology j in the year 2014; the "red" cities are global cities i that achieve a technological advantage for the technology j in the year 2014, going beyond the expectations (expressed by the respective RF, BART, and XGB predictions).[45]

Hence, as a policy implication of the adoption of each of these methods for a similar analysis, policy makers could jointly identify the "green" and "yellow" cities for specific IPCs,[46] and concentrate future investments of such cities in those sectors,[47] allowing them to better express their potential technological advantage.[48]

It is worth remarking that similar insights as those coming from Fig. 6 could be obtained at a more aggregate level (e.g. replacing single observed/predicted elements of the incidence matrix with summations over a subset of IPCs, and the threshold 1 with a larger positive integer). Indeed, although the analysis of individual IPCs, as shown in Fig. 6, is conducted here only for a specific case study, a broader application of this approach to all IPCs provides valuable insights. For example, it can highlight one of the most extensively studied phenomena in innovation economics, providing clues about how competition in various sectors affects innovation.

In generalizing the study reported in Fig. 6, we also examine the relationship between the competition level of an IPC and its number of deactivations (instead of activations) from 2009 to 2014.[49]

To assess the impact of competition, we first quantify the competition level for each IPC by calculating its number of active instances in 2009, i.e. the number of cities with a comparative advantage in that IPC in 2009 (this is also called ubiquity in the economic complexity literature, see Hidalgo and Hausmann (2009)). We then partition the set of the IPCs into quintiles based on this competition metric. The first quintile represents IPCs with the lowest competition level, while the fifth quintile represents those with the highest competition level. Our objective is to establish whether, for each quintile, the average predicted number of deactivations in 2014 is close to the average actual numbers of deactivations in the same year, and whether the latter are higher for quintiles associated with high competition levels with respect to quintiles associated with low competition levels. Deactivations are defined as instances of cities where an IPC is active in 2009 but inactive in 2014.

The results of this analysis are shown in Table 6. They are fairly consistent with the theoretical phenomena described by Akcigit and Van Reenen (2023), which states that very high levels of competition can reduce a firm's market share, sales and profits, making

[44] These methods are also chosen for their characteristic of being easily interpretable, following the influential work by Rudin (2019).

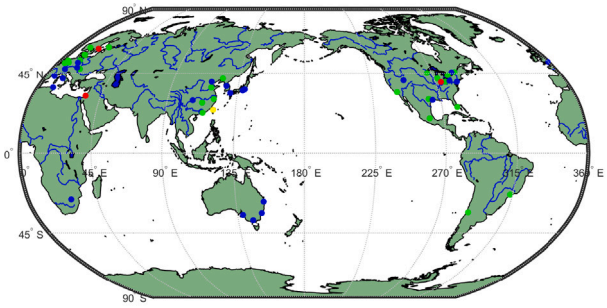
[45] The smaller number of "red" cities for the RF method with respect to the BART and XGB methods are likely due to its better average performance on "activating" elements, as highlighted by Table 2.

[46] I.e. without knowing if the respective true future RTA values will reach or surpass the threshold 1 or not.

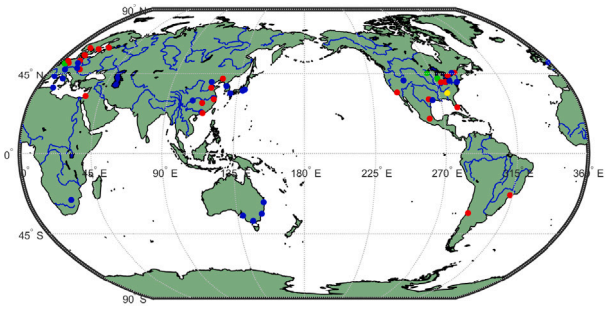
[47] A possible way to achieve this objective could consist in increasing the number of joint patenting activities in collaboration with top inventors coming from global cities having a comparative advantage in such sectors.

[48] It is worth observing that, by construction, no "green" or "yellow" cities would be obtained by using the CE method, confirming its uselessness from a policy perspective.

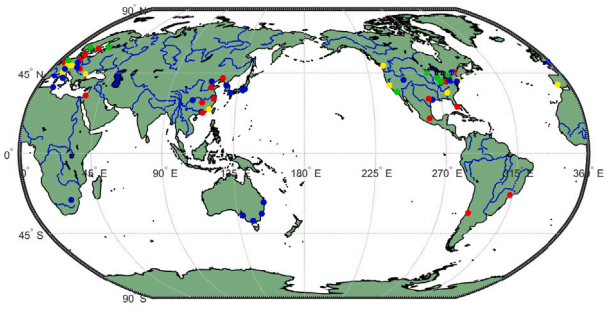
[49] The results of this analysis are reported using the predictions achieved by the RF method, but similar results are obtained with XGB and BART. MC, which achieves the best average performance on the "deactivating" elements of the collection of all test sets, is not considered here because it is more computationally expensive than the other methods, so it is expected to have less policy implications. Moreover, as highlighted by Table 2, the average performance of RF on such elements turns out to be similar to the one of MC.



(a) Predictions achieved by the RF method



(b) Predictions achieved by the BART method



(c) Predictions achieved by the XGB method

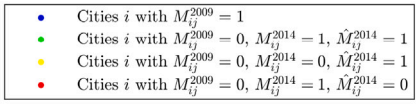


Fig. 6. World map with global cities. The figure shows, for a specific IPC j: in blue, the locations of global cities i, for which $M_{ij}^{2009} = 1$; in green, the locations of global cities i, for which $M_{ij}^{2009} = 0$, $M_{ij}^{2014} = 1$ and $\hat{M}_{ij}^{2014} = 1$; in yellow the locations of the global cities i, for which $M_{ij}^{2009} = 0$, $M_{ij}^{2014} = 0$ and $\hat{M}_{ij}^{2014} = 1$; in red the locations of the global cities i, for which $M_{ij}^{2009} = 0$, $M_{ij}^{2014} = 1$ and $\hat{M}_{ij}^{2014} = 0$. The specific IPC is $j = 105$ – B09C ("Reclamation of Contaminated Soil") – and is chosen to obtain a positive number of "blue", "green", "yellow" and "green" cities for all 3 tree-based supervised ML methods considered in this part of the study. (a) Predictions using the RF method; (b) Predictions using the BART method; (c) Predictions using the XGB method; (d) Legend. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

innovation more difficult (Canare and Francisco, 2021). From a policy perspective, these results could allow policy makers to find IPCs where it might make sense for a city to disinvest (rather than invest in them, which was the case with the comments on Fig. 6). This may, in principle, be related to diversification, as cities may refrain from investing in IPCs where they have little chance of retaining their comparative advantage and reorient their investments towards other IPCs where the results of the supervised ML analysis (tailored to each city) highlight a potential comparative advantage for the specific cities.

Table 6
For each quintile of IPCs (based on their competition levels): average number of active instances in 2009; average number of real deactivations in 2014; average number of predicted (by the RF method) deactivations in 2014.

Quintile of IPC competition level	Average no. of active instances in 2009	Average no. of true deactivations in 2014	Average no. of predicted RF deactivations in 2014
1st	11.196	7.338	8.228
2nd	24.781	14.687	15.390
3rd	33.456	18.464	18.929
4th	42.804	21.744	21.891
5th	61.890	26.375	26.414

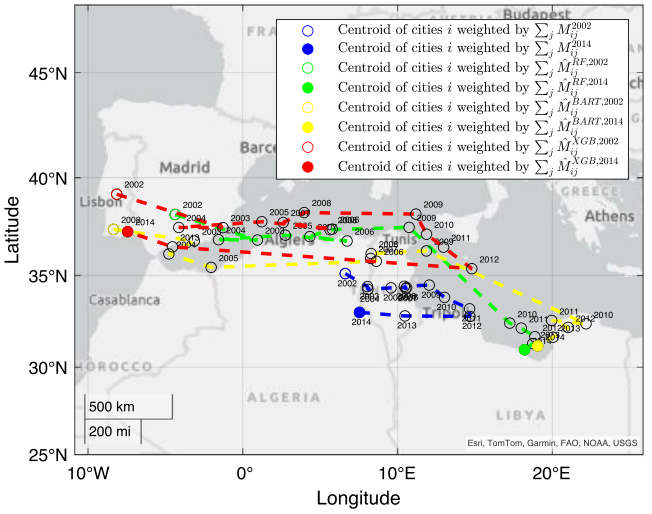


Fig. 7. Evolution with respect to time of the location of the centroid (here taken as the barycenter) of the set of global cities i, for several different choices of the weight given to each city: $\sum_j M_{ij}^t$ (blue dashed line); $\sum_j \hat{M}_{ij}^{RF,t}$ (green dashed line); $\sum_j \hat{M}_{ij}^{BART,t}$ (yellow dashed line); $\sum_j \hat{M}_{ij}^{XGB,t}$ (red dashed line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.3.2. Geographical trajectories of innovation in global cities

As a source of further policy implications, we explore the geographical trajectories of centroids of innovation related to global cities. The need to study innovation trajectories has been emphasized e.g. by Lema et al. (2018). Indeed, such an investigation could help policy makers to gain meaningful information on the evolution of geographical patterns related to innovation over time. Specifically, Fig. 7 shows the evolution over time of the location of the centroid (here as a barycenter) of global cities based on their innovation patterns. The centroid is calculated for different weighting options assigned to each city i:

(i) The number of IPCs for which city i has a true technological advantage in year t (known in the economic complexity literature as its diversity or diversification, see Hidalgo and Hausmann (2009)). This value is represented by $\sum_j M_{ij}^t$ and its evolution over time is shown through a blue dashed line in the figure.

(ii) The number of IPCs for which city i has a predicted technological advantage in year t, represented by $\sum_j \hat{M}_{ij}^t$. Here, $\hat{M}_{ij}^t$ refers to the test-set prediction of M_{ij}^t obtained using either the RF, BART, or XGB method. To distinguish among these three cases, in the following we use the notations $\sum_j \hat{M}_{ij}^{RF,t}$, $\sum_j \hat{M}_{ij}^{BART,t}$, and $\sum_j \hat{M}_{ij}^{XGB,t}$, respectively. For each method, the evolution over time of $\sum_j \hat{M}_{ij}^t$ is shown through a green dashed line in the figure.

To create Fig. 7, the rolling window technique described in Section 4.2 is used. The figure shows similar locations (i.e. in the same part of the globe) of the centroids obtained in the 4 cases, although a greater temporal divergence is observed for the locations of the centroids associated with the RF, BART and XGB predictions with respect to the location of the centroid associated with the ground truth. As a general trend, an initial eastward movement can be observed in all 4 centroids, followed by a westward movement in recent years. Remarkably, all four trajectories depicted in the figure show a much larger variation in the East–West direction than in the North–South direction, which could be related to Diamond's hypothesis on technology diffusion (Bologna Pavlik and Young, 2019). Further results related to the analysis of the trajectories of other suitably defined centroids of innovation are reported in Appendix G of the Online Supplementary Material.

4.3.3. Innovation potential and predictive challenges in key IPC sectors

As a further policy-related analysis, we turn our attention to IPCs, the importance of which was recently highlighted in the work by Hausmann et al. (2024b). Specifically, we compare predicted and actual technological advantages over multiple years. Our goal

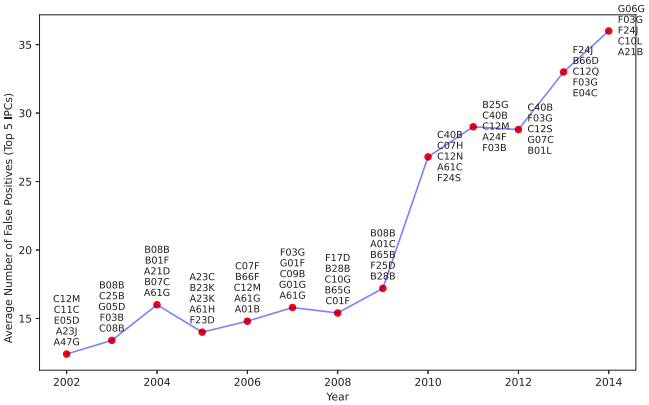


Fig. 8. Top 5 IPCs per year having the highest numbers of false positives, where a false positive is defined as a case for which an IPC is predicted to have $RTA \geq 1$ in a city by at least 2 of the 3 supervised tree-based ensemble ML methods (RF, BART, and XGB), but this does not actually happen.

is to calculate the number of false positives for different IPCs, where a false positive is defined as a case for which an IPC is predicted to have $RTA \geq 1$ in a city by at least 2 of the 3 supervised tree-based ensemble ML methods (RF, BART and XGB), but this does not actually occur. For each year, we identify the top 5 IPCs with the highest number of false positives and calculate the average number of false positives among them. We then create a scatterplot in which each point represents the average number of false positives for the top 5 IPCs in a given year (based on the results of the rolling-window analysis described in Section 4.2), with the corresponding IPC codes appearing near each point. The results obtained are shown in Fig. 8.

Our results are consistent with the recent literature (Tang et al., 2020; Zhu et al., 2023), especially in relation to IPC codes such as C12M ("Apparatus for Enzymology or Microbiology") and G06G ("Analogue Computers"), which were identified as highly innovative in their patent network analysis.

More in general, our results indicate significant hidden potential in IPC codes related to biotechnology,[50] green energy technologies,[51] and traditional sectors,[52] which aligns with broader economic trends. In particular, studies in the field of innovation economics, such as that by Hall and Rosenberg (2010), emphasize the crucial role of biotechnology and the energy sector as drivers of future innovation, while Aghion et al. (2016) highlights the importance of green technologies as key areas for sustainable growth. Furthermore, innovation in traditional sectors supports the notion of technological spillover effects into less expected areas as discussed by Jaffe et al. (1993). These consistent findings confirm the relevance of our results as they highlight the potential of certain IPCs by 2014, which then materialized in line with the findings of the referenced literature (Tang et al., 2020; Zhu et al., 2023).

4.4. Interpretability analysis

Table 5 already provides valuable insights into the hidden technological potential of global cities, by examining the difference between true and predicted GENEPY rankings. However, as is common practice in ML analyses for economics (Athey, 2018), we also include here an interpretability study, focusing specifically on the top 5 cities that experience the largest positive jumps in the RF predicted GENEPY rankings in 2014 when compared to their corresponding true GENEPY rankings in the same year. These cities are identified as those with the largest difference between their highest and lowest positions during this period, moving from a bottom position in the true GENEPY ranking to a top position in the predicted GENEPY ranking.[53] Then, for each selected city of interest, we evaluate the average importance of the input features of the corresponding RF classifier. In the context of the RF method, the average importance of an input feature is evaluated here as the mean of the accumulation of its associated impurity decrease within

[50] C12M ("Apparatus for Enzymology or Microbiology"), C12N ("Microorganisms or Enzymes; Compositions thereof; Propagating, Preserving, or Maintaining Microorganisms; Mutation or Genetic Engineering; Culture Media"), and C12S ("Processes using Enzymes or Micro-Organisms to Liberate, Separate or Purify a Pre-Existing Compound or Composition; Processes using Enzymes or Micro-Organisms to Treat Textiles or to Clean Solid Surfaces of Materials").

[51] F03G ("Spring, Weight, Inertia, or Like Motors; Mechanical-Power-Producing Devices or Mechanisms, not Otherwise Provided for or Using Energy Sources not Otherwise Provided For"), F24J ("Production or Use of Heat not Otherwise Provided For"), and F24S ("Solar Heat Collectors; Solar Heat Systems").

[52] A61G ("Transport, Personal Conveyances, or Accommodation Specially Adapted for Patients or Disabled Persons; Operating Tables or Chairs; Chairs for Dentistry; Funeral Devices"), B08B ("Cleaning in General; Prevention of Fouling in General"), and A01C ("Planting; Sowing; Fertilising"). With "traditional sectors" we refer to industries that are typically considered mature or well-established, with slower rates of technological change compared to cutting-edge fields like information technology or biotechnology. These sectors often involve foundational economic activities such as agriculture, manufacturing, and basic services. For instance, the IPC codes A01C ("Planting; Sowing; Fertilising") and B08B ("Cleaning in General; Prevention of Fouling in General") refer to traditional industries that may not be commonly associated with rapid innovation but still demonstrate hidden potential for technological advancements through the adoption of new methods or technologies.

[53] In other words, these are the cities with the largest positive slopes in the bump chart reported in Figure E.3 in Section E of the Online Supplementary Material.

Table 7
The top five cities by average feature importance for each of the five global cities with the largest positive deviations between the actual GENEPY ranking in 2014 and the GENEPY ranking predicted by the RF in the same year.

Top 5 cities changing in the GENEPY ranking	San Diego	Dublin	Melbourne	Rome	Rio de Janeiro
Top 5 cities according to the average feature importance	Houston	Tallin	Jinan	Beijing	Quebec
	Toulouse	Utrecht	Zurich	Karlsruhe	Bangkok
	Detroit	Lisbon	Philadelphia	Kunming	Detroit
	Hefei	Prague	Brisbane	Seoul	Sacramento
	Stockholm	Athens	Atlanta	Madrid	Paris

each tree of the ensemble (Scornet, 2023).[54] Since, in our application of the RF method, each input feature is associated with a specific city (among the 50 most similar cities to the selected city), we end up with a ranking of cities according to their feature importance for the selected city of interest. The results obtained are reported in Table 7.

This analysis contributes to the literature on Regional Innovation Systems (RIS) by highlighting how shifts in technological complexity can reveal critical insights into regional innovation dynamics. Identifying similar cities based on feature importance contributes to the discussion of path dependence in economic geography, as outlined by Cooke et al. (1997) and Martin and Sunley (2006), by illustrating how certain cities may follow comparable developmental trajectories due to shared characteristics.

The findings obtained in Table 7 resonate with the literature on innovation clusters and urban hierarchies (Porter, 1998), emphasizing how new clusters may emerge as cities gain technological complexity. The significant changes in the ranking positions observed in the cities reported in Table 7 suggest that these cities were emerging as new innovation hubs in 2014, in line with theories of urban hierarchies and cluster formation (Porter, 1998). For instance, Dublin's continued attraction of multinational high-tech companies in 2014 underscores its growing role as a European technology center, due to the presence in such city of the so-called Silicon Docks area (Roberts et al., 2015). It is particularly interesting how the interpretability analysis can catch some well-known patterns, which validate its robustness. For instance, in 2014, Dublin's development in Information Technology (IT) was mirrored by cities like Tallin and Lisbon, which also gained prominence as innovation centers due to favorable policies and skilled workforces (Roberts et al., 2015; Rissola and Sörvik, 2018). These trends underscore how geographically distant cities can follow similar trajectories in their rise as technology-driven economies, illustrating the role of global knowledge networks in fostering regional innovation. Moreover, the identification of similar cities based on feature importance, such as San Diego and Huston, or Melbourne and Brisbane, points to the concept of path dependence mentioned above, where historical trajectories shape current innovation dynamics (Martin and Sunley, 2006). The presence of geographically distant yet technologically similar cities indicates that knowledge spillovers and technological specialization can transcend geographic boundaries, emphasizing the role of global networks in regional innovation.

These results can have practical implications for regional innovation policies, supporting the argument by Rodríguez-Pose (2013) that tailored interventions can enhance the technological capacity of cities and regions. Specifically, the 5 cities reported in Table 7 as those having more potential than their true positions in the GENEPY ranking shows, may have networked with their 5 most similar cities according to the average feature importance, in order to reach their full (hidden) potential.

5. Conclusions

In this paper, an economic complexity approach is used to analyze the technological complexity of global cities. Subsequently, a machine learning approach is applied to predict their future technological competitiveness in specific technological areas. More specifically, our methodology is based on predicting the technological capabilities of cities through a combination of network analysis and machine learning methods. In this work, several supervised machine learning methods are combined with a state-of-the-art network community detection model (namely the Poisson stochastic block model) to predict the future competitiveness of global cities. The generalizability of the combined methods is evaluated on a ground truth derived from an up-to-date dataset of historical patent data with georeferencing to global cities. Random forests, and more generally methods based on ensembles of trees, achieve better predictive performance on most metrics than other supervised machine learning methods to which the same pre-processing is applied. In particular, the ability of Random Forest to work effectively with sparse data makes it particularly effective in overcoming the challenges faced by the literature in predicting the future competitiveness of cities. Therefore, we conclude that random forests, and more generally models based on ensembles of trees, are the first choice for predicting technological complexity, as they can help identify the technological competencies of cities that are the prelude to future technological achievements.[55]

[54] In the context of RFs, "impurity" refers to a measure of how mixed the classes are within a node of a decision tree. Common impurity metrics include Gini impurity (which is the one that we use in this part of the study) and entropy, both of which are minimized when a node contains only a single class. When a feature is used to split the data at a node, it decreases the impurity by making the resulting child nodes more homogeneous, since the objective of a RF is to create nodes that are as heterogeneous as possible among them, while achieving maximum homogeneity within each node (Breiman, 2001). The importance of a feature is thus measured by how much it contributes to reducing impurity across all the nodes in which it is used. Specifically, in RFs, feature importance can be quantified as the mean decrease in an impurity metric (such as Gini impurity or entropy) associated with that feature, accumulated over all the trees in the ensemble. Intuitively, features that lead to larger impurity reductions across the ensemble are more influential in the model's decisions and are therefore considered more "important" for accurate predictions.

Moreover, the results of our study on the trajectories of centroids of global cities, weighted by their innovation measures, are consistent with the dynamic perspectives on innovation trajectories discussed by Balland et al. (2015), where the authors emphasized the importance of proximity and network dynamics in regional innovation development. Moreover, our results are consistent with Frenken et al. (2007)'s emphasis on related variety driving regional growth, as stable innovation centroids may reflect the benefits of regional specialization and knowledge spillovers. These results are also consistent with the seminal work of Audretsch and Feldman (1996b) on R&D spillovers, which emphasized the geographically bounded nature of innovation clusters and underpinned the observed regional concentration of innovation clusters in Section 4.3. Finally, the results are consistent with the findings of the economic complexity literature, particularly with respect to the ubiquity of patenting activity, especially when focusing on the predicted rankings (see Section 4.2).

The predictions of our machine learning approach can be used by policy makers to develop specific strategic innovation policies. By identifying technologies that are currently missing in their technology portfolio, cities can take action to become more competitive in the identified technology areas (Balland et al., 2022). More specifically, by predicting the technological areas in which it is most likely to be competitive (or uncompetitive) in the future, a city can decide on which technological areas to focus (or not focus) its investments, leading to the activation or possible deactivation of the technological advantage in these areas.[56] In this way, the city could better orient itself towards other cities that are active in selected technology sectors and are considered reference cities. This can lead to the development of targeted smart policies that incentivize the city's economic actors to engage in projects that are in line with competitiveness in the identified technology areas and possibly improve collaboration in innovation processes that also involve the reference cities just identified.

Some possible limitations of this approach, however, are that innovation policy can sometimes be exogenous to patent activity and that a large proportion of patent applications do not lead to product innovation (a strategy used by large firms to discourage competition and not drive technology development in the market). Therefore, the predictability of the competitive patenting landscape may not adequately capture the competitiveness of innovation activities in a city.

It is worth noting that in the present work, the features used to train the different supervised machine learning models come exclusively from the incidence matrix. While this may seem like a limitation, there are several reasons for doing so. First, the data used for our machine learning analysis corresponds to that typically analyzed in the context of economic complexity studies, i.e. the elements of the incidence matrix. Second, the results we obtained in several parts of our analysis, in particular the results on the matching between actual and predicted technological complexity (see Table 3 in Section 4.2), turn out to be quite good. Third, based on our pre-processing step described in Section 3.1, a different machine learning model is trained for each city of interest (using data from that city and from the 50 most similar cities identified in this step, see Section 3.2). In this way, each model is tailored to the city in question, so in a sense it can also implicitly take into account other possible features associated with that city. Nevertheless, it is worth noting that the current approach could be adapted with some modifications to incorporate appropriate additional features into the analysis.[57]

Finally, we discuss some possible extensions of our research. First, a possible extension of our analysis is to include more city-specific attributes and more cities in the prediction problem to better capture the heterogeneity of institutional, technological and economic characteristics across cities. Second, a larger variety of machine learning techniques could be used, such as additional dimensionality reduction techniques (Gnecco and Sanguineti, 2009) and constraints encoding prior knowledge about the specific learning task (Gnecco et al., 2013). Another further promising class of models that can be employed involves the use of generative artificial intelligence for predicting conditional probabilities based on extensive information about the subject (Du et al., 2024). A valuable extension of this approach would be to condition with respect to all covariates and the historical data of cities, leveraging the pre-training of these powerful models to enhance predictions in activation/deactivation tasks. Third, the analysis can also be extended over time to a larger number of cities and to other areas of analysis such as scientific production as new data sources become available. Fourth, the additional use of federated/decentralized machine learning methods could reduce the computational burden and mitigate the privacy issues associated with data sharing in the application of machine learning methods. Fifth, another promising direction for future research is the integration of our analysis into an agent-based model to analyze the location choice of factors such as human capital and investments in research and development. For instance, other relevant variables could be

[55] It is worth mentioning that our re-training of models associated with the various supervised machine learning methods for the pair of years (2009, 2014) (based on the average best hyperparameters found for the previous pairs of years) is possible due to the availability of labels for the year 2014. Nevertheless, our methodology is flexible enough to be adapted to the case in which future labels are not available. For instance, one could apply to the final models not only the average best hyperparameters but also the average best parameters found for the previous pairs of years, eliminating the need for re-training. Additionally, one could restrict the training/validation phase to previous periods not overlapping with the last available year in the dataset, whose features are used to predict future elements of the incidence matrix. It is worth mentioning that, while this kind of extension would be straightforward for the case of the first 6 supervised machine learning considered in the work, this is not the case of matrix completion, which – in the present implementation – requires (at least partial) information related to the year to which the prediction themselves refer.

[56] The deactivation of a technological advantage is relevant here, due to the resource constraints of a city, which could lead that city to refocus its investments on other technological sectors with higher potential for the development of the specific city. Deactivation may be associated, e.g. by the existence of a high-level of competition for that IPC, consistently with the findings obtained in Section 4.3.

[57] In the supervised machine learning approach used in this work, where a different model is trained for each city of interest, some care would be required to include meaningful additional features as inputs to the different supervised machine learning models. As described in detail in Section 3.2, the different feature vectors are obtained by varying the IPC, so only IPC-related features could be easily integrated into this approach. In particular, this would exclude geographic features (such as latitude and longitude), which would be constant for each model, or time-varying features, which would require each model to be trained based on data from multiple pairs of years (again, to avoid constant features). Including a large number of additional features would therefore require not training a model for each city and each year. While this is possible in principle, it is beyond the scope of this paper and is the subject of further research.

included in the prediction problem, either as independent or outcome variables, such as those related to immigration patterns of innovators. Sixth, alternative pre-processing methods could be used, to identify subsets of cities most similar to a city of interest, in preparation of the successive supervised machine learning analysis.[58] Seventh, the policy implications derived from the application of our methodology could be the starting point for the development of recommendation systems, providing to each city suggestions of strategic innovation policies tailored to the specific technological profile of that city.

Declaration of competing interest

The authors declare not to have any conflict of interest.

Acknowledgments

The work was supported in part by the "Contributi Liberali 2022" project "Application of Matrix Completion Techniques to the Definition of Economic and Financial Recommender Systems", granted by the Bank of Italy, by the PRIN 2022 project "MAHATMA" (CUP: D53D23008790006), funded by the European Union – Next Generation EU program, and by the PRIN PNRR 2022 project "MOTUS" (CUP: D53D23017470001), funded by the European Union – Next Generation EU program. The authors wish to thank the Editor and three anonymous Reviewers for their constructive feedback, which increased the quality of the article with respect to its first version submitted to the journal.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jebo.2025.107011.

Data availability

Data will be made available on request.

References

Aghion, P., Dechezleprêtre, A., Hemous, D., Martin, R., Van Reenen, J., 2016. Carbon taxes, path dependency, and directed technical change: Evidence from the auto industry. J. Political Econ. 124 (1), 1–51.
Akcigit, U., Van Reenen, J., 2023. The Economics of Creative Destruction. Harvard University Press, Cambridge.
Aljalbout, E., Golkov, V., Siddiqui, Y., Strobel, M., Cremers, D., 2018. Clustering with deep learning: Taxonomy and new methods. arXiv preprint arXiv:1801.07648.
Arthur, W.B., 1994. Increasing Returns and Path Dependence in the Economy. University of Michigan Press.
Athey, S., 2018. The impact of machine learning on economics. In: Agrawal, A., Gans, J., Goldfarb, A. (Eds.), The Economics of Artificial Intelligence: An Agenda. pp. 507–547.
Athey, S., Imbens, G.W., 2019. Machine learning methods that economists should know about. Annu. Rev. Econ. 11, 685–725.
Audretsch, D.B., Feldman, M.P., 1996a. Innovative clusters and the industry life cycle. Rev. Ind. Organ. 11, 253–273.
Audretsch, D.B., Feldman, M.P., 1996b. R&D spillovers and the geography of innovation and production. Am. Econ. Rev. 86 (3), 630–640.
Balassa, B., 1965. Trade liberalisation and revealed comparative advantage. Manch. Sch. 33, 99–123.
Balland, P.A., Boschma, R., Crespo, J., Rigby, D.L., 2019. Smart specialization policy in the European union: Relatedness, knowledge complexity and regional diversification. Reg. Stud. 53, 1252–1268.
Balland, P.-A., Boschma, R., Frenken, K., 2015. Proximity and innovation: From statics to dynamics. Reg. Stud. 49 (6), 907–920.
Balland, P.A., Broekel, T., Diodato, D., Giuliani, E., Hausmann, R., O'Clery, N., Rigby, D., 2022. The new paradigm of economic complexity. Res. Policy 51 (3), 104450.
Balland, P.A., Jara-Figueroa, C., Petralia, S.G., Steijn, M.P., Rigby, D.L., Hidalgo, C.A., 2020. Complex economic activities concentrate in large cities. Nat. Hum. Behav. 4 (3), 248–254.
Beaverstock, J.V., Smith, R.G., Taylor, P.J., 1999. A roster of world cities. Cities 16 (6), 445–458.
Belderbos, R., Benoit, F., Edet, S., Lee, G.H., Riccaboni, M., 2022. Global cities' cross-border innovation networks. In: Cross-Border Innovation in a Changing World: Players, Places, and Policies. Oxford University Press, pp. 160–185.
Bettencourt, L.M., Lobo, J., Strumsky, D., 2007. Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size. Res. Policy 36 (1), 107–120.
Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008 (10), P10008.
Bologna Pavlik, J., Young, A.T., 2019. Did technology transfer more rapidly East-West than North-South? Eur. Econ. Rev. 119, 216–235.
Boschma, R., Iammarino, S., 2009. Related variety, trade linkages, and regional growth in Italy. Econ. Geogr. 85 (3), 289–311.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Buyukyazici, D., Mazzoni, L., Riccaboni, M., Serti, F., 2024. Workplace skills as regional capabilities: relatedness, complexity and industrial diversification of regions. Reg. Stud. 58 (3), 469–489.
Canare, T., Francisco, J.P., 2021. Does competition enhance or hinder innovation? J. Southeast Asian Econ. 38 (1), 24–50.
Chakravarty, D., Goerzen, A., Musteen, M., Ahsan, M., 2021. Global cities: A multi-disciplinary review and research agenda. J. World Bus. 56 (3), 101182.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. J. Artificial Intelligence Res. 16, 321–357.

[58] In principle, the evaluation of the ground truth of the technological complexity indices considered in this work could be used to make such an extension. In this paper, we have preferred not to do so, as the associated true rankings of cities are quite unstable over time, unlike the communities found with the Poisson stochastic block model. This might be an advantage of this model in terms of the interpretability of the subsequent machine learning results.

Chicco, D., Jurman, G., 2020. The advantages of the Matthew's correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21 (1), 1–13.
Chu, B., Qureshi, S., 2023. Comparing out-of-sample performance of machine learning methods to forecast US GDP growth. Comput. Econ. 62 (4), 1567–1609.
Cimini, G., Gabrielli, A., Sylos Labini, F., 2014. The scientific competitiveness of nations. PloS One 9 (12), 1–11, Article no. e113470.
Cooke, P., Uranga, M.G., Etxebarria, G., 1997. Regional innovation systems: Institutional and organisational dimensions. Res. Policy 26 (4–5), 475–491.
Dosi, G., 2023. Innovation as an evolutionary process. In: The Foundations of Complex Evolving Economies: Part One: Innovation, Organization, and Industrial Dynamics. Oxford University Press, pp. 97–171.
Du, H.S., Belderbos, R., Somers, D., 2022. Research versus development: Global cities and the location of MNCs' cross-border R&D investments. Reg. Stud. 56 (12), 2001–2018.
Du, T., Kanodia, A., Brunborg, H., Vafa, K., Athey, S., 2024. LABOR-LLM: Language-based occupational representations with large language models. arXiv preprint arXiv:2406.17972.
Edet, S., 2022. Essays on Innovation Networks and Global Cities (Ph.D. thesis). IMT School for Advanced Studies, Lucca, Italy and Faculty of Economics and Business, KU, Leuven, Belgium, https://e-theses.imtlucca.it/348/1/Edet_phdthesis.pdf.
Edet, S., Panzarasa, P., Riccaboni, M., 2021. Global cities in international networks of innovators. Adv. Complex Syst. 24 (03n04), 2140002.
Feng, L., Wang, Q., Wang, J., Lin, K.-Y., 2022. A review of technological forecasting from the perspective of complex systems. Entropy 24 (6), 787.
Filippone, M., Camastra, F., Masulli, F., Rovetta, S., 2008. A survey of kernel and spectral methods for clustering. Pattern Recognit. 41 (1), 176–190.
Florida, R., Adler, P., Mellander, C., 2018. The city as innovation machine. In: Turok, I., Bailey, D., Clark, J., Du, J., Fratesi, U., Fritsch, M., Harrison, J., Kemeny, T., Kogler, D., Lagendijk, A. (Eds.), Transitions in Regional Economic Development. Routledge, pp. 151–170.
Fontanari, T., Fróes, T.C., Recamonde-Mendoza, M., 2022. Cross-validation strategies for balanced and imbalanced datasets. In: Proceedings of the Brazilian Conference on Intelligent Systems. Springer, pp. 626–640.
Frenken, K., Van Oort, F., Verburg, T., 2007. Related variety, unrelated variety and regional economic growth. Reg. Stud. 41 (5), 685–697.
Fritz, B.S., Manduca, R.A., 2021. The economic complexity of US metropolitan areas. Reg. Stud. 55 (7), 1299–1310.
Gnecco, G., Gori, M., Sanguineti, M., 2013. Learning with boundary conditions. Neural Comput. 25, 1029–1106.
Gnecco, G., Nutarelli, F., Riccaboni, M., 2022. A machine learning approach to economic complexity based on matrix completion. Sci. Rep. 12 (1), 1–10, Article no. 9639.
Gnecco, G., Sanguineti, M., 2009. Accuracy of suboptimal solutions to kernel principal component analysis. Comput. Optim. Appl. 42, 265–287.
Goulet Coulombe, P., 2024. The macroeconomy as a random forest. J. Appl. Econometrics 39, 401–442.
Guimerà, R., Sales-Pardo, M., 2009. Missing and spurious interactions and the reconstruction of complex networks. Proc. Natl. Acad. Sci. 106, 22073–22078.
Hall, B.H., Rosenberg, N., 2010. Handbook of the Economics of Innovation. Elsevier.
Harris, N., 2007. City competitiveness. Originally drafted for a World Bank study of competitiveness in four Latin American cities.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, second ed. Springer.
Hausmann, R., Hidalgo, C., Bustos, S., Coscia, M., Simoes, A., 2014. The Atlas of Economic Complexity: Mapping Paths to Prosperity. MIT Press.
Hausmann, R., Yildirim, M.A., Chacua, C., Hartog, M., Matha, S.G., 2024a. Global Trends in Innovation Patterns: A Complexity Approach. Economic Research Working Paper Series, no. 80, World Intellectual Property Organization (WIPO).
Hausmann, R., Yildirim, M.A., Chacua, C., Hartog, M., Matha, S.G., 2024b. Innovation Policies Under Economic Complexity. Economic Research Working Paper Series, no. 79, World Intellectual Property Organization (WIPO).
He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21 (9), 1263–1284.
Hidalgo, C.A., 2021. Economic complexity theory and applications. Nat. Rev. Phys. 3 (2), 92–113.
Hidalgo, C.A., Hausmann, R., 2009. The building blocks of economic complexity. Proc. Natl. Acad. Sci. 106 (26), 10570–10575.
Holland, P.W., Laskey, K.B., Leinhardt, S., 1983. Stochastic blockmodels: First steps. Soc. Netw. 5 (2), 109–137.
Jaffe, A.B., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. Q. J. Econ. 108 (3), 577–598.
Kachniewska, M., Kowalski, A.M., Szczech-Pietkiewicz, E., 2018. The Competitiveness of Cities: Components, Meaning and Determinants. World Economy Research Institute SGH Warsaw School of Economics, Warsaw.
Kamiya, M., Ni, P., Guo, J., Li, B., Ma, H., Xu, H., Gauntner, L., Allou, S., Aldon, L., Eguino, H., et al., 2020. Global Urban Competitiveness Report (2019–2020). UN HABITAT, Nairobi, Kenya.
Karrer, B., Newman, M.E., 2011. Stochastic blockmodels and community structure in networks. Phys. Rev. E 83 (1), 016107.
Kerkache, M.H., Sadeg-Belkacem, L., Benbouzid-Si Tayeb, F., Ali, A., 2022. Supervised learning using community detection for link prediction. In: Proceedings of the International Conference on Computing Systems and Applications. Cham: Springer International Publishing, pp. 85–94.
Lapatinas, A., Litina, A., Poulios, K., 2022. Economic complexity of cities and its role for resilience. PloS One 17 (8), 1–24, Article no. e0269797.
Lee, C., Wilkinson, D.J., 2019. A review of stochastic block models and extensions for graph clustering. Appl. Netw. Sci. 4 (1), 1–50.
Lema, R., Rabellotti, R., Gehl Sampath, P., 2018. Innovation trajectories in developing countries: Co-evolution of global value chains and innovation systems. Eur. J. Dev. Res. 30, 345–363.
Li, R., Dong, L., Zhang, J., Wang, X., Wang, W.-X., Di, Z., Stanley, H.E., 2017. Simple spatial scaling rules behind complex cities. Nat. Commun. 8 (1), 1–7.
Maggioni, M.A., Uberti, T.E., 2009. Knowledge networks across europe: Which distance matters? Ann. Reg. Sci. 43, 691–720.
Martin, R., Simmie, J., 2008. The theoretical bases of urban competitivenss: Does proximity matter? Rev. Econ. Régionale Urbaine 3, 333–351.
Martin, R., Sunley, P., 2006. Path dependence and regional economic evolution. J. Econ. Geogr. 6 (4), 395–437.
Mewes, L., Broekel, T., 2022. Technological complexity and economic growth of regions. Res. Policy 51 (8), 104156.
Misra, S., Li, H., He, J., 2019. Machine Learning for Subsurface Characterization. Gulf Professional Publishing.
Moreno, R., Paci, R., Usai, S., 2006. Innovation clusters in the European regions. Eur. Plan. Stud. 14 (9), 1235—-1263.
Morrison, G., Buldyrev, S.V., Imbruno, M., Arrieta, O.A.D., Rungi, A., Riccaboni, M., Pammolli, F., 2017. On economic complexity and the fitness of nations. Sci. Rep. 7 (1), 1–11, Article no. 15332.
Nowicki, K., Snijders, T.A.B., 2001. Estimation and prediction for stochastic blockstructures. J. Amer. Statist. Assoc. 96 (455), 1077–1087.
OECD, 2020. Cities in the World: A New Perspective on Urbanisation. OECD Urban Studies, OECD Publishing, Paris.
Peixoto, T.P., 2018. Nonparametric weighted stochastic block models. Phys. Rev. E 97 (1), 012306.
Peixoto, T.P., 2023. Descriptive vs. Inferential Community Detection: Pitfalls, Myths and Half-Truths. Cambridge University Press.
Porter, M.J., 1990. The Competitive Advantage of Nations. Free Press.
Porter, M.E., 1998. Clusters and the new economics of competition. Harv. Bus. Rev. 76 (6), 77–90.
Pugliese, E., Cimini, G., Patelli, A., Zaccaria, A., Pietronero, L., Gabrielli, A., 2019. Unfolding the innovation system for the development of countries: Coevolution of science, technology and production. Sci. Rep. 9 (1), 1–12, Article no. 16440.
Pugliese, E., Zaccaria, A., Pietronero, L., 2016. On the convergence of the fitness-complexity algorithm. Eur. Phys. J. Spec. Top. 225 (10), 1893–1911.
Raimbault, J., Pumain, D., 2023. Innovation dynamics in multi-scalar systems of cities. In: ALIFE 2023: Ghost in the Machine: Proceedings of the 2023 Artificial Life Conference. p. 143.
Ricardo, D., 1817. On the Principles of Political Economy and Taxation. John Murray.
Rigby, D.L., 2015. Technological relatedness and knowledge space: Entry and exit of US cities from patent classes. Reg. Stud. 49 (11), 1922–1937.
Rissola, G., Sörvik, J., 2018. Digital Innovation Hubs in Smart Specialisation Strategies. Publications Office of the European Union, Luxembourg.

Roberts, J., Worrall, J., Burke, E., Connolly, P., 2015. Silicon Docks: The Rise of Dublin as a Global Tech Hub. Liberties Press.
Rodríguez-Pose, A., 2013. Do institutions matter for regional development? Reg. Stud. 47 (7), 1034–1047.
Rozenblat, C., Pumain, D., 2007. Firm linkages, innovation and the evolution of urban systems. In: Taylor, P., Derudder, B., Saey, P., Witlox, F. (Eds.), Cities in Globalization: Practices, Policies and Theories. Routdledge, pp. 130–156.
Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1 (5), 206–215.
Sassen, S., 2001. Cities in the global economy. In: Paddison, R. (Ed.), Handbook of Urban Studies. SAGE, pp. 56–272.
Sciarra, C., Chiarotti, G., Ridolfi, L., Laio, F., 2020. Reconciling contrasting views on economic complexity. Nat. Commun. 11 (1), 1–10.
Scornet, E., 2023. Trees, forests, and impurity-based variable importance in regression. Ann. Inst. Henri Poincare (B) Probab. Stat. 59 (1), 21–52.
Soete, L., 1987. The impact of technological innovation on international trade patterns: The evidence reconsidered. Res. Policy 16, 101–130.
Straccamore, M., Bruno, M., Monechi, B., Loreto, V., 2023. Urban economic fitness and complexity from patent data. Sci. Rep. 13 (1), 1–13, Article no. 3655.
Straccamore, M., Pietronero, L., Zaccaria, A., 2022. Which will be your firm's next technology? Comparison between machine learning and network-based algorithms. J. Phys.: Complex. 3 (3), 035002.
Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., Pietronero, L., 2013. Economic complexity: Conceptual grounding of a new metrics for global competitiveness. J. Econom. Dynam. Control 37 (8), 1683–1691.
Tacchella, A., Zaccaria, A., Miccheli, M., Pietronero, L., 2023. Relatedness in the era of machine learning. Chaos Solitons Fractals 173, 114071.
Tang, Y., Lou, X., Chen, Z., Zhang, C., 2020. A study on dynamic patterns of technology convergence with IPC co-occurrence-based analysis: The case of 3D printing. Sustainability 12 (7), 2655.
Traag, V.A., Waltman, L., Van Eck, N.J., 2019. From louvain to Leiden: Guaranteeing well-connected communities. Sci. Rep. 9 (1), 1–12, Article no. 5233.
Tu, C., Zeng, X., Wang, H., Zhang, Z., Liu, Z., Sun, M., Zhang, B., Lin, L., 2018. A unified framework for community detection and network representation learning. IEEE Trans. Knowl. Data Eng. 31 (6), 1051–1065.
Verginer, L., Riccaboni, M., 2020. Cities and countries in the global scientist mobility network. Appl. Netw. Sci. 5, 1–16.
Verginer, L., Riccaboni, M., 2021. Talent goes to global cities: The world network of scientists' mobility. Res. Policy 50 (1), 104127.
Zaccaria, A., Cristelli, M., Tacchella, A., Pietronero, L., 2014. How the taxonomy of products drives the economic development of countries. PloS One 9 (12), 1–17, Article no. e113770.
Zhang, F., Wang, Y., Liu, W., 2020. Science and technology resource allocation, spatial association, and regional innovation. Sustainability 12 (2), 694.
Zhu, J.X., Sun, M., Wei, S.X., Ye, F.Y., 2023. Characterizing patent big data upon IPC: a survey of triadic patent families and PCT applications. J. Big Data 10 (1), 85.