Indicators and Metrics of Emerging Country-Level STEM Innovation

Kelsey Hollenback[1], James H. Lambert[1], Igor Linkov[2]

[1] University of Virginia

[2] US Army Engineer Research and Development Center

ABSTRACT: A variety of signals, metrics, and data may be pertinent to identifying country-level partners in science and technology innovation, and especially newcomers, within a given technology area. Identifying and determining which metrics can be reliably used is necessary to inform policy. There are extant metrics that could be used to predict innovation; however, a certain level of reliability is required for data to be useful. Data collection must consider the reliability and viability of the data in addition to its relevance in predictive modeling for innovation.

KEYWORDS: Bibliometrics, Big Data, Leading and Lagging Indicators

INTRODUCTION

The definition of innovation varies across disciplines. For the purposes of this paper, we define innovation, at the country level, to be new science and technology that a country creates; and the existence of ability and mechanisms to share and communicate with the financial capacity to do so. Of importance is the nation's *prospective* potential and suitability for collaboration with the United States, which may be predicated on a nation's ability to assist the United States on achieving some research and/or innovation goal.

Previous approaches to national innovation focused primarily on the economic structure of country-specific innovation systems [1]-[2]. Common considerations include a nation's human capital, research infrastructure, information infrastructure, the business and financial environment, and the regulatory

environment. These econometric models provide an objective base for predicting innovation, but they fail to account for many aspects of a nation's research ability and identifying emergent national capabilities.

This innovation may be reflected in a small pocket of unique activity that leverages a specific local event or situation, e.g. Chile and aquaculture. Different context-framing is related to the structure of the science discussed. Equipment-intensive sciences are centralized, e.g. CERN, and human resources therefore gather around the equipment. Science and technology innovation may be top-down and distributed, such as the human genome project, or bottom-up research activities that may be distributed or centralized.

Previous work on identifying these indicators and metrics include the Foresight and Understanding of Scientific Exploration (FUSE) Program [3], an IARPA program. IARPA's goal was to identify the next emerging technology, funding teams to examine data sources and signals. However, today there are new datasets, analytics, and methods, which invite new exploration of this topic.

Ideally, collection and consideration of a small set of metrics for a country would provide an efficient way of identifying potential collaboration partners in a specific science or technology field. A similar approach was used predict the number of medals earned by different countries in the Olympics. Bernard and Busse [4] made statistically relevant and accurate predictions of medal count using only the metrics of national size, wealth, and resources.

METRICS

In this section we survey traditional and non-traditional metrics. Some of these metrics may be short-term, some are longer term. Also, some may be leading indicators, while others are lagging indicators. Five general categories of metrics can be identified: (1) input, such as funding and personnel; (2) program activities, such as lab construction and equipment purchases; (3) program capacities established, such as

patents filed, degree programs founded, and legal framework laid; (4) partner activities of capacity versus capability; and (5) objectives and outcomes, such as economic development or breakthrough achievements.

Some potential categories and sources of metrics are presented below. There are overlaps and interactions amongst these metrics; some of these items are identified.

Migration of Personnel. Internationally, there is a great deal of interest in science human resources statistics. These statistics can help identify emergent areas of importance and research. Migration of people with certain interests and backgrounds to build expertise can provide some insight into what technological and scientific advances that a nation may be targeting or building.

Most countries are unable to effectively collect this form of information, and, if they are, the databases do not currently have the necessary granularity to sufficiently track migration of personnel. Existing datasets, such as the National Science Foundation's survey of earned doctorates and survey of doctorate recipients, has an international component and represents an effort to track recipients once they go overseas.

Countries of origin and industries for U.S. H1B visa applications may give some indication as to which countries are producing expatriate scientists who are moving to the United States. Some data on movement of people that does not involve the U.S. are available from a variety sources including Open Doors in New York, OECD mobility data, IEEE, and social network scanning such as from Facebook location posts and LinkedIn employment history.

Applications for Ph.D. programs, not restricted to admitted students only, could also produce this migration information. A good way to corroborate this particular signal would be to look at Ph.D. applicants from ten years ago and match them to whether countries of origin are emerging in science and technology innovation. There is a lag in building education and eventually building technological and scientific

capacity and capabilities. This situation may also require migration information to determine human resource flows back to country of origin.

Expert Opinion. Expert opinion can be valuable in helping to identify where hidden newcomer countries may exist. But, finding experts is an important step. Surveying academic literature to identify a U.S. expert(s) at the top of the field may be a start. Contacting them directly and asking about individuals and countries that are best for collaboration, is a rough but potentially very accurate method for generating candidate partner nations. This becomes a metric, if tools like Delphi approaches are used. Contacting existing research networks, likewise, and iterating on that process to triangulate outstanding researchers and from what countries they originate is a similar method. Like all expert judgment, this is a subjective metric, and methods to collate expert judgment should be used.

Existing Collaborations. Google and Bing search API (application programming interface) could be used to locate university faculty pages and examine with whom these faculty are collaborating internationally. This form of data scraping and mining can be useful to identify various flows of human resources in addition to existing and emergent collaborations.

GitHub holds records of members that copy a repository and or collaborate with each other. It is a unique resource for identifying emergent and current collaborations. Tracking who is producing code and where the collaborations are originating is a method for identifying potential areas of emerging innovation. GitHub is widely used by industry, which is both a positive and a negative. Similar websites such as Source Forage, Black Duck, and Depsy, may also help to identify patterns of collaboration.

Strategic public policy toward science/technology. Innovation at a country level requires long-term strategies and policies. Proxy measures in terms of strategic level investments and plans could include a variety of measures. Some of this information may be hidden, some information may be very obvious on

the specific targeted technology. For example, defense spending overall can be general, but much of it may be for research and development. But, in that case the transparency of what specific technology investments are occurring will be low. Other measures include grant award data, regional programs such as technology parks, and standardization may be indicators of scientific infrastructure investment.

One of the issues and concerns is that some metrics may or may not be linked to innovations. For example, political stability and intellectual property are not necessarily tightly coupled with innovation at the country level.

National commerce data is a rich source of indicators for scientific infrastructure investment. These include industry data, subsidies and tax breaks, and basic science government-industrial joint projects. The Frost & Sullivan investment risk measure, whose calculations include inflation, 5-year outlook, and election results, is a potential indicator. There is a tool called the 'Frost Radar' [5], although it is based on company level analyses, aggregating to the national level may be possible, or at least, similar metrics and indicators can be utilized at the national level.

Laboratory, research and development, and education capital spending, are data and indicators that may be available from the World Bank - for example, on foreign direct investment (FDI). The World Competitiveness Report, better known as the Davos Report, contains in the footnotes data about the amount of time required to set up a company in a given country. These indicators may provide insight into how quickly countries can develop and exploit innovations.

Patents per capita may produce useful data but this metric may be problematic for many reasons discussed later. Momentum and acceleration of patents may also indicate momentum and acceleration of innovation. The European Patent Office sells turnkey patent systems, and examining who is purchasing these systems may reveal who is making investments in science infrastructure.

Insurance coverage and insurance premiums may help in quantifying value to the enterprise. This insurance coverage may identify potential risks and major investments by industries.

Shipping manifests, intellectual property rights, and licensing fees may reveal areas of innovation activity. But industry is difficult to disaggregate using this data. For example, for electronics, it may be difficult to discern information about entertainment industry versus scientific equipment. Outside of the U.S., the information is difficult to obtain.

Tech transfer agreements are useful, but tend not to be disclosed in public records. U.S. company investments into foreign companies are easy to track if the company is publicly traded, but again, differentiating investment in consumables vs. investment in science is very difficult to do. International private investment and crowdfunding platforms like Kickstarter are not captured; and may be sources of additional non-traditional type metrics on innovation.

Publications and citations. Publications and citations are examples of publicly available information that can show patterns of country-level innovation. It is within this area that a significant amount of data is being analyzed to help trace various fields and areas of scientific development. The metrics in this group can be evaluated quantitatively using various statistics and bibliometrics.

As an early measure of innovation emergence are conferences and conference publications. Usually, early research work is presented at these conferences and may serve as leading indicators. For example, a survey of technology conference attendance and conference proceedings can yield which countries are sending scientists to communicate and share knowledge and findings that may indicate emerging innovation. This would require access to conference information, or at least some collection and analysis of various

conferences. Some conferences have specific themes and if many are dominated by a country or themes, some initial linkages can be made.

Journal paper submissions may also be indicators. One approach is to not restrict these submissions only to accepted papers but also include rejected papers. Rejected papers may be due to the relatively novelty of a topic without much rigorous investigation due to novelty. But, many times this information is difficult to obtain. A technique to build a database on this might be to potentially survey journal editors and then present the data in the aggregate to preserve anonymity.

Author affiliation, to determine country location, as listed on SCOPUS and Web of Science is a promising source of data. But, there are limits since the SCOPUS and Web of Science databases cover a subset of journals.

Some countries have databases about academic and scientific research that is searchable. For example, Brazil's LATTES system and the ORCID adoption rate by country may be done within specific disciplines.

Another metric, at the institutional level, are institution journal subscriptions. The types of journals may provide insights into what fields and innovations are most important for an institution, and at an aggregated level for countries. These subscriptions may to an extent be a proxy for submissions and downloads. A limitation is that journal packages are negotiated at a national level and may include some journals which are never or only rarely accessed. Thus this type of data may be misleading.

There are disparate results in terms of country level involvement in these databases and standard journal publication processes. Non-Western journals, for example in China and Africa, or the 15,000 national publications from BRIC countries, are not captured by Web of Science. The Web of Science is a very exclusive database of only the top journals and conferences included, potentially excluding many

publications. This selectivity may cause a loss of capturing emerging trends. Also, it may be a lagging indicator since some of the work has matured greatly by the time it appears in the Web of Science databases.

To counter some of these issues, those who do not write in English have started to formulate their own databases. Such "unseen science" may include regionally useful information that does not reach the point of global significance, e.g. fruit tree diseases of Southeast Asia. A key question is that, since some countries do not emphasize research publications as much as the U.S. and other industrialized nations, what activity is being done instead that can be measured?

A related set of information to journal and conference proceeding publications is to consider citation metrics. Scholars may be grouped into various disciplines (e.g. Google Scholar allows use of keywords, as do other databases for keywords used in publications). If there are clear disciplines or topics that are covered, measures such as citations per country and per capita can provide some insight to country specific focus. There are also statistics that can be provided and adjusted on a per-discipline basis, since not all disciplines have the same level of citations. There may also be opportunities to consider not only current situations, but evolving situations – 'first and second derivatives' of number of citations.

Crises. Nation-specific chronic problems can also guide potential innovations for various regions of the world or countries. Types of infectious disease and crop fungi are examples. These chronic problems may require establishing centers or institutes to address them. Examples may include major new ministry-level science program; similar to an Apollo or Manhattan Project. Other examples include the energy crisis in Japan and global climate change, Israel and desalination, and Saudi Arabia and renewables. Such projects require engagement with the international community, for example via the United Nations, and may include shared vulnerabilities and acceleration; they represent problems that need to be solved with urgency and immediacy. This type of information may be gathered through government policy documents, blogs, social

media, or the popular press. This approach would require significant effort and linkages to various other information.

Capacity and capability building. In terms of analogy, capacity can be equated to owning a bicycle, whereas capability translates into the ability to ride a bicycle. Capacity building includes a number of potential metrics to determine if and where it is occurring. The number and variety of university degrees offered and the addition of new university departments, especially in publically supported institutions, are two potential metrics. Investment in various labs to serve as conduits for knowledge application would be a capacity – capability link. A resource in this situation is the classification scheme developed by RAND to measure science and technology capacity for 150 countries of the world [6].

Leadership. Leadership figures such as CEOs, community leaders, and nonprofit leaders who have degrees or backgrounds in STEM fields may indicate an area emerging in science and technology innovation. This connects to the "key man theory" in regional economics. Civic entrepreneurs, however, are difficult to find because they are often "behind the scenes" and generally want to keep as low a profile as possible. Rewards for leadership figures may not be financial; instead, they may be equated to status or perks. Leaders in society who have a strong innovation and science background may serve as potential indicators of societies and countries on the verge or following through on innovation.

Corruption metrics. Corruption metrics such as the Transparency International Corruption Index and UN 1540 compliance data might be a good proxy for ability to follow through on strategic decision-making commitments. However, corruption may be used as a means for innovation or a way to move toward technology and away from agriculture. In high-trust societies, corruption may be effective; in low-trust societies, corruption undermines trust. Depending on the cultural context, too, bribery may be expected and not viewed as bribery.

Prizes and awards. Awarding of mathematics and science innovation prizes such as the Fields Medal, X-Prize, or various U.S. National Aeronautics and Space Administration (NASA) prizes may indicate areas emerging in science and technology innovation.

USING THE METRICS

One approach to identifying emerging nations is to sort countries into the categories of scientifically proficient, scientifically emerging, and scientifically lagging, and then examine lagging and proficient countries for political stability, science and technology indices, science and technology spending and related legislation, and the migration of science and technology human capital. This generates a list of potential candidate nations for partnership, which can be examined for indicators for movement and other existing metrics, with the awareness that some metrics may not apply. Published literature, for example, will not provide a reliable indicator for early stage emerging nations.

Another approach is to create a conditional flowchart, with each considered country first identified as either a highly innovative country or a low innovative country. Metrics can then be combined into a logic model with accompanying probabilities. This approach has the advantage that it could potentially be computerized into a decision support tool.

Finally, if the question of best candidate for partnership is urgent and time-sensitive, the best method for identifying emerging nations may simply be to contact experts and elicit expert judgment.

Ideally, a group may be formed that could scan a specific science or technology community and produce a periodic bulletin, such as Wentworth Institute of Technology or Wheeling Jesuit University. This is difficult for the United States to accomplish because its science and technology activities are decentralized.

Fast followers are better able to accomplish this task, but it is very difficult to do as a science/technology leader.

OPPORTUNITIES AN CHALLENGES

An opportunity exists for policymakers to leverage big data and vast computing power to mine datasets and derive insights. The introduction of computer power and machine learning such as semi-supervised training may allow signals and metrics to be mined wholesale. For example some matching and analytics programs may be developed for finding a country-country pairing on a university website or in a curriculum vitae, or by training in descriptions of science parks with the goal of identifying previously unknown science parks. Data sources, however, may be biased, and it is unknown how to quantify that bias; research on bias determination to help in accuracy of predictions and identification schemes would be needed. Machine learning may allow combinations of metrics that have not been combined before, but because they have not been combined before, there is not a good understanding of what those combinations mean. Web scraping, along with automated natural language processing, may be able to identify innovation clusters. Yet, training set bias would be very difficult to overcome.

Further, one might use computational techniques to identify clusters via remote sensing or anomaly detection, but prior knowledge of a country context would be necessary to establish a baseline as anomalies are detected as change from baseline. It would also be interesting to draw a sample frame from Web of Science/SCOPUS and then compare with GitHub to try to correlate measures to see how measures from original data sources comport with what is already known.

However, large datasets may become cumbersome to manage and utilize completely. There is also a need to narrow the set of metrics to a manageable quantity. It is inefficient to use hundreds of metrics and unnecessarily complex models, when few key metrics can produce comparable results. The value of

information gained by each additional indicator will likely diminish, especially when the costs of data collection are considered.

There are also a number of ongoing challenges faced by practitioners. One area is factoring in social aspects and indicators for a country's research ability. Past work on national innovation metrics indicates that the environment a country creates, specifically one where individuals with curiosity and talent can flourish, is important to fostering innovation potential. More subjective indicators are valuable prediction components as a way to factor in these social and societal influences. One example of a semi-subjective indicator set is Hofstede's cultural dimension theory [7]. The cultural dimensions provide country data relating to six different aspects of culture, such as "uncertainty avoidance" and "long-term versus short-term orientation".

The biggest barrier to identification of emerging nations in science and technology innovation is language. There is a Chinese version of Web of Science, non-Chinese authors frequently publish in high-quality Chinese journals, which are not easily accessible to examination from the United States. Thus, multilingual systems are needed to be able to decipher emergent innovations in non-English languages.

Finally, an ongoing challenge is the use of patent databases and publication databases. There are significant issues in the use of patent data. Access, type, language, and purpose of patents are only some of the concerns. Similarly, significant issues exist in the use of publication data. For example, much of this data is biased toward countries that have strong publications. As mentioned previously a significant portion of world researchers do not publish in many of the journals. Some scientific and technological domains, for example computer science, are not covered by traditional metrics such as SCOPUS and Web of Science because the most innovative work is not published. Also, much of this information are lagging indicators, at least for basic research publications. Quantitative metrics may have one to three years of lag time and require field work to obtain.

CONCLUSIONS

This paper explored the promises and challenges associated with metrics for identifying innovation partners. The following are some practical takeaways for practitioners and policymakers working in this field:

- When working with metrics, remember that some metrics may be leading indicators, while others may be lagging. All metrics and datasets have pros and cons which much be considered.
- While machine learning and other "big data" approaches can be leveraged, it is important to remember that these approaches are not perfect. One must be mindful of bias in training sets.
- Signals and metrics may be region-specific, e.g. signals and metrics that are valid for France may not be valid for Cameroon. Some indicators and metrics may be more appropriate for regions of the world.
- No single measure, index, or model will successfully identify emerging innovation in science and technology. Multiple independent measures and models need to be corroborated. A mix of qualitative and quantitative data will likely need to be utilized.
- Signals and metrics for innovation will likely identify more potential partnerships than are realistically feasible to enact. Decision models for prioritizing partnerships are needed, in addition to policies to incentivize and promote lasting relationships.

REFERENCES

1. Hogan, M., Gallaher, M. (2018). "Quantitative Indicators for Country-Level Innovation Ecosystems." Publication No. OP-0051-1805. RTI Press.
2. EBRD (2014) "Drivers of Innovation," (2014). European Bank for Reconstitution and Development.
3. https://www.iarpa.gov/index.php/research-programs/fuse

4. Bernard, A.B., and Busse, M.R. (2004). Who wins the Olympic games: economic resources and medal totals. *The Review of Economics and Statistics*, Vol. 86, No. 1, pp. 413-417.
5. https://ww2.frost.com/research/frost-radar/
6. Wagner, C.S., Brahmakulam, I.T., Jackson, B.A., Wong, A., Yoda, T. (2001). "Science & Technology Collaboration: Building Capacity in Developing Countries?" MR-1357.0-WB. Santa Monica, CA: RAND Corporation.
7. Hofstede, G. (1991). *Cultures and Organizations: Software of the Mind*. London: McGraw-Hill.

Acknowledgements: This work was supported by the National Science Foundation under Grant 1848669 "Assessing International Collaboration Opportunities for Science and Technology Innovation: Methods and Approaches". The opinions expressed herein are those of the authors alone, and not necessarily of their affiliated institutions.