RIETI Discussion Paper Series 24-E-075

Quantifying the Differences in Innovation Processes in China, Japan and the United States by Document Level Concordance between Patents and Web Contents

MOTOHASHI, Kazuyuki
RIETI

ZHU, Chen
University of Tokyo

RIETI
Research Institute of Economy, Trade & Industry, IAA

The Research Institute of Economy, Trade and Industry
https://www.rieti.go.jp/en/

RIETI Discussion Paper Series 24-E-075
October 2024

Quantifying the Differences in Innovation Processes in China, Japan and the United States by Document Level Concordance between Patents and Web Contents [i]

Kazuyuki Motohashi (RIETI, University of Tokyo)
Zhu Chen (University of Tokyo)

Abstract

While innovation performance at country level has been analyzed using a variety of STI indicators, the relationship between them such as the patent-new product relationship is under-investigated. Historically, the relationship between technology and industrial output has been analyzed using technology-industry concordance matrices, but the granularity of output information is bounded by the industrial classification system. In this study, we use the text information in both patent and product-related keywords extracted from company's web site contents to come up with detailed concordance information between technology and products, and compare them across three countries, China, Japan and the United States. First, we apply a dual attention model to extract product/service information from web page information. Then, using the textual information of both patent abstracts and product/service keywords, we develop a machine learning model to predict products/services from a particular type of technology. Then, we use this transformation model (from technology to product) to understand the difference in innovation processes of the three countries.

Keywords: text analysis, patent, web mining, innovation process, international comparison
JEL codes: O31, O57

[i] This study is conducted as a part of the Project "Research on Digital Innovation Models" undertaken at the Research Institute of Economy, Trade and Industry (RIETI). The draft of this paper was presented at the RIETI DP seminar for the paper. I would like to thank participants of the RIETI DP Seminar for their helpful comments. The authors also acknowledge financial support from MEXT/JSPS KAKENHI (Grant Number: 23K25540)

1. Introduction

Understanding the difference of innovation process across countries is important, particularly for technology and innovation policy discussion in international organizations such as OECD. One of major functions of such international fora is cross learning of innovation policies, or the best practice of a particular type of policy. The applicability of a policy of one country to the other depends on the similarity of economic and institutional conditions between these counties. Therefore, a series of comparative analysis of the innovation system have been conducted to understand its commonalities and the differences across OECD countries. (OECD, 1999; OECD, 2002; OECD 2005).

The innovation system is made up of components, relationships, and attributes (Carlsson et al., 2002). The main component is a firm to conduct R&D for new products and processes, but a public research institute including university is also an important player in national innovation system (Lundvall, 1992; Nelson, 1993). The relationship stands for both market and non market transactions. For example, the market transactions include formal collaborative research between a firm and a university, while there can be also an informal interaction between them at an academic conference. Finally, the attributes are the properties of the components and the relationships between them. At the end, the way that the components interact each other shapes the attributes of the system where the system level innovation performance is determined.

Empirically, the national innovation system can be analyzed by variety of STI indicators measuring the activities of its components and interactions. For example, the attribute of Japan's innovation system is profiled to be strong in industrial system and knowledge investments, but weak on system performance. It is because in Japan, the R&D of business sector is quite active together with strong patent indicators, while it is weak in collaborative innovation with public research institutions (OECD, 2005).

However, such characteristics of Japan is attributed its industry structure as well. Japan is strong in automotive industry, where substantial R&D is conducted in general, but a less collaboration with university is found, when it is compared to pharmaceutical industry for example. Under the Pavitt Taxonomy on technology, automotive industry is categorized into scale intensive industries, while pharmaceutical industry in grouped into science based industry, where distinct features of production technology and innovation process to the market in each group are found (Pavitt, 1984). As a result, sectoral system of innovation is conceptualized by looking into sector specific characteristics of the attributes of innovation actors and interactions (Breschi and Malerba, 1997).

This paper provides new findings on the measurement of innovation process, to compare three countries. A novel feature of our methodology is based on fine grained textual information at firm, allowing us to separate cross country differences from cross industry ones. First, the product information at firm is obtained by keyword extraction from its web page, by using the dual attention model. (Motohashi and Zhu, 2023). Then, the transformation matrix from technology (text embedded

vector patent abstract) to product (the same for extracted keywords) is estimated for Japan, US and China, individually. Finally, the difference index based on the cosine similarity between the actual and imputed product vector by industry and country is proposed to compare the innovation process across countries and industries.

The organization of this paper is as follows. The next section is for related literature of this paper, and the methodology part together with the dataset to be used in this paper follows. Then, the construction of the transformation matrix from technology to product to characterize the innovation process is presented. Furthermore, the section for proposing a indicator to reflect the difference across countries and firm type (startups, established domestic firms or multinationals) is provided. Finally, this paper concludes with a summary of our findings and some future works.

2. Related Literature

2.1. Technology industry concordance matrix

The relationship between (patented) technology and industrial output have been quantified by the technology industry concordance matrix. Canadian patent examiners assigned patent filings to industries of origin and used codes between 1972 and 1995, which gives the information on the linkage of a particular type of technology with product/service information as the input and/or output. Everson and Putnam (1988) used this information to come up with the concordance table with eight IPC sections to 25 industries (Yale Technology Concordance (YTC)). In 2002, Johnson provided an additional concordance between the IPC and ISIC codes, referred to as the OECD Technology Concordance (OTC). Johnson (2002) used the industry of origin (IOO) and industry of use (IOU) codes as the basis for the YTC and translates them into the Canadian SIC system. To make the results compatible with international data, the Canadian SIC system was translated into ISIC codes in the second step.

Another approach proposed by Schmoch et al. (2003) uses the industrial classification of patent applicants. Specifically, these authors assigned IPC codes to 44 industrial fields based on NACE industries and the patent activities of approximately 3000 firms active in these industry sectors. This methodology for the technology industry concordance table has been applied by several scholars in subsequent studies such as Dorner and Harhoff (2017) in Germany and Ikeuchi et al. (2017) in Japan. Neuhausler et al. (2019) compared existing studies to evaluate performance differences between datasets and methodologies.

2.2. Patent information and its beyond for innovation process assessment

Another stream of related studies is textual analysis of patent information to identify its potential applications. Yoon et al. (2014) explored technological opportunities by constructing technology and product morphologies from collected patent data. Wang and Chen (2019) detected potential

opportunities based on the notion that emerging technologies are outliers in a patent corpus. Undeniably, patent data are useful for technology opportunity discovery analysis; however, a single data source cannot present the most up-to-date and non-technological information (e.g., market side information). To address this issue, the trademark can be an alternative proxy for evaluating market-side innovation since it is assigned to designated products and services, protecting the intended business usage (Sandner and Block, 2011). In this light, Lee and Lee (2017) introduced a modified collaborative filtering approach to explore new business opportunities by combining patent and trademark data. Jeong et al. (2019) investigated the commercialization strategy of new technology by analyzing design patents and trademarks.

Besides, other studies have shown that web-based data such as technical news, Wikipedia articles, and company websites can also be informative supplements for patent data (Arora et al., 2013; Kwon et al., 2018; Park and Geum, 2022). Finally, Choi and Kwon (2023) proposed a text-mining based approach to identify innovation opportunities in the field of smart grids by integrating science (i.e., papers), technology (i.e., patents), and business (i.e., ProQuest) databases.

2.3. Comparative analysis of innovation process : Country vs Sector

The world becomes flatter but it is not completely flat (Ghemawat, 2007). There are substantial differences in economic institutions, formal and informal rules in business activities, at country level, which contributes to the variations of innovation process across countries. Hall and Soskice (2001) develop a theory of comparative innovation advantage, based on their argument on variety of capitalism. That is, a comparative advantage in radical innovation is found for LMEs (liberalized market economies), such as US and UK, while an incremental innovation is more fitted to the economic institutions in CMEs (coordinated market economies). The concept of national innovation system pays an attention also to the role of home market (Lundvall et al,.2002). Therefore, the pattern of international competitiveness by product from trade statistics influences the production and innovation system at country level.

At the same time, there is also a view that the way that new knowledge is generated for innovation activities, or the technology regime, is not constant across countries. For example, Nelson and Winter (1982) described that there are two types of technology regime, an entrepreneurial and routinized ones. Malerba and Orsenigo (1997) pointed out three conditions to determine the technology regime, i.e., technology opportunities, appropriability conditions and cumulativeness in knowledge generative process, and found that the data on industry-specific technological conditions (technology regime) are remarkably consistent across countries. Therefore, they proposed that the innovation system should be evaluated at the sectoral level (sectoral innovation system), taking into account those conditions to explain the type of technology regime.

In addition, these technology conditions should be evaluated dynamically along the stages of

industry lifecycle. In the early stages of an industry's lifecycle, where new knowledge is relatively important (entrepreneurial technology regime), a diverse range of product innovations is important for industrial dynamics. However, as dominant designs or standard product configurations are established, knowledge generation process is routinized and the main source of innovation shifts to process innovation (Utterback and Abernathy, 1975). Accordingly, new firms play a dominant role in the emerging embryotic stage of a new industry, whereas large established firms become the main players in the process of industry maturity (Klepper, 1997).

3. Data and methodology

3.1. Data

In this study, we use company websites and patent portfolios to come up with text information on each firm's products and services. To collect company web data, the initial step is to retrieve the homepage URLs of listed companies. The homepage URLs for U.S. and Chinese companies were obtained from Moody's DataHub, while those for Japanese companies were sourced from the Japanese Financial Service Agency. In total, we scraped web data from 6,893 U.S.-listed companies, 3,829 Japanese-listed companies, and 4,335 Chinese-listed companies. Kinne and Axenbeck (2020) proposed a web mining framework to collect company web content, viewing a website as a company's online presence. The highest level is the root URL (i.e., homepage), with web content comprising (1) the textual information and (2) subpages directly linked to it (e.g., www.company-name.com/products). Subpages on different domains, such as www.x.com/company-name, are excluded.

As for technological inputs of each firm, we collected companies' patent portfolios from multiple patent databases, including the United States Patent and Trademark Office (USPTO), the Japan Patent Office (JPO), and the China National Intellectual Property Administration (CNIPA). In summary, for the U.S., 3,118 companies had at least one patent, while 3,775 companies had none. For Japan, 1,789 companies had at least one patent, while 2,040 companies had none. For China, 3,588 companies had at least one patent, and 747 had none.

3.2. Product Keyword Retrieval Using Dual-Attention Model

A critical issue with company web data is that, while it contains valuable product information, irrelevant content is also included, such as details about factory locations and recruitment activities. This noise may obscure the subsequent analysis of innovation opportunity discovery. In terms of this, Motohashi and Zhu (2023) proposed a dual-attention model that automatically identifies product-related keywords in companies' web content. The model retrieves product keywords using two attention layers: page-level attention, which selects product-related URLs, and word-level attention, which identifies product keywords on the selected web pages. The model is trained with a binary objective, distinguishing between high-tech and non-high-tech companies, where high-tech companies

are defined as those with at least one patent.

As the dual-attention model is initialized with pretrained word vectors prior to training, it is essential to first align the pretrained word vectors from the U.S., Japan, and China in order to effectively train the model on the combined US-JP-CN dataset. Following Conneau et al. (2018), we aligned these pretrained multilingual word vectors by using a set of anchor pairs. The pretrained multilingual word vectors are sourced from FastText word vectors, which are pretrained on the Common Crawl dataset[1]. The alignment is guided by anchor pairs, which consist of a set of high-quality CN-US and JP-US word translation pairs. The fundamental idea is to use the pretrained U.S. semantic space as the benchmark and align the pretrained Chinese (CN) and Japanese (JP) semantic spaces with it.

Table 1 presents the quality of the aligned word vectors and evaluates their performance using AC@N. AC@N measures the frequency with which the correct answer is ranked within the top N results, providing an indication of the accuracy and effectiveness of the alignment process. For example, in a JP-US alignment task, if the Japanese word "パソコン" (computer) is correctly ranked within the top 5 English translations, the model would score an AC@5 for that instance. Indeed, FastText also provides aligned word vectors; however, it does not offer alignment for the JP-US case. Additionally, we compared the CN-US alignment generated by us to that provided by FastText, and our results significantly outperformed FastText's in terms of alignment quality. Finally, we initialized the dual-attention model using the aligned US-CN-JP word vectors and trained it on the combined datasets from the three countries.

Sample size:10000	AC@1	AC@5	AC@10	AC@100
JP-US	0.18	0.33	0.40	0.68
CN-US	0.22	0.39	0.46	0.73
CN-US (by FastText)	0.05	0.11	0.15	0.39

Table 1. Evaluation of the Quality of Aligned Word Vectors

3.3. Mapping Technology to Product Using Concordance Matrices

The product keywords extracted by the dual-attention model are utilized to represent the web portfolio in place of the raw content. The effectiveness of these extracted product keywords will be evaluated and discussed in the subsequent section. We then capture the linkage between technology and product by estimating a technology-to-product matrix A, such that

$$p_i = At_i$$

[1] https://fasttext.cc/docs/en/crawl-vectors.html

$$\min_{A} \sum_{i \in S} (p_i - At_i)$$

where p_i and t_i represent the company's product and technology vectors, respectively, and S represents a set of companies with at least one patent. In this study, the product and technology vectors for a given company are generated by vectorizing its web portfolio (extracted product keywords) and patent portfolio using the aligned word embedding model for the three countries. The econometric indicators presented in the following sections will be constructed based on various technology-to-product matrices.

4. Results

4.1. Training process and web content extraction

In this study, attention weights derived from the dual-attention model were employed to extract product-related web keywords. Given that company webpages are written in multiple languages, we first applied a pre-trained language identification model[2], which returns a discrete probability distribution of the languages present in the document. We retained webpages primarily written in English, Chinese, and Japanese. Besides, unlike English, Japanese and Chinese lack spaces between words. To address this, we employed open-source tokenization tools such as MeCab[3] for Japanese and Jieba[4] for Chinese to segment the original web documents. The collected web content was then further processed by removing stop words and punctuation, ensuring a cleaner dataset for subsequent modeling. After that, we applied cutting and padding strategies to standardize the input format for the model (i.e., a company's web portfolio). The web portfolio for each company was set to 32 webpages, with each webpage consisting of 768 tokens. The training target is a binary classification label, distinguishing 8,495 high-tech companies from 6,562 non-high-tech companies.

4.2. Evaluation of extracted web content

In this subsection, we evaluate the quality of the extracted web keywords through a classification task aimed at discerning high-tech companies. We compare the extracted web keywords with the raw web content and those extracted by KeyBERT, which is an efficient keyword extraction method based on BERT embeddings. To do so, we constructed LDA topic models to generate company vector representations based on the extracted web keywords and the raw web content, with the number of topics set to 16. Additional experiments confirmed that the evaluation results remained stable across different choices of topic numbers. We applied commonly used machine learning classification models,

---

[2] https://fasttext.cc/docs/en/language-identification.html

[3] https://pypi.org/project/mecab-python3/

[4] https://github.com/fxsjy/jieba

including Logistic Regression (LR) and Random Forests (RF), with default settings. The labeled datasets were split into training and test sets in proportions of 0.7 and 0.3, respectively. The results were evaluated based on four metrics: precision, recall, F1 score, and accuracy. Table 2 presents the training and test results for the extracted web keywords using two different methods, as well as for the raw web content. The relatively low performance of KeyBERT can be attributed to its unsupervised nature for keyword extraction, which makes it susceptible to being influenced by noisy or irrelevant words. Indeed, based on several experiments, a more effective approach was found by combining the dual-attention model with KeyBERT. In this approach, after filtering product-related keywords using the dual-attention model, the filtered keywords are further processed by KeyBERT, which selects the top 40% of the most relevant keywords. As shown in the table, the combined method achieves the highest accuracy, with 0.88 on the training set and 0.78 on the test set. Therefore, we will use the keywords extracted by this combined approach for the subsequent analysis.

Metrics	Train				Metrics	Test			
	Raw	KeyBERT	Dual-attn	Dual-attn+KeyBERT		Raw	KeyBERT	Dual-attn	Dual-attn+KeyBERT
(1) LR					(1) LR				
Precision	0.69	0.71	0.73	0.74	Precision	0.68	0.69	0.72	0.73
Recall	0.68	0.69	0.73	0.73	Recall	0.67	0.67	0.72	0.73
F1 score	0.68	0.68	0.73	0.73	F1 score	0.67	0.66	0.71	0.73
Accuracy	0.68	0.69	0.73	0.73	Accuracy	0.67	0.67	0.72	0.73
(2) RF					(2) RF				
Precision	0.73	0.79	0.83	0.88	Precision	0.71	0.73	0.75	0.78
Recall	0.73	0.78	0.83	0.88	Recall	0.71	0.73	0.75	0.78
F1 score	0.73	0.78	0.83	0.88	F1 score	0.71	0.73	0.75	0.78
Accuracy	0.73	0.78	0.83	0.88	Accuracy	0.71	0.73	0.75	0.78

Table 2. Comparison of the classification results

4.3. Evaluation of technology-to-product matrices

In this study, we estimate several tech2prod matrices by minimizing the sum of squared residuals, where the residuals represent the differences between the actual and transformed product vectors. These matrices are trained separately for different countries and company types to capture context-specific technology-product relationships. The three country types are the U.S. (US), Japan (JP), and China (CN), while the three company types include high-growth, established, and multinational

companies. High-growth companies are selected based on different stock market segments, such as the Growth Market for Japan, NASDAQ for the U.S., and ChiNext and STAR Market for China. On the other hand, established companies refer to mature firms, such as those listed on the main board of Chinese stock exchanges. Finally, multinational companies are identified based on whether a company has overseas divisions, as determined using Moody's DataHub. In total, we train nine tech2prod matrices, corresponding to the combinations of the different country and company types.

During the training process, we split the data into training and test sets in proportions of 0.7 and 0.3. We carefully chose the number of epochs, one of the most critical parameters affecting model performance. The optimal number of epochs was selected to prevent the matrix from overfitting the training dataset. Table 3 presents the evaluation of the estimated technology-to-product matrices. It is important to note that the accuracy is measured by the mean cosine similarity between the true product vectors and the transformed product vectors. This metric reflects the alignment between the predicted and actual product representations.

Metrics	Train			Metrics	Test		
	Accuracy	#Instances	#Epoches		Accuracy	#Instances	#Epoches
(1) US				(1) US			
High-growth	0.89	950	40	High-growth	0.88	408	40
Established	0.87	1108	40	Established	0.85	476	40
Multinational	0.89	1222	30	Multinational	0.87	525	30
(2) CN				(2) CN			
High-growth	0.88	947	30	High-growth	0.86	407	30
Established	0.89	1269	30	Established	0.87	545	30
Multinational	0.86	658	30	Multinational	0.84	282	30
(3) JP				(3) JP			
High-growth	0.92	68	90	High-growth	0.89	30	90
Established	0.94	1153	90	Established	0.89	495	90
Multinational	0.89	364	90	Multinational	0.86	156	90

Table 3. Evaluation of the estimated technology-to-product matrices

5. Comparison of transformation process across countries and firm's types

In this section, we present our empirical results on the difference of technology to product transformation process across countries and firm's type. A fundamental idea of this section is to calculate the difference in a product vector of certain country and firm type, and an imputed product

vector of the same certain and firm type, estimated by a transformation matrix by different country or firm type. Specifically, we calculated the following cosine similarity based IP_Diff (Innovation Process Difference) indices as bellow.

$$IP\_Diff_{t,i,c1-c2} = ArcCOS(P_{t,i,c1}, \widehat{P_{t,i,c1/c2}}) \text{ where } \widehat{P_{t,i,c1/c2}} = A_{t,c2} * T_{t,i,c1}$$

(cross firm type difference within country by industry)

$$IP\_Diff_{t1-t2,i,c} = ArcCOS(P_{t1,i,c}, \widehat{P_{t1/t2,i,c}}) \text{ where } \widehat{P_{t1/t2,i,c}} = A_{t2,c} * T_{t1,i,c}$$

(cross firm type difference within country by industry)

Where

$P\_(t,i,c)$: product vector of type "t", industry "i" and country "c"

$T\_(t,i,c)$: technology vector of type "t", industry "i" and country "c"

$A\_(t,c)$: technology convergence matrix of type "t" and country "c"

The figure 1 shows the results for the difference indices across countries within each firm type for all industries. First, the cross country differences is relatively smaller for Japan and US, as compared to the ones involving China (China-Japan and China-US). Both of Japan and US are developed countries with relatively matured product and factor market conditions, as compared to China which is an emerging country under economic development. Therefore, the innovation process of Japanese firms and US ones are supposed to have similar each other. Second, the cross country difference for multinationals is the smaller than that of other type of firms for all countries. Such firm is operating not only in its home country, but other countries, so that it is influenced not only the economic institutions in its home country, but also others. Accordingly, the difference in its innovation process becomes smaller as compared to that of the other types, operating in its home country only.

(Figure 1)

The figure 2 shows the results for the difference indices across firm types within each country for all industries. The product life cycle theory suggests that more variance in product category is found in early stage of life cycle, as compared to in is matured stage and a type of innovation shifts from product innovation to process innovation (Utterback and Abernathy, 1977). Accordingly, the major player in the market changes from startup firms to established large firms (Klepper, 1997). Therefore, the transformation process from technology to product is different across startup firms and established ones (existing firms and multinationals). Therefore, the difference between existing firms (EX) and multinationals (MU) is supposed to be smaller as compared to other pairs involving growing startup firms (GR). This pattern is found in Japan and US, while the difference between growth firms (GR) and existing firms (EX) is the smallest. It may be due to the fact that China has more turbulent market environment where an exist firm are also entrepreneurial as is the case for growth firm. Or, a

multinational is very different from domestic players, because the China's economic institution is quite distinct in the world. Another finding in the Table 2 is that the difference (around 25-30 degrees) is relatively smaller as compared to that in Figure 1 (around 40-50 degrees). In addition, it is found that the order of the different index varies by country. Therefore, the home country does matter with the characteristics of the transformation process from technology to product.

(Figure 2)

The next question is to what extent such across country difference can be explained by industry structure in each country. Figure 3 shows the cross country difference for existing firms by industry. As is shown in the Figure 1, the difference between Japan and US is relatively smaller than those of the pairs involving China for all industries. It is found that that this pattern is quite consistent across industry. Furthermore, the difference index is very stable over industries within the pair (40-45 degrees for JP-US, around 50 degrees for CN-JP and for CN-US). Therefore, the difference across country for innovation process can be explained by country specific factors, instead of difference by industry and its composition in each country.

(Figure 3)

The Figure 4 shows the difference index between growth firms and existing ones within each country. In contrast to the Figure 2, the index varies by industry, as well as by country. In Japan, this indicator is the highest for all industries, which comes from services, such as wholesale/retail, FIRB (Finance, Insurance, Real estate and Business services) and scientific services. In contrast, large numbers can be found in primary sector (agriculture, forestry and mining), food and petroleum industry for US. The pattern of relatively high indices in primary and manufacturing sector can be found also in China.

(Figure 4)

In the Figure 4, the difference in innovation transformation process varies by industry, and such variation significantly varies by country. More specifically, the cross industry variation in the innovation process indicators in the US is relatively smaller, as compared to those of Japan and China. In contrast, for both Japan and China, the indicators in primary sector, food and beverages and petroleum products are relatively small, while those of service sectors such as wholesale, retail and FIRB are relatively large. The difference in innovation process between startup and existing firms shows that the products of both types of firms are different even if they are using the same technologies.

The product lifecycle theory suggests that this indicator reflect the heterogeneity in the scope of innovation between startup and large firms. Therefore, cross country variation of this indicator shows that the state of innovation competition between startup and large firm is influenced by country specific factors, as well as industry specific ones, particularly in the case of China and Japan. In contrast, when the firm grows as an established one, such cross industry variation cannot be found (in Figure 3), where cross country institutional difference does matter dominantly.

6. Conclusion

This paper presents a new methodology to quantify the innovation process (from technology to product) and to compare it across countries and firm types by industry, using advanced natural language technique. We apply this methodology to the datasets for listed companies in China, Japan and the United States. The results are interpreted by product life cycle theory and the national and sectoral innovation system concepts. It is found that the results are generally consistent to the predictions derived by such theories.

Our empirical analysis shows that the difference of the innovation process in established domestic firms is explained by cross country difference within each industry, and not much cross industry differences are found. In contrast, we can find a great heterogeneity across industry for the comparison between startup firms and established ones. An innovation opportunity, new product development based on existing technology, can be found in the field with greater differences. Our study suggests that there is a great opportunity to learn across countries for new way of innovation transformation of particular type of technology.

However, a major limitation of our study is that we apply our methodology to only three countries, and only qualitative face validation is made. Therefore, it is important to apply this methodology for multiple countries, such as OECD countries, and quantitative assessment of the results by cross country and industry regression analysis with potential factors explaining the differences should be proceeded. In addition, our analysis deals with the innovation process only at firm, and no information of public research institutions such as universities are taken into account. It is important to incorporate scientific progress, particularly to science based industries where technology opportunities are driven by scientific findings. Another venue of our study is to incorporate research paper information into our framework to look into the role of scientific progress on firm' innovation process.

Reference

Carlsson, B., Jacobsson, S., Holmen, A. and Rckne, A. (2002), Innovation systems: analytical and methodological issues, Research Policy, 31(2), 233-245

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2018). Word translation without parallel data. Proceedings of the International Conference on Learning Representations.

Breschi, S., Malerba, F., (1997), Sectoral Innovation Systems: Technological Regimes, Schumpeterian Dynamics, and Spatial Boundaries. In: Edquist, C. (Ed.), Systems of Innovation: Technologies, Institution and Organisations Pinter, London, Washington.

Dorner, M., Harhoff, D., 2017. A Novel Technology-Industry Concordance Table Based on Linked Inventor-Establishment Data. SSRN Electronic Journal.

Ghemawat, P., (2007), *Redefining Global Strategy: Crossing Borders in a World where Differences Still Matter,* Harvard Business School Press, Boston, MA

Hall, P. A. and D. Soskice (2001), *Variety of Capitalism, The International Foundations of Comparative Advantage*, Oxford University Press

Ikeuchi, K., Motohashi, K., Tamura, R., Tsukada, N., 2017. Measuring Science Intensity of Industry Using Linked Dataset of Science, Technology and Industry, RIETI Discussion Paper, 17-E-056.

Johnson, D., 2002. The OECD Technology Concordance (OTC). Patents by Industry of Manufacture and Sector of Use. OECD STI Working Papers, 2002/5, Paris.

Klepper, S., 1997. Industry life cycles. Ind. Corp. Chang. 6 (1), 145–181.

Klevorick, A.K., Levin, R.C., Nelson, R.R., Winter, S.G., 1995. On the sources and significance of interindustry differences in technological opportunities. Res. Policy 24, 185–205.

Kortum, S., Putnam, J., 1997. Assigning patents to industries: tests of the Yale Technology concordance. Econ. Syst. Res. 9 (2), 161–176.

Kinne, J., Axenbeck, J., 2020. Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. Scientometrics 125 (3), 2011–2041.

Libaers, D., Hicks, D., Porter, A.L., 2010. A taxonomy of small firm technology commercialization. In: Industrial and Corporate Change.

Lundvall, B.-Å. (Ed.), 1992. National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning. Pinter Publishers, London.

BÅ Lundvall, B Johnson, ES Andersen, B Dalum (2002), National systems of production, innovation and competence building, Research policy, 31(2), 213-231

Malerba, F., and Orsenigo, L., (1997), Technological Regimes and Sectoral Patterns of Innovative Activities, Industrial and Corporate Change, 6(1) 83-118

Motohashi, K. and Zhu, C. (2023), Identifying technology opportunity using dual-attention model and technology-market concordance matrix, *Technology Forecasting and Social Change*, 197, December 2023,122916

Nelson, R.R. (Ed.), 1993. National Systems of Innovation. A Comparative Analysis. Oxford University Press, Oxford.

Nelson, R.R., Winter, S.G., 1982. An Evolutionary Theory of Economic Change. Belknap Press, Cambridge, London, MA.

Neuhausler, P., Frietsch, R., Kroll, H., 2019. Probabilistic concordance schemes for the re-assignment of patents to economic sectors and scientific publications to technology fields. In: Discussion Papers Innovation Systems and Policy Analysis Nr. 60. Karlsruhe: Fraunhofer ISI.

OECD (1999), *Managing National Innovation Systems*. OECD, Paris. France

OECD (2002) *Dynamising National Innovation Systems*, OECD, Paris. France

OECD (2005), *Governance of Innovation Systems, Volume 1: Synthesis Report*, OECD, Paris, France

Pavitt, K. (1984). Patterns of technical change: towards a taxonomy and a theory. Research Policy, 13, 343–374.

Schmoch, U., 2008. Concept of a Technology Classification for Country Comparisons – Final Report to the World Intellectual Property Organization (WIPO). URL. htt p://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/pdf/wipo_ipc_technology.pdf. (Accessed 4 February 2016).

Utterback, J.M., Abernathy, W.J., 1975. A dynamic model of process and product innovation. Omega 3 (6), 639–656.

Figure 1: Difference index across countries by firm type (for all industries)

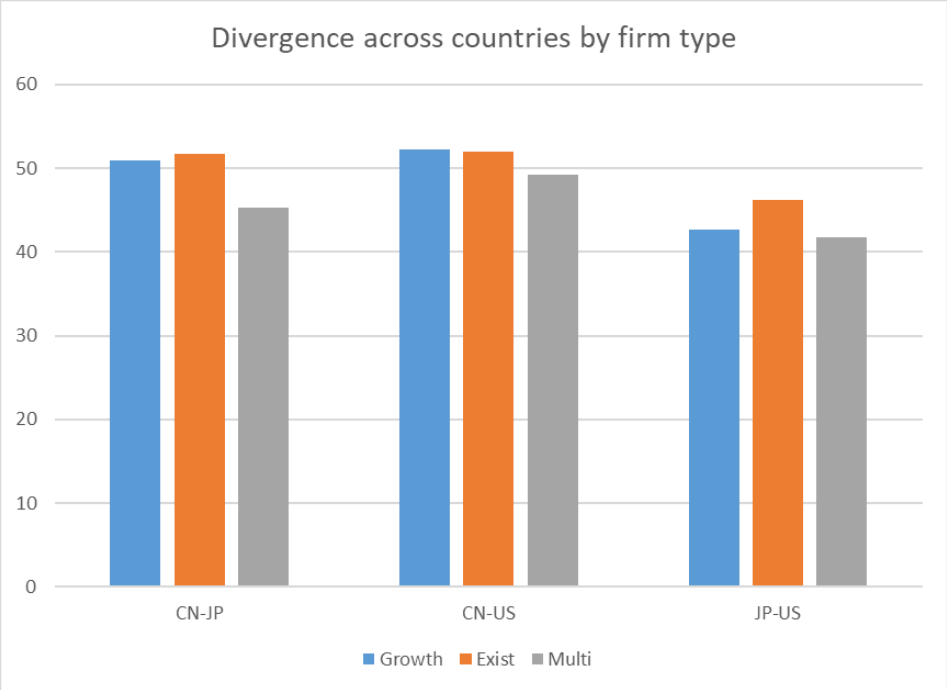

Figure 2: Difference index across firm types by country (for all industries)

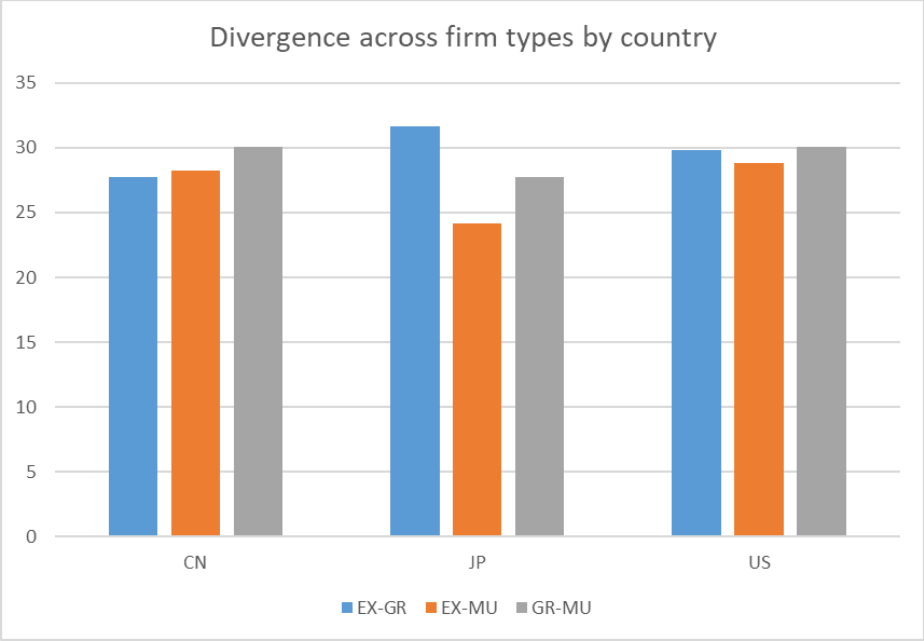

Figure 3: Cross country difference for existing firms by industry

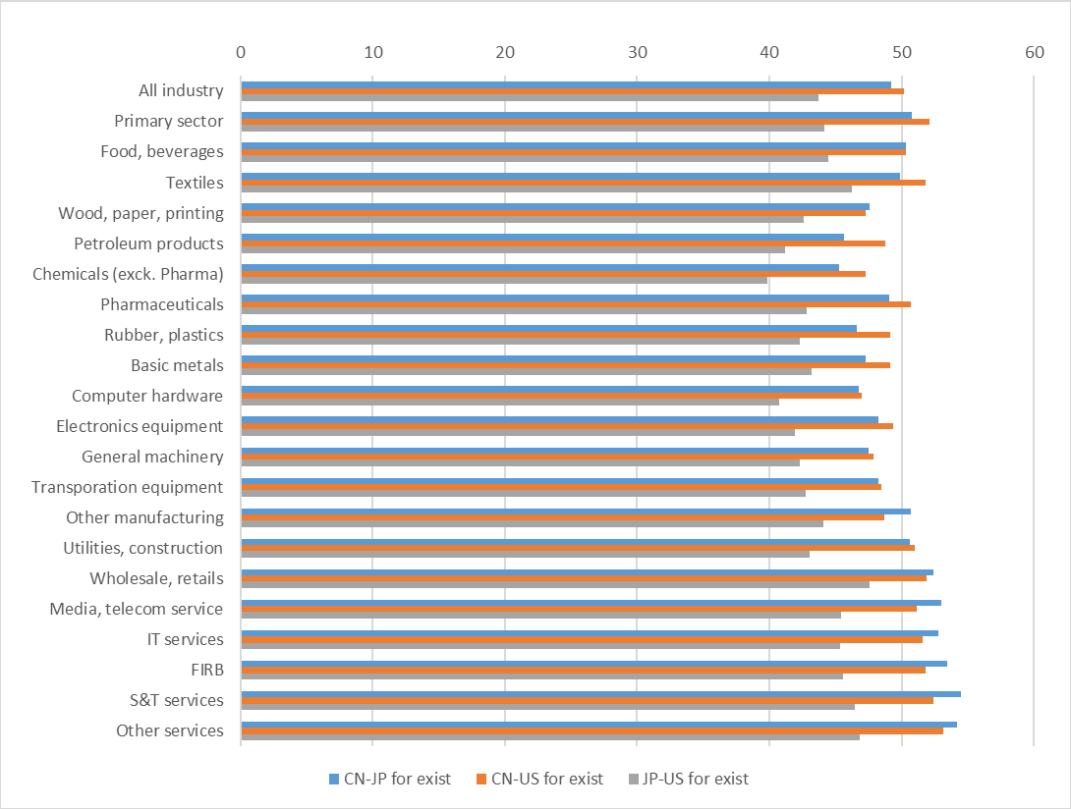

Figure 4: Within country, between growth and existing firm by industry