RIETI Discussion Paper Series 23-E-052

Firm-level Determinants of Cross-border Data Flows: An econometric analysis based on a variable selection technique

ITO, Banri
RIETI

TOMIURA, Eiichi
RIETI

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 23-E-052
July 2023

Firm-level determinants of cross-border data flows: An econometric analysis based on a variable selection technique*

Banri ITO
Research Institute of Economy, Trade and Industry,
Aoyama Gakuin University

Eiichi TOMIURA
Research Institute of Economy, Trade and Industry,
Hitotsubashi University

Abstract

As digital trade involving data transfer including cross-border e-commerce is expanding, firms are actively collecting and utilizing data. This study examines the dynamics of corporate activities related to cross-border data flows based on our questionnaire surveys in 2019 and 2021 of Japanese firms regarding their data collecting activities. The firm entry and exit figures related to foreign data collection activities are quite extreme. Firms active in data collection overseas tend to be more productive. We further explore the variables that are strongly associated with the entry into data collection activities by using the least absolute shrinkage and selection operator technique for variable selection (LASSO). In addition to productivity and firm size, we found that foreign direct investment stock, skill development expenses, intangible assets, and service trade intensity are especially useful for explaining the firms' entry into overseas data collection activities.

Keywords: cross-border data flows, productivity, lasso, firm-level data
JEL classification: F14; F23; O33

* This study is conducted as a part of the Project "Empirical analysis of firms amidst globalization, digitization and the COVID-19 pandemic" undertaken at the Research Institute of Economy, Trade and Industry (RIETI).
The draft of this paper was presented at a DP seminar of the Research Institute of Economy, Trade and Industry (RIETI). I would like to thank the participants of the RIETI DP Seminar for their helpful comments.
This study utilizes the micro data of the questionnaire information based on "the Basic Survey of Japanese Business Structure and Activities" which is conducted by the Ministry of Economy, Trade and Industry (METI), and "Communication Usage Trend Survey" by the Ministry of Internal Affairs and Communication.

1. Introduction

With the rapid increase in digital trade that involves data transfer, such as cross-border e-commerce, data free flow has become important in international transactions. Many reports from international organizations have confirmed this trend.

[1] Reflecting its importance, regulations on cross-border data flows have been introduced in many countries around the world. Our survey of Japanese firms in 2019 show that only a limited fraction of the total population of firms but many of the firms engaged in cross-border data transfer are affected by such regulations (Tomiura et al. 2019). However, the identification of the firm attributes characterizing digital trade has been still limited due to the data availability. Therefore, most previous studies related to cross-border data flows have relied on aggregated cross-country data (Ferracane and van der Marel, 2021; Spiezia and Tscheke, 2020; Gupta et al., 2022). Unlike the previous empirical analyzes based on cross-country data, this study examines the dynamics of corporate activities related to cross-border data transfer based on the firm-level data of our own questionnaire survey in 2019 and 2021 of Japanese firms regarding their data activities.

In international trade literature, the dynamics of entry and exit have been analyzed for exports, imports, and foreign direct investment (FDI), but there is almost no analysis using firm-level data for cross-border data transfer. From our survey on Japanese firms in the manufacturing and service sectors, we found the following notable facts. First, the entry and exit of foreign data collection

[1] OECD (2020) and UNCTAD (2021) are recent notable examples of these reports.

activities are quite intense. Second, cross-border data transfers require high productivity of firms as theoretically formalized and empirically confirmed in the literature on firm heterogeneity and international trade since Bernard and Jensen (1995). This line of argument suggests that firms that engage in data collection activities overseas also engage in domestic activities, and few firms collect data only overseas. In addition, firms that collect data abroad are the most productive, followed by firms that collect data domestically and those that do not.

In line with previous studies on firm heterogeneity and international trade, firm productivity may be one strong determinant, but there may be other significant determinants of entry into offshore data collection. The feature of this study is to identify the determinants of entry into overseas data collection by applying a variable selection technique by LASSO (Least absolute shrinkage and selection operator) regression based on machine learning. As no economic theory models formalizing firms' data collection activities have been established, the application of LASSO is useful for our empirical analysis. The results of LASSO regression succeeded in extracting a handful of specific elements from among 40 various corporate attributes. These are mainly represented by the stock of FDI and the firm size, but the shift to services trade and the corresponding reskilling were also the attributes with strong predictive power. In addition, we also find that these determinants of entry into overseas data collection are distinctively different from the determinants of entry into domestic data collection.

The remainder of this paper is organized as follows: Section 2 elaborates on the data for Japanese firms' data collection activities and the descriptive analyses, Section 3 explains the empirical strategy using a variable selection technique undertaken by LASSO, Section 4 presents the estimation results of the LASSO, Section 5 adds a complementary analysis based on different statistics, and Section 6 highlights conclusions drawn from the analyses.

2. Data

To collect data on corporate activities related to cross-border data transfers, we conducted a questionnaire survey for Japanese firms in 2019 and 2021. We sent the questionnaire to all large- and mid-sized firms, which are defined as firms with 50 or more employees and capital of 30 million yen or more, in manufacturing, wholesale, and information-related service industries. In the first survey, we distributed our survey questionnaires to 19,790 firms in April 2019.[2] We collected responses from 4,227 firms. A second survey for 22,948 firms was conducted in February 2021[3]. We received responses from 6,722 firms. The response rate was 21.6% in 2019 and 29.3% in 2021, respectively, relatively high for an academic survey. For "data" in our survey, we have in mind the continuous and conscious collection of data, such as purchase history of

[2] The "Survey of cross-border data flows of firms" was conducted by the Tokyo Shoko Research Co., Ltd. (TSR) for our research project at the Research Institute of Economy, Trade and Industry (RIETI). We sent the questionnaire to all firms in these sectors covered by the METI's *Basis Survey of Japanese Business Structure and Activities* (*Keizaisangyosho Kigyo katsudo kihon chosa* in Japanese).
[3] The "Survey of globalization and reduced face-to-face contacts during the COVID-19 pandemic" was conducted by TSR for our research project at RIETI. We sent the questionnaire to all firms in the manufacturing and wholesale sectors covered by the METI's BSJBSA.

customers, machine maintenance records, or personal data of employees, through the firm's own daily business operations, but explicitly exclude purchases of prepared databases.[4] The question text about data collection in our survey is as follows.

Please circle the one that best applies to your company's collection of data through its business operations. The "data" to be collected shall refer to raw information (raw data) before being processed or edited into a format such as a database.

1. Continuously collecting digital data both in Japan and overseas

2. Digital data is continuously collected in Japan, but not overseas

3. Digital data is continuously collected overseas, but not in Japan

4. Digital data is not consciously and continuously collected both in Japan and overseas.

The detailed results of our surveys are summarized in Tomiura et al. (2020) for the 2019 survey and Tomiura et al. (2021) for the 2021 survey. The distribution of responding firms to the question on data collection in each survey is shown in Table 1. The largest number of respondents are those that do not continuously collect data, followed by firms that are engaged in data collection only

[4] While the exact definition of "data" has not been established in this context, there are several discussions on the conceptualization. For example, see Gonzáles and Jouanjean (2017).

in Japan, and the fewest are firms that are engaged in overseas data collection.[5] This ordering is observed in both surveys. However, it is noteworthy that the percentage of firms engaged in data collection has dramatically increased from 30% to over 50%. In particular, the percentage of firms engaged in overseas data collection has doubled. The drastic changes in the last two years raise a question about what kind of firms started collecting overseas data.

Table 1. Response distribution regarding data collection activities

	2019		2021	
No dataflows	2891	71.8%	2967	44.7%
Domestic	691	17.2%	2096	31.6%
Dom&Overseas	443	11.0%	1574	23.7%
Total	4025	100.0%	6637	100.0%

The feature of this paper is to identify the corporate attributes that encourage or hinder entry into overseas data collection by capturing changes between two points in time regarding the data collection activities of firms. For this reason, we form a panel dataset of firms that consistently responded to surveys conducted at two different points in time. As a result, we were able to construct balanced panel data for 2,108 firms.

The next step is to cross-tabulate these 2,108 firms for the two years. Panel (a) in Table

[5] Since there are only a limited number of firms engaged in data collection only overseas, these are merged with the firms engaged in both domestic and overseas data collection.

2 shows the results in total with the 2019 status in the rows and the 2021 status in the columns. Firms on the diagonal line of the table, that is, firms with no change in status, account for a total of 52.2% (40.9+6.5+4.8), but the share of these firms is not necessarily high. The remaining half of the firms have rather changed their status, and more than 30% (21.1+11.1) of the firms have started collecting data. On the other hand, only 10% (6.9+2.6) of firms have withdrawn from data collection. Focusing on overseas data collection, 14.2% (11.1+3.1) of firms have newly entered over the past two years, which is significantly higher than the 5.6% (2.6+3.0) of firms that have withdrawn. One may expect that the distribution is different according to respondents' industry affiliation. Panel (b) shows the distribution for manufacturing while Panel (c) displays that for non-manufacturing. Over the two periods, a higher proportion of manufacturing firms engaged in overseas data collection than non-manufacturing firms. However, no significant difference in distribution is observed, and therefore the dynamics of entry into overseas data collection by the manufacturing firms seems to be almost the same as that of the non-manufacturing firms.

Table 2. Distribution in status change from 2019 to 2021

(a) Total	2021			
2019	No dataflows	Domestic	Dom&Overseas	Total
No dataflows	862 (40.9%)	444 (21.1%)	234 (11.1%)	1540
Domestic	146 (6.9%)	136 (6.5%)	66 (3.1%)	348
Dom&Overseas	55 (2.6%)	64 (3.0%)	101 (4.8%)	220
Total	1063	644	396	2108 (100%)

(b) Manufacturing				
2019	No dataflows	Domestic	Dom&Overseas	Total
No dataflows	547 (39.7%)	285 (20.7%)	161 (11.7%)	993
Domestic	96 (7.0%)	92 (6.7%)	45 (3.3%)	233
Dom&Overseas	38 (2.8%)	39 (2.8%)	76 (5.5%)	153
Total	681	416	282	1379 (100%)

(c) Non-manufacturing				
2019	No dataflows	Domestic	Dom&Overseas	Total
No dataflows	315 (43.2%)	159 (21.8%)	73 (10.0%)	547
Domestic	50 (6.9%)	44 (6.0%)	21 (2.9%)	115
Dom&Overseas	17 (2.3%)	25 (3.4%)	25 (3.4%)	67
Total	382	228	119	729 (100%)

Previous studies have often pointed out the relationship between firm's globalization and productivity both theoretically and empirically. Table 3 follows these previous studies and displays the average productivity of firms corresponding to each cell in Table 2 in terms of the logarithm of labor productivity (Y/L) calculated as the value-added per total employee in 2019. The firm-level data for labor productivity is retrieved from METI's Basic Survey of Japanese Business Structure and Activities in 2019 (hereinafter BSJBSA for short, or *Keizaisangyosho*

Kigyo Katsudo Kihon Chosa in Japanese).[6]

Table 3. Status change from 2019 to 2021 and productivity in 2019

(a) Total	2021			
2019	No dataflows	Domestic	Dom&Overseas	Total
No dataflows	1.706	1.747	1.847	1.739
Domestic	1.703	1.864	1.830	1.790
Dom&Overseas	1.878	1.933	1.986	1.944
Total	1.714	1.790	1.879	1.769

(b) Manufacturing	2021			
2019	No dataflows	Domestic	Dom&Overseas	Total
No dataflows	1.684	1.749	1.824	1.727
Domestic	1.685	1.830	1.786	1.761
Dom&Overseas	1.888	1.914	1.916	1.917
Total	1.731	1.803	1.910	1.752

(c) Non-manufacturing	2021			
2019	No dataflows	Domestic	Dom&Overseas	Total
No dataflows	1.744	1.745	1.897	1.765
Domestic	1.738	1.935	1.925	1.852
Dom&Overseas	1.854	1.963	2.201	2.025
Total	1.788	1.843	2.039	1.802

In the panel (a) Total, as shown in the total columns or rows for both years, the ordering of productivity is clearly reflected in the firm's decision of data collection activities. In other

[6] METI conducts this survey annually by imposing legal reporting obligation for all mid- or large-sized firms as defined above. Firms are required to report the previous year's information on a non-consolidated firm basis.

words, firms engaged in overseas data collection have the highest productivity, followed by firms engaged in domestic data collection, and firms not engaged in data collection have the lowest productivity. A similar ordering applies to the relationship between the status change during the two years and the productivity before the change. As shown in the top row, the productivity of firms that have started data collection is higher for firms that have entered overseas data collection (1.847) than for firms that have entered data collection in Japan (1.747). As shown in the second row, among firms engaged in domestic data collection in 2019, the productivity of firms that started collecting data overseas in 2021 (1.830) was higher than the average (1.790), but slightly lower than firms that continued to engage in data collection in Japan. The difference here is not clear. Regarding the firms that were already engaged in overseas data collection in 2019, as shown in the third row, the productivity of firms that continued in 2021 is highest (1.986), while the productivity of firms that have withdrawn from overseas data collection is relatively low (1.933 or 1.878).

Panels (b) and (c) show the average productivity of manufacturing firms and non-manufacturing firms for each corresponding status. Regardless of manufacturing or non-manufacturing firms, as shown again in the top row, regarding the entry of non-data-collecting firms, the productivity of firms entering overseas data collection is the highest, followed by domestic data collection entrants and non-entrants. Our findings on differences in productivity

according to these data-collecting activities suggest that firm attributes in 2019 are associated with the changes in status. Further analysis of what firm attributes, including productivity, influence decisions on data collection activities is provided in the following sections.

3. Empirical strategy

Exploring the determinants of firms' overseas data collection is essential for our understanding of digital trade, but several factors make it empirically difficult. First, it is not clear which firm attributes are theoretically meaningful. A series of studies on firm's globalization have demonstrated both theoretically and empirically that productivity and firm size determine exports and FDI. If we assume that certain fixed costs will be incurred in data collection activities such as data center and server installation, data management, etc., like export and FDI, it is conceivable that highly productive firms that can cover fixed costs can participate in overseas data collection. Based on this idea, productivity, and firm size can be considered as one of the candidates for the determinants. On the other hand, as other factors are not theoretically clear, overlooking important variables in the estimation may cause omitted variable bias or we may end up with including unnecessary explanatory variables. In sum, we have no established orthodox theory models for the firm's data collection decision to guide our empirical analysis. To overcome these shortcomings, we adopt a variable selection technique by LASSO (least absolute shrinkage and selection operator) generalized by Tibshirani (1996). Rather than estimating a reduced model that

adds ad hoc firm attributes, it is possible to identify variables that are strongly related to entry into overseas data collection among various firm attributes and improve the predictive power of the model.

We use LASSO logistic regression model to identify the factors that determine firms' entry into overseas data collection in a binary selection framework. The LASSO logistic regression estimator is expressed as:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{1}{N} \{-y_i(\beta_0 + \mathrm{x_i}\beta') + \ln\{1 + \exp(\beta_0 + \mathrm{x_i}\beta')\}\} + \lambda \sum_{j=1}^{p} |\beta_j|$$

where N is the number of observations, y_i is the outcome variable defined as a binary variable that takes a value of 1 for firms that did not engage in overseas data collection in 2019 but did in 2021, and 0 for firms that did not engage in overseas data collection during this period, β_0 is the constant term, $\mathrm{x_i}$ is the vector of covariates, β is the vector of coefficients, and the last term is the penalty function with λ which is the LASSO penalty parameter.

The reason for using LASSO in this study is to select explanatory variables that are highly associated with this outcome variable. We use 43 explanatory variables for candidate firm attributes. Appendix Table A1 lists them together with their definitions. All listed firm attributes were collected and constructed from the aforementioned METI's BSJBSA. In addition, we

account for industry-specific effects by using two-digit level industry dummy variables. The sample includes 33 industries, and the information services industry which has the largest number of observations is treated as a benchmark. The accurate coefficients obtained by choosing suitable explanatory variables are generated from the values that minimize the λ of the penalty function. In determining the λ, the cross-validation (CV) is computed after estimating the coefficient for each λ for the 100 grids. CV is obtained by randomly splitting the data into 10 folds and regresses between the folds by using the variables in the model for each λ. The average value of mean squared error (MSE) calculated after 10 regressions is called the CV average prediction error or CV function, and the λ with the smallest CV function is chosen.

4. Results

We conduct a LASSO logit regression on the determinants of entry into overseas data collection. In addition, in order to clarify whether the mechanism of entry into overseas data collection differs from entry into domestic data collection, the LASSO regression for binary selection using the same set of 40 covariates is also performed for the entry into data collection in Japan.

First, we show the results regarding the entry of overseas data collection. Figure 1 plots the CV function, and λ, which is determined at the value that minimizes the CV function. In this case, λ_{cv} with the minimum CV function is 0.011, and 11 variables and 10 industry dummies were selected.

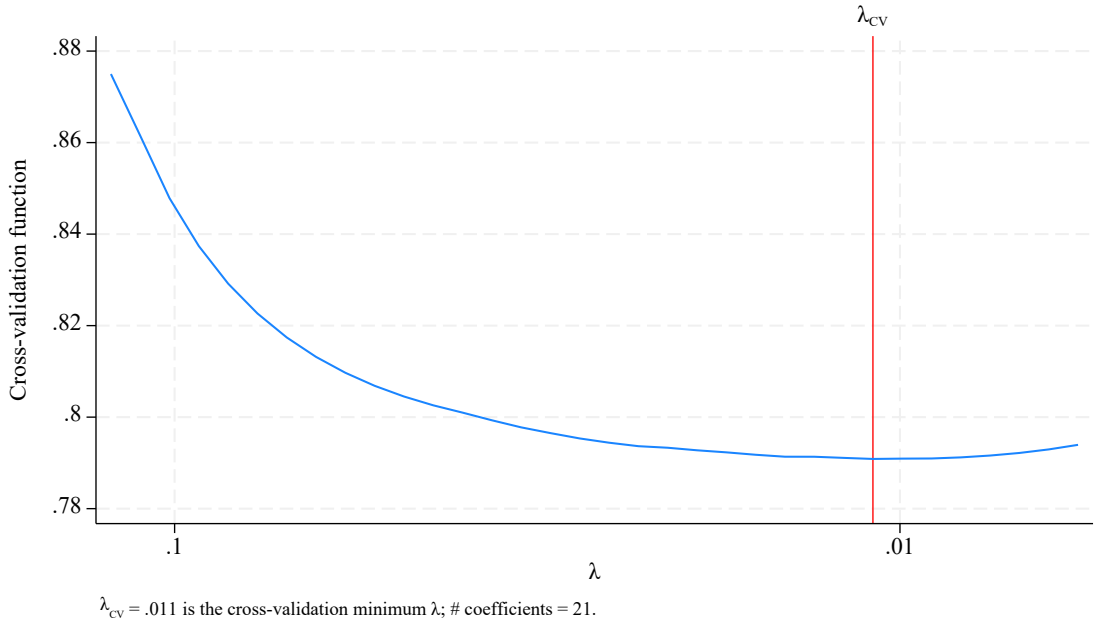

Figure 1. Cross-validation plot from the model for overseas data collection

Table 4 shows the 11 covariates and 10 industry dummies selected by the LASSO procedure. The selection of productivity and sales as firm size is consistent with what has been identified as the main determinant in a series of studies on firm globalization. In addition, the size of intangible assets per employee is also related to entry into overseas data collection. As shown by the fact that FDI stocks and intra-firm service exports are selected, cross-border data transfer is tied to the globalization of firms, especially multinational enterprises (MNEs). In addition, firms that actively develop their human capabilities, import services from unrelated foreign firms, outsource manufacturing to unrelated foreign suppliers, and outsource services to own foreign affiliates tend to start overseas data collection. Regarding the composition of employees, firms with a high share of employees in the planning department at the head office are more likely to start overseas data collection. Given the results of these variable selections, it is thought that firms

that are expanding overseas, are more service-oriented within the firm, and are more active in developing their skills will be more likely to participate in overseas data collection. Regarding the industry dummy variables, 10 industries out of the 32 dummies were found to have a significant difference from the information service industry. Since a positive sign can be found even in the manufacturing industries such as rubber products and electrical machinery, it is not necessarily the case that firms in the information service industry, which are likely to be digitized, are leading the way in overseas data collection. Figure 2 visualizes how these variables are selected in response to λ. The vertical dashed line denotes $\lambda_{\mathrm{cv}} = 0.011$ with the minimum CV function. In particular, it is strongly related to the size of foreign investment. Figure 2 visualizes how these variables are selected in response to different values of λ. Paths of each variable mean that variables with earlier entries have stronger predictive power, indicating that they are important variables to be included in the model. It is clear from this figure that the predictive power of the size of the FDI stock is outstanding.

This finding of strong relation with FDI is consistent with our previous reports from our survey. In the 2019 survey, we ask about the counterparts in cross-border data transfer and find that the data transfers between overseas affiliates and their MNE parents occupy the majority. This finding of strong FDI effect also indicates that the issue of cross-border data flow in general and the impact of regulation on cross-border data transfer, in particular, has still been mainly a concern of multinationals, not for the universe of firms including small-sized domestic firms.

Table 4. Selected determinants of overseas data collection

	Lasso logit	
	Coef.	Odds ratio
lnY_L	0.014	1.014
lnSales	0.061	1.063
lnN_L	0.060	1.061
lnFDI	0.540	1.717
Age	-0.028	0.972
CB_int	0.031	1.031
L_p	0.022	1.022
ExS_intra	0.167	1.181
ImS_arms	0.017	1.017
FO_p	0.093	1.098
FI_s	0.012	1.012
Mfg of lumber and wood products	0.002	1.002
Mfg of plastic products	0.013	1.013
Mfg of rubber products	0.070	1.072
Mfg of fabricated metal products	0.021	1.022
Mfg of electrical machinery	0.034	1.035
Mfg of info and com electronics equip	-0.019	0.981
Mfg of transportation equipment	0.077	1.080
Service incidental to internet	0.021	1.021
Wholseale trade (apparel)	0.033	1.034
Scientific research institutions	0.055	1.056
Constant	-1.792	0.167

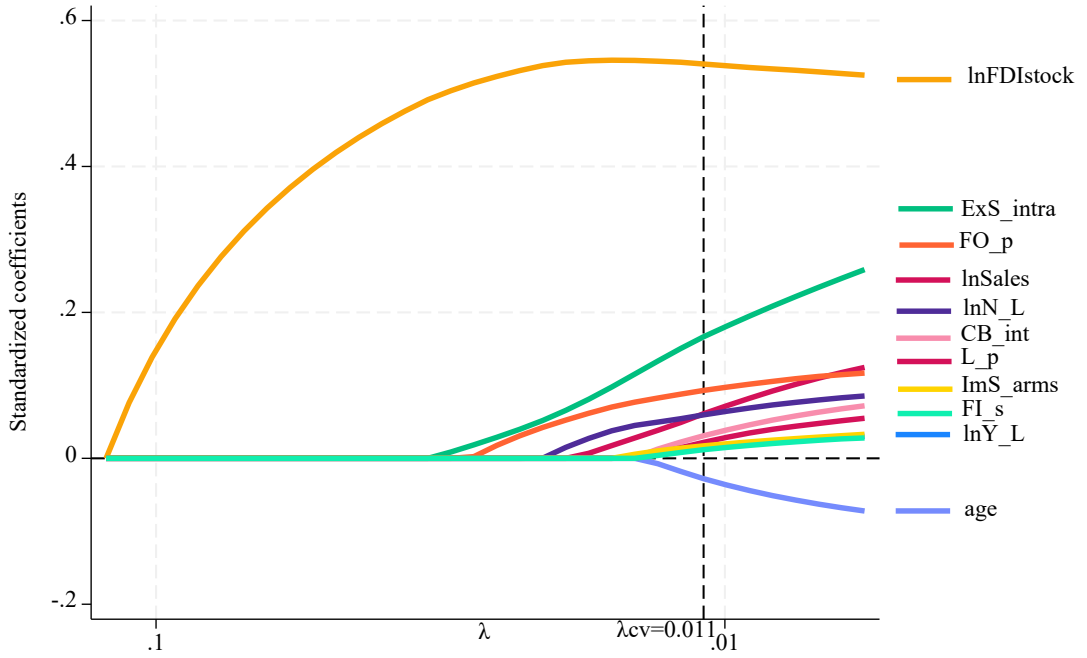

Figure 2. Coefficient path for the model of overseas data collection

Next, we discuss the results of similar LASSO regressions for the binary selection of domestic data collection to compare with that of overseas data collection. In this case, for the binary variable, we assign 0 to firms that did not collect data in 2019 and 1 to firms that entered domestic data collection in 2021. Again, the 43 variables in Appendix Table A1 are applied to candidate explanatory variables. The coefficient of the penalty function, λ was determined to be 0.023 as shown in Figure 3, and four effective coefficients were detected. Those coefficients are shown in Table 5. Corporate attributes that have strong predictive power for entry into domestic data collection include the young age of the firm, the high ratio of employees in the restaurant business, and the high intensity of service imports from foreign unrelated firms. However, the

magnitude of these impacts is limited according to the coefficient path in Figure 4. The firm size appears to be the only dominant factor driving the entry into data collection domestically. We also notice the impact of the COVID-19 pandemic on firms in our second survey, as the employment share of restaurant business, which is seriously affected by lockdown policy during the pandemic, in a firm is detected as one of the significant factors predicting the firm's domestic data collection. Taken together, these results show that the entry mechanisms for domestic data collection and overseas data collection are sharply different. In order to enter the overseas data collection, it is suggested that corporate globalization and service development and corresponding reskilling are required.

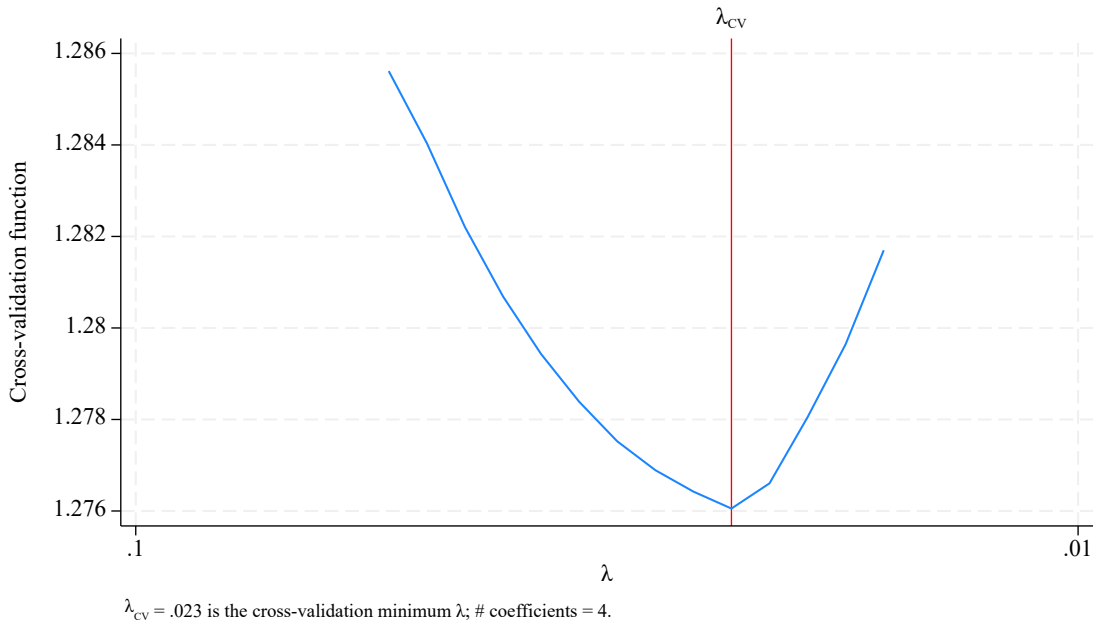

Figure 3. Cross-validation plot from the model for domestic data collection

Table 5. Selected determinants of domestic data collection

	Coef.	Odds ratio
lnSales	0.136	1.146
Age	-0.009	0.991
L_rest	0.012	1.012
ImS_arms	0.017	1.017
Constant	-0.664	0.515

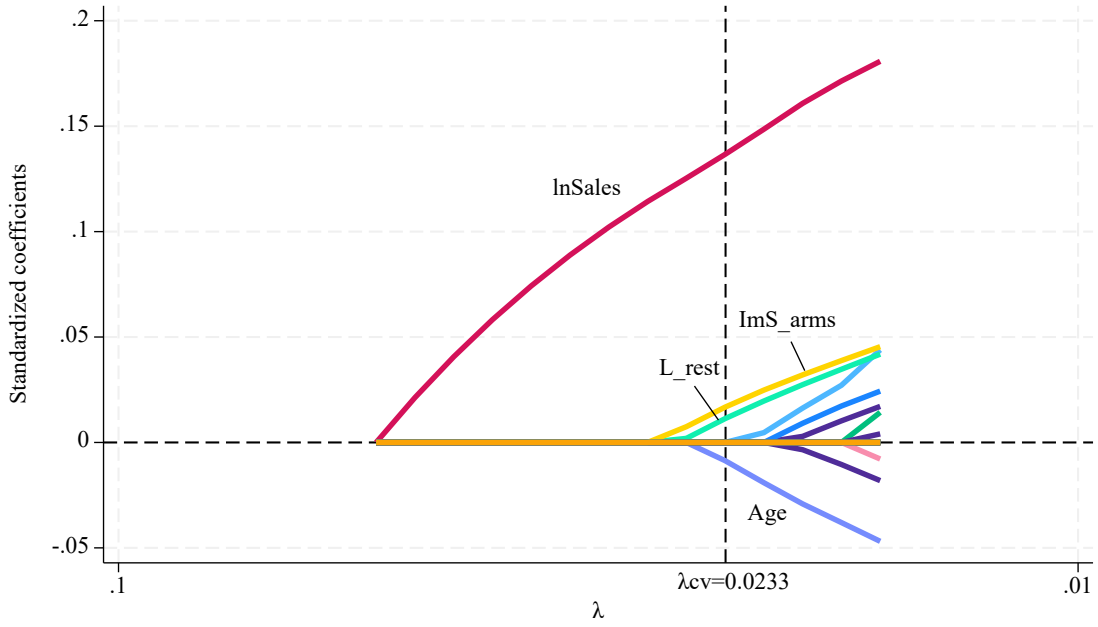

Figure 4. Coefficient path for the model of domestic data collection

5. Complementary analysis

5.1. Additional variables of digitization

Variable selection by LASSO regression in our research relies on 43 variables collected from METI's BSJBSA. Although BSJBSA has many variables related to corporate financial

information and organizations, it has relatively limited digital-related variables. Given that data collection activities are likely to be closely related to the adoption of other digital technologies, we should also add indicators for the digital technologies. In this section, we conduct a complementary analysis to check whether adding digital-related indicators to the candidate variables for the LASSO does not change the main results. Communication Usage Trend Survey (hereafter CUTS) by the Ministry of Internal Affairs and Communications in 2019 is used for this supplementary digital-related variables. Unfortunately, the roster information of CUTS is not available, so we tried to combine it with the BSJBSA by using the information on capital, sales, operating income, and the prefecture where the corporate headquarters are located. As a result, although the number of sample firms that were successfully combined with CUTS is 117, it was possible to add the following digital-related indicators to the candidate variables for these firms: the status of the introduction of artificial intelligence, the status of the introduction of cloud services, the status of the introduction of telework, the usage rate of telework, and the status of data security measures. Table 6 shows selected variables resulting from LASSO regression using additional variables from CUTS. We found that there is not much change from the main result. Indicators of FDI and international trade continue to be relevant to entry into overseas data collection.

Table 6. Selected determinants of overseas data collection with additional covariates

	Lasso logit	
	Coef.	Odds ratio
lnFDI	0.349	1.418
Ex_arms	0.052	1.053
FDI_nb_int	0.020	1.020
Constant	-1.485	0.227

5.2. Determinants of withdrawal from data collection

As shown in Table 2, a certain number of firms have also experienced withdrawal from data collection. To see if there is any difference from the entry mechanism examined in the previous section, we perform the variable selection by LASSO using the same candidate variables used in the entry analysis. In the case of withdrawal, out of the 220 firms engaged in overseas data collection in 2019, shown in the third row of Table 2, 119 firms withdrew from overseas data collection while 101 firms continued in 2021. We examined the determinants of withdrawal by using the binary variable that is coded with 1 for the withdrawing firms. As displayed in Table 7, all selected variables show negative signs, indicating that firms with lower FDI stock, intra-firm exports, and share of the number of overseas affiliates tend to stop collecting overseas data. The results suggest that the withdrawal from data collection is mainly explained by the reduced commitment of MNEs abroad.

Table 7. Selected determinants of withdrawal from overseas data collection

	Lasso logit	
	Coef.	Odds ratio
lnFDI	-0.289	0.749
Ex_intra	-0.134	0.875
FDI_nb_int	-0.140	0.869
Mfg of plastic products	-0.008	0.992
Constant	0.168	1.182

As shown in the second row of Table 2, there are also a certain number of withdrawals from domestic data collection. We examine whether how the withdrawal mechanism differs using a binary variable coded with 0 for 136 firms that continued to collect data in Japan and 1 for firms that withdrew from domestic data collection in 2021. Unlike withdrawal from overseas data collection, it seems that various corporate and industrial attributes are related to withdrawal from domestic data collection. In particular, the magnitude of the impact of lower profit margins is remarkably large. The deterioration of the business environment due to the COVID-19 crisis seems to have a strong impact on the decline in the profit margin, resulting in withdrawal from domestic data collection. Non-introduction of cloud services is also strongly related to withdrawal. It is suggested that the withdrawal from data collection is closely related to the use of services related to data storage.

Table 8. Selected determinants of withdrawal from domestic data collection

	Lasso logit	
	Coef.	Odds ratio
lnY_L	-0.033	0.968
HQ_int	-0.007	0.993
F_ratio	0.014	1.015
L_ib	-0.051	0.950
Im_intra	0.005	1.005
ExS_arms	-0.026	0.974
ExS_intra	-0.040	0.961
ImS_arms	-0.018	0.982
FO_p	0.047	1.048
Profit	-0.392	0.676
Cloud	-0.126	0.881
Mfg of food	-0.122	0.885
Mfg of beverages, tobacco and feed	-0.023	0.977
Mfg of textile products	0.127	1.135
Mfg of ceramic products	0.087	1.091
Mfg of non-ferrous metals	-0.026	0.975
Mfg of info and com electronics equip	0.050	1.051
Wholesale trade (building materials)	0.099	1.104
Misc wholesale trade	-0.120	0.887
Constant	0.041	1.041

6. Conclusions

With the rise of digital trade, the collection and utilization of data has been becoming more significant. Collecting data in foreign markets is essential for gaining access to digital export markets, but it is not clear what kind of firm attributes encourage or hinder overseas data collection. We used variable selection techniques by machine learning-based LASSO regression to discover firm attributes that strongly determine entry into overseas data collection. Our results

show that the multinationals making FDI and the service-oriented firms tend to start collecting data overseas. The reason why FDI is the main reason for entry into overseas data collection is that the establishment of overseas bases such as data centers and servers and/or the intensive information sharing between subsidiaries and parent headquarters is indispensable for maintaining global activities of MNEs. The importance of servitization is shown in our LASSO results by service trade, stock of intangible assets, and overseas outsourcing of manufacturing tasks. As another noteworthy result, skill development within firms will be required in order to shift from conventional trade in goods and services to digital trade.

All these findings on the significant determinants of cross-border data transfer show that globalization, especially FDI, is among the critical precursors of overseas data collection activities. We also find that the servitization of corporate activities and skill upgrading of employees are important for entry into overseas data collection. The introduction of regulations on cross-border data flows should have serious impact on our economies, as these service-intensive multinationals with skillful employees tend to be large, productive, innovative, and pay high wages in many countries. While this paper focuses on the determinants of the entry into data collection among various firm attributes observed before the entry, it will be informative to discuss which firm attributes will be affected by cross-border data flows if relevant data are collected in future independent studies.

References

Bernard, A., & Jensen, J. 1995. Exporters, jobs, and wages in U.S. Manufacturing, 1976–1987. Bookings Papers on Economic Activity, Microeconomics, 26: 67–119.

Ferracane, M. F. & van der Marel, E. 2021. Regulating Personal Data: Data Models and Digital Services Trade. Policy Research Working Paper; No. 9596. World Bank, Washington, DC.

Gonzáles, L., & Jouanjean, M. 2017. Digital trade: Developing a framework for analysis, OECD Trade Policy Papers No.205, OECD Publishing.

Gupta, S., Ghosh, P., & Sridhar, V. 2022. Impact of data trade restrictions on IT services export: A cross-country analysis, *Telecommunications Policy* 46(9), 102403.

Organisation for Economic Cooperation and Development (OECD). 2020. Measuring the economic value of data and cross-border data flows, OECD Digital Economy Papers No.297.

Spiezia, V. & Tscheke, J. 2020. International agreements on cross-border data flows and international trade: A statistical analysis, OECD Science, Technology and Industry Working Papers, No. 2020/09, OECD Publishing, Paris, https://doi.org/10.1787/b9be6cbf-en.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*. Series B (Methodological), 58(1), 267–288.

Tomiura, E., Ito, B., & Kang, B. 2019. Effects of regulations on cross-border data flows: Evidence from a survey of Japanese firms. Discussion Paper No.19-E-088, Research Institute of Economy, Trade, and Industry, Tokyo

Tomiura, E., Ito, B., & Kang, B. 2020. Characteristics of firms transmitting data across borders: Evidence from Japanese firm-level data. Discussion Paper No.20-E-048, Research Institute of Economy, Trade, and Industry, Tokyo.

United Nations Conference on Trade and Development (UNCTAD). 2021. *Digital Economy Report 2021 Cross-border data flows and development: For whom the data flow*, UN.

Appendix Table A1. Covariates list

Variables	Description
ODC	1 for firms that did not engage in overseas data collection in 2019 but did in 2021, and 0 for firms that did not engage in overseas data collection during this period
lnY_L	Log of labor productivity (value-added per employee)
lnSales	Log of total sales
HQ_int	Headquaters intensity measured as the N of employees in headquarters over total employee
lnK_L	Log of tangible fixed asset per employee
lnN_L	Log of intangible fixed asset per employee
RD_int	R&D expenditures over total sales
Patent	The N of patents
SGA_int	Selling, general and administrative expenses over total sales
RegL	The N of regular workers over the total N of workers
lnFDI	Log of overseas investment in stock
FDI_nb_int	The N of foreign affiliates over total N of domestic plants and foreign affiliates
Age	Firm age since established year
ICT_int	Information processing/communication costs over total sales
F_ratio	Foreign capital ratio
CB_int	Costs for capacity-building over total sales
DE	Debt to equity ratio
LB	Debt to asset ratio
L_p	The N of employee in Planning Dept. of HQ over the N of empoyees in HQ
L_ip	The N of employee in Information Processing Dept. of HQ over the N of employees in HQ
L_rd	The N of employee in R&D Dept. of HQ over the N of employees in HQ
L_ib	The N of employee in International Business dept. of HQ over the N of employees in HQ
L_aap	The N of employee in Administration/Accounting/Personnel dept. of HQ over the N of employees in HQ
L_mmeg	The N of employee in Manufacturing/Mining, Electricity/Gas Division of HQ over the N of employees in HQ
L_com	The N of employee in Commercial Division of HQ over the N of employees in HQ
L_rest	The N of employee in Restaurant Business Division of HQ over the N of employees in HQ
L_serv	The N of employee in Service Business Division of HQ over the N of employees in HQ
L_info	The N of employee in Information Service Division of plants over the total N of employees
L_other	The N of employee in other divisions of plants over the total N of employees
Ex_intra	Intra-exports in goods over total sales
Ex_arms	Arms's-length exports in goods over total sales
Im_intra	Intra-imports in goods over total sales
Im_arms	Arms's-length imports in goods over total sales
ExS_intra	Intra-exports in services over total sales
ExS_arms	Arms's-length exports in services over total sales
ImS_intra	Intra-imports in services over total sales
ImS_arms	Arms's-length imports in services over total sales
FO_p	Costs for foreign outsourcing (to unrelated foreign suppliers) in production tasks over total subcontract cost for productio
FI_p	Costs for foreign inourcing (to related foreign suppliers) in production tasks over total subcontract cost for production
FO_s	Costs for foreign outsourcing (to unrelated foreign suppliers) in service tasks over total subcontract cost for services
FI_s	Costs for foreign inourcing (to related foreign suppliers) in service tasks over total subcontract cost for services
Tech_ex	Net technology exports (incl patent, utility model, design right, and copyright)
Profit	Profit over total sales
Cloud	1 for firms that introduce cloud services, and otherwise 0
Ind	2digit industry dummy variables