Propagation of Shocks in Individual Firms Through Supplier–Customer Relationships

Ryoji Sato1  ·  Takayuki Mizuno2,3

Received: 20 May 2022 / Accepted: 29 August 2022 / Published online: 1 October 2022
© Springer Japan KK, part of Springer Nature 2022

Abstract
We quantify the magnitude of shock that propagates individual firms through direct supplier–customer relationships. First, we construct machine learning models that predict a firm’s sales growth rate based on corporate attributes and sales information of the firm and its suppliers/customers. The prediction models indicate that not only macroeconomic factors, such as the year and country, but also sales fluctuation of suppliers/customers are important predictors of the firm’s sales growth rate. Second, we plot the change in the predicted sales growth rates in accordance with those of suppliers/customers using a partial dependence plot. Thus, we quantify how much a firm’s sales growth rate changes in accordance with the changes of its suppliers/customers, namely, the magnitude of shock propagation. Finally, we verify the magnitude of shock propagation by comparing it with the sales growth rate of firms that have suppliers/customers negatively impacted by Hurricane Sandy in the U.S. in 2012. The comparison indicates that there is no significant difference between them and further demonstrates that we can simulate how much the shock that occurred in the disaster-affected firms propagates to their transaction firms.

Keywords  Supply chain  ·  Propagation of shock  ·  Sales growth rate  ·  Machine learning

1 Property Risk Engineering Department, Tokio Marine dR Co., Ltd., 1-5-1, Otemachi, Chiyoda-ku, Tokyo, Japan
2 Information and Society Research Division, National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, Japan
3 Department of Informatics, The Graduate University for Advanced Studies, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, Japan
1 Introduction

The loss of a firm’s production capacity due to accidents or natural disasters can propagate through supplier–customer relationships to both upstream and downstream firms [1–3]. Many firms were compelled to suspend operations and reduce production due to supply chain disruption by the Great East Japan Earthquake in 2011 (Nikkei Newspaper, May 3, 2011). In the supply chain networks, individual firm shocks are widely diffused through hub firms with a very large number of links. The diffused shocks will circle back around and return to the original firm. It was pointed out that such a swinging back phenomenon prolongs the deterioration of the original firm’s performance [5]. In addition to such natural disasters, the increase in tariffs in the US–China trade war [4, 6, 7] and the suspension of economic activities caused by the worldwide COVID-19 pandemic [8–11] caused economic losses to both upstream and downstream sectors. Several recent studies have substantiated the shock propagation in supply chains using big data of the supplier–customer relationships and calculating the coefficients of ordinary least squares (OLS) regression as a measure [12–14]. These studies used the firms’ sales growth as the objective variable and the presence (or number) of damaged firms among their suppliers/customers as the main explanatory variable.

This approach, however, does not quantitatively reveal how much a firm’s sales growth rate decreases in accordance with the decrease in that of its suppliers/customers. One reason is that the explanatory variable is not the suppliers'/customers’ sales growth but the presence of damaged firms among the suppliers/customers. Another reason is that OLS focuses not on the accuracy of the regression but only on the value of the regression coefficients. Here, supervised machine learning methods can construct regression models that ensure generalized performance, even for out-of-sample data. By analyzing these regression models, we can clarify how much the sales growth of suppliers/customers contributes to that of a focused firm.

Our study aims to clarify the magnitude of shock propagation quantitatively through the supplier–customer relationships. Since corporate sales are determined by a complex combination of several macroeconomic factors and individual corporate factors, it is difficult to predict fluctuation in sales with high accuracy using a simple prediction model with a small number of parameters. We use supervised machine learning methods to construct models that predict the firm’s sales growth rate based on various parameters, such as corporate attributes and sales information of the firm and its suppliers/customers. The prediction models indicate that not only macroeconomic factors such as the year and country but also a firm’s fluctuation in sales are important to predict that of its transaction firms. We then plot the change in the predicted sales growth rate in accordance with that of the suppliers/customers using partial dependence plots (PDP). Thus, we quantify how much a firm’s sales growth rate changes in accordance with the change in that of its suppliers/customers, namely, the magnitude of shock propagation.

Our manuscript is organized as follows. Section 2 defines shock propagation through supplier–customer relationships in this study and describes how to
construct the prediction models using supervised machine learning methods. We also explain how to verify the magnitude of shock propagation by comparing it with the sales growth rate of firms that have suppliers/customers damaged by an actual disaster event. Section 3 describes the dataset we used and how to make data subsets for investigating shock propagation by industry type and by direction of propagation. Section 4 shows the prediction performance of the constructed models and the contribution of suppliers'/customers’ information in the models. Finally, we show the magnitude of shock propagation derived from the models and demonstrate that we can simulate the magnitude of shock propagation owing to a disaster event. We make our concluding remarks in Sect. 5.

2 Method

2.1 What is Shock Propagation Through Supplier–Customer Relationships?

In the manufacturing industry, accidents originate from human operation errors, and disasters such as typhoons, floods, and earthquakes result in damage to buildings and facilities, electric outage, and gas supply interruption. This leads to the termination of the supply of materials and parts to customers, which then affects the production of firms that are not directly damaged. Similarly, the degradation of customers’ production activities also degrades the suppliers’ production activities. We define such marked degradation of corporate activities as “shock.”

When considering supply chain networks, suppliers and customers can be described as being connected by links. Shock propagates from the supplier’s node to the customer’s node and vice versa when the production activity of either the supplier or the customer is degraded. Previous studies [12–14] have proved that both upstream and downstream propagation is important for understanding shock in supply chain networks. Therefore, we quantify the shock propagation as downstream and upstream (i.e., from suppliers and customers) in this study.

Degradation of corporate activities does not always occur as a result of damaged buildings and facilities when the firm belongs to the non-manufacturing industry and previous studies often excluded non-manufacturing firms from their analyses [12, 14]. Provided that we do not restrict the target events to physical and exogenous events such as accidents and natural disasters, the degradation of corporate activities occurs in the non-manufacturing industry and the shock can also propagate to connected firms. Therefore, our study quantifies the shock propagation for both firms belonging to the manufacturing industry and firms belonging to the non-manufacturing industry.
### 2.2 Prediction Models Using Supervised Machine Learning Methods

Let \( x_i^{(n)} \) be the annual sales of the \( i \)-th firm \((i = 1, 2, \ldots, N)\) in the year \( n \) and let \( \ln x_i^{(n)} \) be the natural logarithm of \( x_i^{(n)} \) and \( \Delta \ln x_i^{(n)} := \ln x_i^{(n)} - \ln x_i^{(n-1)} \). We consider the following approximation of \( \Delta \ln x_i^{(n)} \) with the function \( f \) as

\[
\Delta \ln x_i^{(n)} = f\left( x_i^{(n-1)}, \Delta \ln x_i^{(n-1)}, \max \left( x_{i_1}^{(n-1)}, \ldots, x_{i_k}^{(n-1)}, \ldots, x_{i_M}^{(n-1)} \right),\right)
\]

\[
\frac{1}{M} \sum_{k=1}^{M} \Delta \ln x_{i_k}^{(n-1)}, \frac{1}{M} \sum_{k=1}^{M} \Delta \ln x_{i_k}^{(n-1)}, M, \alpha_i, \alpha_{i'}, n
\]

where \( i_k \in (1, 2, \ldots, N) \) \((k = 1, 2, \ldots, M)\) is the firm that is linked with the \( i \)-th firm as its Tier 1 supplier/customer, \( M \) is the number of suppliers/customers of the \( i \)-th firm, and \( \alpha_i \in \mathbb{R}^2 \) is the vector whose components correspond to the characteristics of the \( i \)-th firm: the country and category of industry. Similarly, \( \alpha_{i'} \) is the vector of the characteristics of the \( i' \)-th firm, which has the highest sales in the year \( n-1 \) among the suppliers/customers of \( i \)-th firm, namely, \( i' := \max \bigcup_{k=1,2,\ldots,M} X_{i_k}^{(n-1)} \). \( n \) is the year of the sales growth rate that we approximate, from 2012 to 2016. For example, when we approximate the sales growth rate in 2016, \( n \) is 2016 and \( n-1 \) is 2015. We call the variables that function \( f \) has the explanatory variables. We also call \( \Delta \ln x_i^{(n)} \) the objective variable and the \( i \)-th firm the target firm.

---

**Fig. 1** Schematic of the equation to approximate the sales growth rate of the \( i \)-th firm (which we call the target firm) in year \( n \). One information used in the approximation is the target firm’s own information in year \( n-1 \). The other information is of its Tier 1 suppliers/customers in the years \( n \) and \( n-1 \).
The main explanatory variable in our study is \( (1/M) \sum_{k=1}^{M} \Delta \ln x_i^{(n)} \), which represents the fluctuation in sales (including the shock) of suppliers/customers in year \( n \). \( x_i^{(n-1)} \) and \( \max(x_i^{(n-1)}, x_{i_2}^{(n-1)}, \ldots, x_{i_M}^{(n-1)}) \) are intended to reflect the size of the target firm and its suppliers/customers. \( \Delta \ln x_i^{(n-1)} \) and \( (1/M) \sum_{k=1}^{M} \Delta \ln x_i^{(n-1)} \) are intended to reflect the recent growth trend of the target firm and its suppliers/customers. \( M \) is expected to reflect the difference in the shock propagation depending on the number of suppliers/customers that the target firm has. \( \alpha_i, \alpha_f, \) and \( n \) are intended to reflect macroeconomic growth trends according to a combination of the country, category of industry, and year. Figure 1 shows a schematic of the equation to approximate the objective variable \( \Delta \ln x_i^{(n)} \).

\( f \) is determined by the supervised machine learning method. The first is the least absolute shrinkage and selection operator (LASSO) regression [15], which is a linear regression method in machine learning. The second is CatBoost [16], which is a Gradient Boosting Decision Tree method. Either machine learning method optimizes its parameters to minimize the error between the left and right sides of Eq. (1); we call this function the prediction model. The number of training data is set as four times the validation data and we implement the cross-validation five times.

Furthermore, we use a Partial Dependence Plot (PDP) [17] to visualize how much the sales growth rate of the target firm changes in accordance with that of its suppliers/customers in the prediction models. PDP visualizes high-dimensional functions and can illustrate how the objective variable depends on one explanatory variable in the machine learning model. Let \( f(x_1, x_2, \ldots, x_p) \) be a machine learning model with \( p \) explanatory variables. We consider visualizing the values of the objective variables according to the \( i \in [1, p] \)th explanatory variable \( x_i \). \( F(x_i) \), which represents the PDP, is defined as

\[
F_i(x) = \frac{1}{N} \sum_{j=1}^{N} f(x_j^1, \ldots, x_{i-1}^j, x_i, x_{i+1}^j, \ldots, x_p^j),
\]

where \( x_i^j \) denotes the value of the \( i \)th explanatory variable in the data \( j \in [1, N] \).

### 2.3 Verification of Magnitude of Shock Propagation

We derive the relationship between the sales growth rate of a target firm and that of its suppliers/customers using the supervised machine learning methods and PDP. The relationship is the magnitude of shock propagation. To verify the derived magnitude of shock propagation, we compare it with the shock propagation observed in an actual disaster event. The disaster event we used for comparison is Hurricane Sandy, which struck the East Coast of the United States in October 2012 and caused widespread damage, mainly in New York and New Jersey, by strong winds and storm surges. The total economic loss is estimated at about 77.4 billion USD, which is the fourth-largest loss in the history of meteorological disasters in the United States as of July 2021 [18].
The procedure of this comparison is as follows. First, we define the area considered directly affected by Hurricane Sandy and, conversely, the area that was free from damage by Hurricane Sandy. Following the method of Kashiwagi et al. [14], we define the disaster-affected area as the counties that are categorized as “Very High” by the Federal Emergency Management Agency (FEMA) [19]. Meanwhile, we define counties that are not categorized as “Very High,” “High,” or “Moderate” as non-affected areas (see the appendix for a map of the defined disaster-affected areas and non-affected areas). We then define the firms whose headquarters are in the disaster-affected area and whose sales growth rate is negative as “disaster-affected firms.” We extract the \( i \)-th target firm whose headquarters are in the non-affected area and whose suppliers/customers include the disaster-affected firms. In accordance with Eq. (1), we calculate \( \Delta \ln x_i(n) \) and \( (1/M) \sum_{k=1}^{M} \Delta \ln x_i(n) \) and obtain the relationship between the sales growth rates of the target firm and its suppliers/customers. Finally, we compare this relationship with the magnitude of shock propagation derived from the machine learning models.

### 3 Data

This study mainly uses the Supply Chain Relationships data provided by FactSet Research Systems Inc.\(^1\) This dataset includes a time series of information on pairs of firms that have business relationships, such as a supplier and a customer or a supplier and a distributor (in the following, the distributor is identified with the customer in our study). The dataset maximizes the coverage of the business relationships by referring to direct relationships (relationships disclosed by the reporting firm) and inverse relationships (disclosed not by the reporting firm but by firms that do business with the reporting firm). The dataset’s source is publicly available information, such as securities reports and websites, and it covers only listed firms and firms that have business relationships with them. The business relationships are recorded in units of a parent firm and all transactions between subsidiaries are grouped in their parent firms.

This study also utilizes the dataset of annual sales of firms, names of countries in which the firms’ headquarters are located, and industrial codes, which correspond to levels 1–6, namely, the FactSet Revere Business Industry Classifications System (RBICS) [20]. We tag the firms in the dataset with annual sales in USD and country name. The firms are also tagged with level 2 of RBICS and categorized into “manufacturing” and “non-manufacturing” industries (see the appendix for the industry classification definition).

To investigate the shock propagation by industry type and by direction of propagation, we generate the data subsets by the following steps. We extract pairs of firms with business relationships as supplier–customer at least for some time between 2010 and 2016. This extraction is implemented because some relationships are

---

\(^{1}\) [https://www.factset.com/marketplace/catalog/product/factset-supply-chain-relationships](https://www.factset.com/marketplace/catalog/product/factset-supply-chain-relationships).
missed by incorrect records even when the transactions are actually continuing. We prepare 20,773 target firms with suppliers and 18,935 target firms with customers from those pairs of firms. Provided that there is a mixture of manufacturing and non-manufacturing firms in the suppliers/customers of a target firm, we exclude the target firm from the following analysis. The generated four data subsets are

subset (1): target firms with their suppliers in the manufacturing industry,

subset (2): target firms with their customers in the manufacturing industry,

subset (3): target firms with their suppliers in the non-manufacturing industry, and

subset (4): target firms with their customers in the non-manufacturing industry.

In addition, we generated another data subset as

subset (5): target firms with their suppliers and customers in the manufacturing or non-manufacturing industry.

Data subset (5) includes supplier—target firm—customer linkages and both manufacturing and non-manufacturing firms (data with a mixture of manufacturing and non-manufacturing firms in one hierarchy are excluded). The number of target firms in data subsets (1)–(5) is 17,625, 18,506, 24,405, 21,974, and 26,992, respectively. The objective variable and explanatory variables are prepared for each data subset as described in Sect. 2.2. We compute the cumulative relative frequencies of the countries for target firms and their suppliers/customers, leave the top 80% of the countries as they are, and aggregate the remaining countries as “others.” See the appendix for the summary of statistics and the number of target firms by the country, the industry, and the year in each data subset.

Finally, we remove outliers and oversample the data to improve the imbalances in each data subset. Sales growth rates of firms concentrate around zero and the number of firms is extremely small at the positive and negative tails of the distribution of the sales growth rates. Using such imbalanced data, machine learning models have good performance only in the range of the sales growth rate with high data volume. To address this problem of imbalanced data, we first remove the target firms and their suppliers/customers whose sales growth rate is out of the quartiles as outliers. Second, we apply the synthetic minority over-sampling technique (SMOTE) [21]. For a datum belonging to a minority group, SMOTE randomly selects its neighbor from the $k$-nearest neighbors and interpolates a new datum between the two selected data. The procedure for applying SMOTE to each data subset is as follows. Each data subset is divided into training and validation data vs. test data $= 8:2$. Then, the
training and validation data are divided into the training data and validation data for fivefold cross-validation. Next, we divide each data into eight ranges, such as \( \Delta \ln x_i^{(n)} \in [\pm 0.5k \cdot \sigma, \pm 0.5(k - 1) \cdot \sigma] \) \((k = 1, 2, 3)\) (double-sign corresponds) and outside of \( \pm 1.5\sigma \). SMOTE oversamples the data in each range to align the number of data with the group that has the most data.

### Table 1
The performance of machine learning models obtained using LASSO regression and CatBoost

| Subset no. | prediction algorithm          | RMSE (validation) | RMSE (test) | Correlation (validation) | Correlation (test) |
|------------|-------------------------------|-------------------|-------------|--------------------------|-------------------|
| (1)        | (Baseline)                    | –                 | 0.132       | –                        | –                 |
|            | LASSO                         | 0.123             | 0.121       | 0.429                    | 0.429             |
|            | CatBoost                      | 0.115             | 0.113       | 0.516                    | 0.525             |
|            | CatBoost (w/o suppliers’ variables) | 0.116         | 0.114       | 0.501                    | 0.509             |
| (2)        | (Baseline)                    | –                 | 0.150       | –                        | –                 |
|            | LASSO                         | 0.136             | 0.137       | 0.438                    | 0.429             |
|            | CatBoost                      | 0.129             | 0.129       | 0.509                    | 0.506             |
|            | CatBoost (w/o customers’ variables) | 0.130         | 0.131       | 0.496                    | 0.484             |
| (3)        | (Baseline)                    | –                 | 0.150       | –                        | –                 |
|            | LASSO                         | 0.132             | 0.132       | 0.478                    | 0.495             |
|            | CatBoost                      | 0.123             | 0.121       | 0.564                    | 0.591             |
|            | CatBoost (w/o suppliers’ variables) | 0.124         | 0.123       | 0.551                    | 0.578             |
| (4)        | (Baseline)                    | –                 | 0.170       | –                        | –                 |
|            | LASSO                         | 0.148             | 0.150       | 0.497                    | 0.483             |
|            | CatBoost                      | 0.141             | 0.141       | 0.558                    | 0.560             |
|            | CatBoost (w/o customers’ variables) | 0.143         | 0.143       | 0.538                    | 0.542             |

“Baseline” is the accuracy when all prediction values are set to zero. “CatBoost (w/o suppliers’ variables)” and “CatBoost (w/o customers’ variables)” show the accuracy when suppliers’/customers’ variables (i.e., the third through sixth and eighth arguments in the function \( f \) of Eq. 1) are excluded. “Validation” is the average accuracy of cross-validation with the best hyperparameters. “Test” is the accuracy of applying the model to the test data.
4 Results

4.1 Prediction Performance of Machine Learning Models

Table 1 shows the prediction performance of the machine learning models, which are obtained by applying LASSO regression and CatBoost to data subsets (1)–(4). The performance is shown in Root Mean Square Error (RMSE) and the correlation coefficient between the target firms’ true and predicted sales growth rate. “Baseline” shows the RMSE value when all predicted values are set to zero. The LASSO regression and CatBoost show better RMSE values than the Baseline, and it is concluded that the learnings are successful. We can confirm that the models avoid over-learning because the RMSE values of the prediction with the validation data are similar to that with test data. The correlation coefficients obtained by the LASSO regression and CatBoost with test data are in the range of (0.429, 0.495) and (0.506, 0.591), respectively, and these values also indicate that
the learning is successful. These accuracies also show that the models obtained using CatBoost have higher prediction performance than that obtained using LASSO regression. Linear regression has been used to investigate the significance of regression coefficients in the previous studies mentioned above. On the other hand, our machine learning models indicate that CatBoost is better than linear regression in prediction accuracy.

Next, Fig. 2 shows the prediction result obtained by the CatBoost models when they are applied to the test data. This figure shows that there is a correlation between the target firms’ true and predicted sales growth rates, as shown in Table 1. Meanwhile, the larger the absolute values of the true sales growth rate become, the larger the deviation of the predicted sales growth rate from the true one becomes. In the case of data subset (1), the RMSE value in intervals \((-0.1, 0.0]\) and \((0.0, 0.1]\) are 0.069 and 0.071. On the other hand, the RMSE value for intervals \((-0.2, -0.1]\) and \((0.1, 0.2]\) are 0.116 and 0.122, and for intervals \((-\infty, -0.2]\) and \((0.2, +\infty]\) are 0.189 and 0.187, respectively.

Figure 3 shows the regression coefficients of the model obtained using LASSO regression on data subset (1).

The explanatory variable of interest is the suppliers’/customers’ sales growth rate in year \(n\) (“Salesgrowth_supplier\((n)\)” or “Salesgrowth_customer\((n)\)”), and its regression coefficients show relatively large values. This result indicates that the sales growth rate of target firms decreases in accordance with the decrease
in that of their suppliers/customers; this is consistent with the shock propagation through the supplier–customer relationship. We also confirmed that this tendency similarly holds true for other data subsets. Table 1 shows the performance of the model obtained using CatBoost when the explanatory variables related to suppliers/customers are excluded from the explanatory variables. The prediction performance decreases without explanatory variables related to the suppliers/customers (i.e., the third through sixth and eighth arguments in the function $f$ of Eq. 1), as shown in Table 1. This result indicates that the information about suppliers/customers is significant for predicting the sales growth rate of target firms.

Finally, Fig. 4 shows the importance of explanatory variables of the models obtained using CatBoost on each data subset.

Regarding data subsets (1)–(4), the ranking of the contribution of the explanatory variables tends to be similar to that of the regression coefficients in the LASSO regression models. In the case of data subset (5), the sales growth rate of customers in year $n$ is slightly more significant than that of suppliers, but there is no marked
difference. The significant variables together with the sales growth rate of suppliers/customers are the year $n$ ("predicted_year"), the country in which the target firm is located ("Country_target"), the sales growth rate of the target firm in the previous year ("Salesgrowth_target\(n-1\)"), and the category of industry of the target firm ("Industry_target"). This tendency indicates that the macroeconomic trend, expressed as a combination of a year, a country, an industry, and the firm’s own

**Fig. 5** PDPs drawn from the machine learning models obtained using CatBoost on each data subset. “Average rate of change” indicates the average rate of change in the sales growth rate of the target firms when their suppliers'/customers' sales growth rate varies from −0.1 to 0.1 in steps of 0.02.
recent sales growth, also contribute as much as or more significantly than the suppliers'/customers' sales growth rate.

### 4.2 Magnitude of Shock Propagation Derived from Machine Learning Models

This section describes the magnitude of shock propagation calculated using the machine learning models. Figure 5 shows the PDPs drawn from the machine learning models obtained using CatBoost on each data subset.

The x-axis is the suppliers'/customers’ sales growth rate. It varies from $-0.1$ to $0.1$ in steps of $0.02$ and the average value of objective variables is calculated at each step. We set the number of samples as $500$ and randomly extracted $500$ data from all the training, validation, and test data of each data subset. There is a positive correlation between the target firm’s sales growth rate and that of its suppliers/customers in each data subset, although the slope differs in each interval. This correlation indicates that the sales growth rate of target firms decreases in accordance with the decrease in that of their suppliers/customers through supplier–customer relationships. In data subsets (1)–(4), the average changes in the sales growth rate of target firms are $0.173$, $0.192$, $0.163$, and $0.227$, respectively. This result indicates that, for example, in data subset (1), our machine learning model predicts that the sales growth rate of a target firm decreases by $1.73\%$ on average when that of its suppliers decreases by $10\%$. Provided that these PDPs are

| Subset no. | Statistic | Damaged/non-damaged | 2012   | 2013   | 2014   | 2015   | 2016   |
|-----------|-----------|---------------------|--------|--------|--------|--------|--------|
| (1)       | Mean      | Damaged             | 0.036  | -0.012 | 0.027  | 0.019  | 0.019  |
|           |           | Non-damaged         | 0.040  | 0.037  | 0.042  | 0.003  | 0.020  |
|           | $T$ value | -                   | 0.059  | -2.643*| -1.142 | 1.029  | 0.506  |
|           | $P$ value | -                   | 0.953  | 0.010  | 0.257  | 0.307  | 0.615  |
| (2)       | Mean      | Damaged             | 0.032  | 0.001  | 0.052  | 0.013  | 0.014  |
|           |           | Non-damaged         | 0.03   | 0.032  | 0.045  | 0.01   | 0.014  |
|           | $T$ value | -                   | 0.141  | -2.181*| 0.575  | 0.192  | -0.004 |
|           | $P$ value | -                   | 0.889  | 0.032  | 0.567  | 0.848  | 0.997  |
| (3)       | Mean      | Damaged             | 0.035  | 0.037  | 0.047  | 0.018  | 0.026  |
|           |           | Non-damaged         | 0.051  | 0.062  | 0.064  | 0.039  | 0.037  |
|           | $T$ value | -                   | -1.583 | -2.368*| -1.777 | -1.886 | -1.141 |
|           | $P$ value | -                   | 0.115  | 0.019  | 0.077  | 0.061  | 0.255  |
| (4)       | Mean      | Damaged             | 0.025  | 0.03   | 0.044  | 0.02   | 0.027  |
|           |           | Non-damaged         | 0.044  | 0.046  | 0.049  | 0.036  | 0.037  |
|           | $T$ value | -                   | -2.038*| -1.613 | -0.684 | -1.672 | -1.121 |
|           | $P$ value | -                   | 0.043  | 0.108  | 0.495  | 0.096  | 0.264  |

*Difference between these sales growth rates is statistically significant at the 5% significance level
decomposed into the category of industry, there are no marked differences among
the inclination of the PDPs for each category of industry.

The lower panel of Fig. 5 shows the PDPs drawn from the machine learning
model obtained using CatBoost on data subset (5). The left half of these panels
shows the PDP with the sales growth rate of suppliers as the $x$-axis and the right half
shows the PDP with the sales growth rate of customers as the $x$-axis. The average
rates of change are 0.177 and 0.224, respectively. The inclination of customers’ sales
growth rate is slightly larger than that of suppliers, but the difference is not marked.

### 4.3 Verification of the Magnitude of Shock Propagation Using Disaster Event

This section explains the result of comparing the magnitude of shock propagation
obtained in the previous section with that in an actual disaster event. First, Table 2
shows the sales growth rates of the suppliers/customers averaged over the disaster-
affected area and those averaged over the non-disaster-affected area in each year $n$.

Since Hurricane Sandy struck in Oct. 2012, we measured its impact on the sales
growth rate in 2013. Table 2 shows that the sales growth rate of the firms averaged
over the disaster-affected area is certainly lower than that averaged over the non-affected area in 2013. In particular, in data subsets (1)–(3), the sales growth rate averaged over the disaster-affected area is statistically significantly lower than that in the non-affected area at the 5% significance level. This result indicates that Sandy indeed affected the sales growth rate of the firms in the disaster-affected area. We conclude that Sandy is an appropriate subject for investigating the magnitude of shock propagation. In subset (4), the impact of hurricane appears to be more significant in the sales growth rate of firms a year earlier than in the other subsets.

Next, we obtain the relationship between a target firm’s sales growth rate and that of its suppliers/customers, including the disaster-affected firms. Figure 6 shows the regression lines of the relationship superimposed on the PDP shown in Fig. 5.

The slopes of the regression lines in each data subset (1)–(4) are 0.252, 0.095, 0.361, and 0.708, respectively. Provided these slopes are not markedly different from the average rate of change of the PDP, the magnitude of shock propagation derived from our machine learning models is reasonable. To confirm this, we first check that both the slope of the regression line and the average rate of change of PDP are not zero by a \( t \) test. As a result, only data subsets (3) and (4) contradict the hypothesis that they are zero at the 5% level of significance. For these two data subsets, we implement the \( t \)-test on the hypothesis that the slope of the regression line is equal to the slope at the points plotted in the PDP. As a result, this hypothesis is correct in either data subset at the 5% level of significance. Thus, we conclude that the magnitude of shock propagation derived from our machine learning models is reasonable compared with an actual disaster event. This conclusion indicates that we can simulate how much the shock that occurred in the disaster-affected firms propagates to their transaction firms.

5 Conclusion

This study quantified the magnitude of shock that propagates individual firms through direct supplier–customer relationships. To achieve this aim, we constructed supervised machine learning models that predict the sales growth rate of individual firms by using firm-level big data, which includes supplier–customer relationships, sales, and several corporate attributes. Then, we plotted the change in the predicted sales growth rates in accordance with those of suppliers/customers, which indicated that 16.3–22.7% of the shock on suppliers/customers propagates to their transaction firms. Meanwhile, the models also show that the performance of suppliers/customers is only one of the factors that affect their transaction firm’s performance; the macro growth trends depend on the year, country, and category of industry and the firm’s own recent growth trend has a strong impact. This finding suggests that firms do not need to overly fear the risk of shock propagation from their suppliers/customers compared to the macroeconomic environment surrounding them or their own growth trend. We also found that these machine learning models do not depend on the firm’s industry and direction of shock propagation. Finally, we compared the magnitude of
shock propagation derived from the machine learning models with that observed in an actual disaster event. This comparison showed that our models accurately evaluate the magnitude of shock propagation in accordance with the scale of the natural disaster, indicating that these models work as simulators of shock propagation.

The recent trade war between the U.S. and China and the various economic impacts of COVID-19 are likely to increase risks in the global supply chain. The contribution of suppliers/customers performance in our machine learning models probably increases year by year and thus investigation of the annual change in the contribution is a future task.

Appendix

This appendix firstly shows the map of the defined disaster-affected area and the non-affected area described in Sect. 2.3 (Fig. 7).

We next show the detail of data subsets (1)–(4) described in Sect. 3: the summary of statistics (Table 3) and the number of target firms by country, category of industry, and year (Tables 4, 5, 6, 7).

Fig. 7 This figure denotes the damage level by Hurricane Sandy. The colors in this figure correspond to the impact analysis classification created by the Federal Emergency Management Agency (FEMA). The purple counties indicate areas of “Very High” impact, which we define as disaster-affected areas. Similarly, red counties are “High,” yellow counties are “Moderate,” and counties that are not either of these three colors are defined as non-affected areas by the authors.

© Springer
Table 3  Summary statistics of continuous variables in each data subset (1)–(4)

| Subset no. | Variables | Mean | S.D. | Min      | Median | Max      |
|------------|-----------|------|------|----------|--------|----------|
| (1)        | Objective variable: | 0.001 | 0.112 | −0.273 | 0.004 | 0.275 |
|            | Sales growth rate of target firm (n) | 0.022 | 0.127 | −0.295 | 0.022 | 0.345 |
|            | Explanatory variables: | 0.023 | 0.107 | −0.347 | 0.018 | 0.402 |
|            | Sales growth rate of suppliers (n − 1) | 0.002 | 0.087 | −0.322 | 0.001 | 0.327 |
|            | Sales growth rate of suppliers (n) | 2.872 | 0.805 | −1.678 | 2.857 | 5.434 |
|            | Log of sales of target firm (n − 1) | 3.844 | 0.981 | −1.126 | 4.013 | 5.434 |
|            | Log of sales of suppliers (n − 1) | 0.553 | 0.524 | 0.000 | 0.477 | 2.613 |
| (2)        | Objective variable: | 0.000 | 0.122 | −0.322 | 0.002 | 0.328 |
|            | Sales growth rate of target firm (n) | 0.022 | 0.140 | −0.347 | 0.020 | 0.407 |
|            | Explanatory variables: | 0.023 | 0.091 | −0.295 | 0.020 | 0.345 |
|            | Sales growth rate of suppliers (n − 1) | 0.000 | 0.076 | −0.272 | 0.001 | 0.274 |
|            | Sales growth rate of suppliers (n) | 2.800 | 0.827 | −1.678 | 2.767 | 5.434 |
|            | Log of sales of target firm (n − 1) | 4.229 | 1.037 | −0.301 | 4.561 | 5.434 |
|            | Log of sales of suppliers (n − 1) | 0.623 | 0.525 | 0.000 | 0.602 | 2.041 |
| (3)        | Objective variable: | 0.012 | 0.123 | −0.316 | 0.016 | 0.318 |
|            | Sales growth rate of target firm (n) | 0.032 | 0.129 | −0.303 | 0.032 | 0.359 |
|            | Explanatory variables: | 0.045 | 0.103 | −0.339 | 0.049 | 0.443 |
|            | Sales growth rate of suppliers (n − 1) | 0.023 | 0.096 | −0.346 | 0.028 | 0.396 |
|            | Sales growth rate of suppliers (n) | 2.967 | 0.842 | −1.455 | 2.970 | 5.686 |
|            | Log of sales of target firm (n − 1) | 3.969 | 1.081 | −1.455 | 4.144 | 5.686 |
|            | Log of sales of suppliers (n − 1) | 0.594 | 0.515 | 0.000 | 0.477 | 2.705 |
| (4)        | Objective variable: | 0.016 | 0.138 | −0.349 | 0.018 | 0.398 |
|            | Sales growth rate of target firm (n) | 0.038 | 0.145 | −0.340 | 0.036 | 0.443 |
|            | Explanatory variables: | 0.023 | 0.093 | −0.301 | 0.031 | 0.359 |
|            | Sales growth rate of customers (n − 1) | 0.005 | 0.085 | −0.315 | 0.009 | 0.318 |
|            | Sales growth rate of customers (n) | 2.882 | 0.899 | −1.455 | 2.842 | 5.686 |
|            | Log of sales of target firm (n − 1) | 4.396 | 0.977 | −0.425 | 4.677 | 5.686 |
|            | Log of sales of customers (n − 1) | 0.683 | 0.521 | 0.000 | 0.602 | 2.693 |

n is the year of the sales growth rate that the machine learning model predicts, which takes from 2012 to 2016.
### Table 4  Number of target firms by country in each data subset

| Subset no. | Country       | Freq. | Percent |
|------------|---------------|-------|---------|
| (1)        | United States | 3608  | 20.3    |
|            | Japan         | 3265  | 18.4    |
|            | Taiwan        | 1393  | 7.8     |
|            | South Korea   | 1321  | 7.4     |
|            | India         | 1215  | 6.8     |
|            | others        | 6964  | 39.2    |
| (2)        | United States | 6559  | 26.7    |
|            | Japan         | 3020  | 12.3    |
|            | United Kingdom| 1097  | 4.5     |
|            | China         | 978   | 4.0     |
|            | Canada        | 826   | 3.4     |
|            | others        | 12,055| 49.1    |
| (3)        | United States | 3941  | 21.1    |
|            | Japan         | 3927  | 21.0    |
|            | South Korea   | 1542  | 8.3     |
|            | Taiwan        | 1479  | 7.9     |
|            | China         | 944   | 5.1     |
|            | others        | 6847  | 36.7    |
| (4)        | United States | 5559  | 25.1    |
|            | Japan         | 2951  | 13.4    |
|            | United Kingdom| 989   | 4.5     |
|            | China         | 969   | 4.4     |
|            | Canada        | 771   | 3.5     |
|            | others        | 10,865| 49.2    |
### Table 5
Number of target firms by industry on data subsets (1) and (2), which include the manufacturing industry

| Industry                        | (1) Freq. | Percent | (2) Freq. | Percent |
|---------------------------------|-----------|---------|-----------|---------|
| Consumer products and services  | 3439      | 19.4    | 2597      | 13.9    |
| Transportation equipment        | 1758      | 9.9     | 1905      | 10.2    |
| Chemical materials              | 1734      | 9.8     | 1915      | 10.3    |
| Machinery manufacturing         | 1391      | 7.8     | 1738      | 9.3     |
| Technology hardware             | 1224      | 6.9     | 1353      | 7.2     |
| Metal and mineral products      | 1153      | 6.5     | 1463      | 7.8     |
| Construction materials          | 1015      | 5.7     | 773       | 4.1     |
| Biopharmaceuticals              | 985       | 5.5     | 963       | 5.2     |
| Electronic components           | 962       | 5.4     | 1212      | 6.5     |
| Medical devices                 | 769       | 4.3     | 690       | 3.7     |
| Manufacturing equipment and services | 634 | 3.6 | 904 | 4.8 |
| Electrical equipment and power systems manufacturing | 593 | 3.3 | 767 | 4.1 |
| Industrial products and materials distribution | 590 | 3.3 | 467 | 2.5 |
| Aerospace and defense industry manufacturing | 382 | 2.2 | 486 | 2.6 |
| Containers and packaging        | 379       | 2.1     | 541       | 2.9     |
| Textile products                | 289       | 1.6     | 307       | 1.6     |
| Forestry and paper products materials | 209 | 1.2 | 185 | 1.0 |
| Plastic and rubber materials    | 145       | 0.8     | 240       | 1.3     |
| Scientific healthcare products  | 115       | 0.6     | 174       | 0.9     |
# Table 6: Number of target firms by industry in subsets (3) and (4) (i.e., non-manufacturing industry)

| Industry                                               | (3)    |         | (4)    |         |
|--------------------------------------------------------|--------|---------|--------|---------|
|                                                        | Freq.  | Percent | Freq.  | Percent |
| Consumer retail                                        | 2781   | 11.3    | 1618   | 7.3     |
| Banking                                                | 2773   | 11.3    | 1237   | 5.6     |
| Consumer hospitality                                   | 1800   | 7.3     | 1053   | 4.8     |
| Real estate                                            | 1616   | 6.6     | 2040   | 9.2     |
| Utilities services                                     | 1587   | 6.5     | 1243   | 5.6     |
| Information technology services                        | 1441   | 5.9     | 1471   | 6.7     |
| Construction and engineering services                  | 1236   | 5.0     | 2031   | 9.2     |
| Consumer media                                         | 1194   | 4.9     | 1085   | 4.9     |
| Insurance                                              | 1181   | 4.8     | 830    | 3.8     |
| Transportation and logistics services                  | 1138   | 4.6     | 1275   | 5.8     |
| Software                                               | 1113   | 4.5     | 1598   | 7.2     |
| Telecommunications services                            | 1051   | 4.3     | 950    | 4.3     |
| Investment                                             | 932    | 3.8     | 688    | 3.1     |
| Electronic transaction processing and clearing-house activities | 813    | 3.3     | 1111   | 5.0     |
| Upstream energy                                        | 742    | 3.0     | 984    | 4.5     |
| Mining materials                                       | 657    | 2.7     | 594    | 2.7     |
| Healthcare services                                    | 644    | 2.6     | 444    | 2.0     |
| Downstream energy                                      | 449    | 1.8     | 427    | 1.9     |
| Specialty finance                                      | 391    | 1.6     | 303    | 1.4     |
| Administrative and support services                    | 362    | 1.5     | 391    | 1.8     |
| Midstream energy                                       | 236    | 1.0     | 259    | 1.2     |
| Diversified businesses                                 | 150    | 0.6     | 166    | 0.8     |
| Outsourced biomedical services                         | 130    | 0.5     | 147    | 0.7     |
| Integrated oil and gas energy                          | 59     | 0.2     | 68     | 0.3     |
| Waste management services                              | 59     | 0.2     | 91     | 0.4     |
Acknowledgements This study was conducted as collaborative research of Tokio Marine dR Co., Ltd. and the National Institute of Informatics. This work was supported by JSPS KAKENHI Grant numbers 21H01569, 18KK0316.

Declarations

Conflict of interest The corresponding author states that there is no conflict of interest on behalf of all authors.

References

1. Gabaix, X. (2011). The granular origins of aggregate fluctuations. Econometrica, 79(3), 733–772. https://doi.org/10.3982/ECTA8769.
2. Acemoglu, D., Carvalho, V. M., Ozdaglar, A., & Tahbaz-Salehi, A. (2012). The network origins of aggregate fluctuations. Econometrica, 80(5), 1977–2016. https://doi.org/10.3982/ECTA9623.
3. Baqee, D. R. (2018). Cascading failures in production networks. Econometrica, 86(5), 1819–1838. https://doi.org/10.3982/ECTA15280.
4. Flaaen, A., & Pierce, J. (2019). Disentangling the Effects of the 2018–2019 Tariffs on a Globally Connected US Manufacturing Sector. Finance and Economics Discussion Series (FEDS). https://doi.org/10.17016/FEDS.2019.086.
5. Inoue, H., & Todo, Y. (2019). Firm-level propagation of shocks through supply-chain networks. Nature Sustainability, 2, 841–847. https://doi.org/10.1038/s41893-019-0351-x.
6. Huang, Y., Chen, L., Sibo, L., & Heiwai, T. (2020). Trade networks and firm value: Evidence from the US-China Trade War. SSRN: https://ssrn.com/abstract=3227972. https://doi.org/10.2139/ssrn.3227972

| Table 7 | Number of target firms by year $n$ |
|---|---|---|---|
| Subset no. | $n$ (year) | Freq. | Percent |
| (1) | 2012 | 3452 | 19.4 |
| | 2013 | 3724 | 21.0 |
| | 2014 | 3696 | 20.8 |
| | 2015 | 3509 | 19.8 |
| | 2016 | 3385 | 19.1 |
| (2) | 2012 | 5008 | 20.4 |
| | 2013 | 5264 | 21.5 |
| | 2014 | 5098 | 20.8 |
| | 2015 | 4666 | 19.0 |
| | 2016 | 4499 | 18.3 |
| (3) | 2012 | 3710 | 19.9 |
| | 2013 | 3899 | 20.9 |
| | 2014 | 3854 | 20.6 |
| | 2015 | 3664 | 19.6 |
| | 2016 | 3555 | 19.0 |
| (4) | 2012 | 4568 | 20.7 |
| | 2013 | 4743 | 21.5 |
| | 2014 | 4638 | 21.0 |
| | 2015 | 4155 | 18.8 |
| | 2016 | 4000 | 18.1 |
7. Amiti, M., Kong, S. H., & Weinstein, D. (2020). The Effect of the U.S.-China Trade War on U.S. Investment. *National Bureau of Economic Research, Working Paper No. 27114*. https://doi.org/10.3386/w27114

8. Gerschel, E., Martinez, A., & Mejean, I. (2020). Propagation of Shocks in Global Value Chains: The Coronavirus Case. *IPP Policy Brief*, (53).

9. Baqaee, D., & Farhi, E. (2020). Supply and Demand in Disaggregated Keynesian Economies with an Application to the Covid-19 Crisis. *Working Paper Series. National Bureau of Economic Research*. https://doi.org/10.3386/w27152

10. Hassan, T. A., Hollander, S., Lent, L. V., Schwedeler, M., & Tahoun, A. (2020). Firm-Level Exposure to Epidemic Diseases: COVID-19, SARS, and H1N1. *National Bureau of Economic Research, Working Paper No. 26971*. https://doi.org/10.3386/w26971

11. Meier, M., & Pinto, E. (2020). COVID-19 Supply Chain Disruptions. *Discussion Paper Series - CRC TR 224, Discussion Paper No. 239, Project C 02.*

12. Barrot, J.-N., & Sauvagnat, J. (2016). Input specificity and the propagation of idiosyncratic shocks in production networks. *The Quarterly Journal of Economics, 131*(3), 1543–1592. https://doi.org/10.1093/qje/qjw018.

13. Carvalho, V. M., Nirei, M., Saito, Y. U., & Tahbaz-Salehi, A. (2021). Supply chain disruptions: Evidence from the Great East Japan Earthquake. *The Quarterly Journal of Economics, 136*(2), 1255–1321. https://doi.org/10.1093/qje/qjaa044.

14. Kashiwagi, Y., Todo, Y., & Matous, P. (2018). Propagation of economic shocks through global supply chains: Evidence from hurricane sandy. *Review of International Economics, 29*(5), 1186–1220. https://doi.org/10.1111/roie.12541.

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

16. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *NIPS’18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 6639–6649).

17. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

18. National Oceanic and Atmospheric Administration. (2021). Billion-dollar weather and climate disasters: Events. https://www.ncdc.noaa.gov/billions/events. Accessed 25 July 2021.

19. Blake, E. (2013). Hurricane Sandy. https://www.nhc.noaa.gov/outreach/presentations/Sandy2012.pdf. Accessed 15 Nov 2020.

20. Phillips, R. L., & Ormsby, R. (2016). Industry classification schemes: An analysis and review. *Journal of Business & Finance Librarianship, 21*(1), 1–25. https://doi.org/10.1080/08963568.2015.1110229.

21. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *The Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.