Customs Import Declaration Datasets

Chaeyoon Jeong
KAIST
Daejeon, Republic of Korea
lily9991@kaist.ac.kr

Sundong Kim
Institute for Basic Science
Daejeon, Republic of Korea
sundong@ibs.re.kr

Jaewoo Park
Korea Customs Service
Daejeon, Republic of Korea
jaeus@korea.kr

Yeonsoo Choi
Korea Customs Service
Daejeon, Republic of Korea
yschoi0817@korea.kr

ABSTRACT

Given the huge volume of cross-border flows, effective control of trades becomes more crucial in customs administrations. However, limited accessibility of the customs datasets hinders the progress of open research, and lots of member countries have not benefited from the recent progress. In this paper, we introduce an import declarations dataset to facilitate the collaboration between the domain experts in customs administrations and data science researchers. The dataset contains 54,000 artificially generated trades with 22 key attributes, and it is synthesized with CTGAN while maintaining correlated features. Synthetic data has several advantages. First, releasing the dataset is free from restrictions that do not allow disclosing the original import data. Second, the fabrication step minimizes the possible identity risk which may exist in trade statistics. Lastly, the published data follow a similar distribution to the source data so that it can be used in various downstream tasks. With the provision of data and its generation process, we open baseline codes for fraud detection tasks, as we empirically show that more advanced algorithms can better detect frauds.

CCS CONCEPTS

• **Social and professional topics** → **Taxation**; • **Applied computing** → **E-government**.

KEYWORDS

Import Declarations, Synthetic Data, Tabular Data, Customs Fraud Detection, Correlation Analysis

1 INTRODUCTION

Customs clearance is the process of getting permission from the customs administrations to either move goods out of a country (export) or bring goods into the country (import). The customs declarant declares the good to the customs office, and permission is given only when the declaration is legitimate. If the shipment exceeds the threshold value ($150 in Korea), the customs impose tariffs on the item. Once the tariff is collected, they allow the goods to be taken out.

Given the enthusiasm around the use of data and the possibilities offered by artificial intelligence [9], the adoption of new technology is relatively slow in the customs community. The primary reason is the lack of publicly available data. Disclosure of import declaration data outside customs is strictly prohibited because of its sensitivity. Only authorized departments or institutions could conduct research internally, and there is no visible community effort.

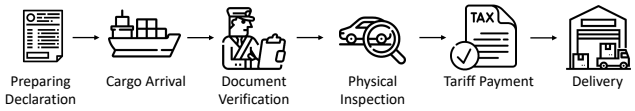


Figure 1: Import clearance process

This leads us to design synthetic data that can be open to the public. The dataset contained in this paper includes 54,000 artificially generated trades with 22 attributes. Using a tabular synthesizer with post-processing techniques, we maintain that the distribution and correlation in the synthetic dataset are similar to the source dataset. Empirical results on fraud detection demonstrate the potency of the data. Meanwhile, the data is used for competition in three universities to develop algorithms that can be applied in practice. We conclude the paper by discussing possible scenarios to use the data and summarizing necessary thoughts on the data synthesis. The data and code can be found in https://github.com/Seondong/Customs-Declaration-Datasets.

2 DATA DESCRIPTION

2.1 Data Schema

The tabular form dataset consists of 54,000 import declarations, where each row describes the report of a single item. Among 62 attributes in the import declaration form,[1] the data includes 22 representative attributes without overlapped or less essential one. The first 20 values are prepared in the declaration stage of the customs clearance, while the rest two attributes are labeled after the inspection. *Fraud* indicates whether the inspected result of the actual imported goods conflict with its declaration. *Critical fraud* is a case that may threaten society's public safety or stability, such as copyright infringement, drug smuggling, or false declaration of the origin of goods. Categorical attributes and their values follow the handbook provided by KCS, which contains trade codes used for filling out import and export declarations in Korea.[2] Detailed data descriptions and example values are shown in Table 1.

2.2 Data Reliability

We conduct statistical tests between the source data and the synthetic dataset to show whether the two data come from the same distribution. The Chi-Squared test and the Two-Sample Kolmogorov-Smirnov test are used. Both tests compare individual attributes from the source data with the corresponding one from the synthetic data and report the average similarity score ranging from 0 to 1. The Chi-Squared test compares the distributions of two categorical attributes. The output for each column indicates the probability that

Conference'17, July 2017, Washington, DC, USA
2022. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] Import declaration format and explanation: https://bit.ly/import-declaration-form
[2] https://www.data.go.kr/data/3040477/fileData.do

Table 1: Data description

Attribute	Value	Explanation
Declaration ID	97061800	Primary key of the record
Date	2020-01-01	Date when the declaration is reported
Office ID	13	Customs office that receives the declaration (e.g., Seoul regional customs)
Process Type	B	Type of the declaration process (e.g., Paperless declaration)
Import Type	11	Code for import type (e.g., OEM import, E-commerce)
Import Use	21	Code for import use (e.g., Raw materials for domestic consumption, from a bonded factory)
Payment Type	11	Distinguish tariff payment type (e.g., Usance credit payable at sight)
Mode of Transport	10	Nine modes of transport (e.g., maritime, rail, air)
Declarant ID	L77JJEG	Person who declares the item
Importer ID	HQ0W7JA	Consumer who imports the item
Seller ID	PBP2MYI	Overseas business partner which supplies goods to Korea
Courier ID	MWIDNS	Delivery service provider (e.g., DHL, FedEx)
HS10 Code	0901210010	10-digit product code (e.g., 090121xxxx = Coffee, Roasted, Not Decaffeinated)
Country of Departure	JP	Country from which a shipment has or is scheduled to depart
Country of Origin	JP	Country of manufacture, production or design, or where an article or product comes from
Country of Origin Indicator	B	Way of indicating the country of origin (e.g., B = Mark on package)
Tax Rate	8.0	Tax rate of the item (%)
Tax Type	A	Tax types (e.g., FTA Preferential rate)
Net Mass	1262.0	Mass without any packaging (kg)
Item Price	1437418.0	Assessed value of an item (KRW)
Fraud	1	Any fraudulent attempt to reduce the customs duty? (0/1)
Critical Fraud	1	Critical case which may threaten the public safety (0/1)

Table 2: Statistical test results indicate that the synthetic data and the source data come from similar distribution.

Metric	Score
Chi-Squared Test	1.000
Two-Sample Kolmogovrov-Smirnov test	0.772

the two attributes are sampled from the same distribution. The Two-Sample Kolmogorov-Smirnov test is used to compare cumulative distribution functions of two continuous variables. The output for each column indicates one minus the maximum distance between two CDFs. We used synthetic data evaluation functions built in the Synthetic Data Vault library [10]. As shown in Table 2, our synthetic dataset is quite indistinguishable from the source dataset.

In addition, we compare the distribution of each attribute in two datasets. Figure 2 show histograms of representative features (*Tax Rate, Net Mass, Critical Fraud*) in the dataset are analogous to that of the source dataset. Figure 3 illustrates that the correlation between the attributes in the synthetic data is similar to that of the source data. Figure 4 shows representative features in the downstream fraud detection task performed on each dataset by using the XGBoost model [1]. The tendency of feature importance score is analogous in both datasets. *Impoter ID, Item Price, Net Mass, Declarant ID*, and *HS10 Code* were considered important with high scores. This indicates that the relationships which playes important role in the actual customs task are well represented.

3 DATA GENERATION

3.1 Preprocessing

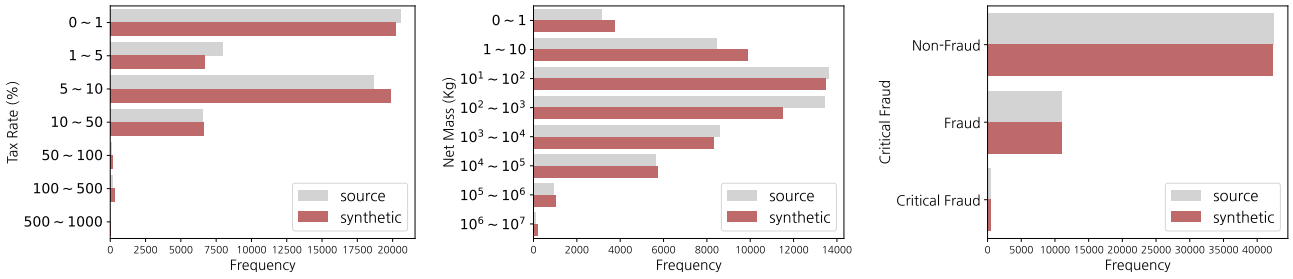
Among 24.7 million customs declarations reported for 18 months between January 1, 2020, and June 30, 2021, we used the inspected (i.e., labeled) part of the declarations to synthesize our data. Inspected items account for a relatively small percentage of the total, but they are more accurate, all validated by field officers. We designate it as the source data throughout the paper. Identifiable information such as the importer name in the source data is anonymized into *Importer ID*. The price of goods traded between vendors (i.e., *Item Price*) can be problematic when fully disclosed, so we add some Gaussian noise to the average price of each category of item.

3.2 Generating Data with CTGAN

We used conditional tabular GAN (CTGAN) [13] from the Synthetic Data Vault library to generate the data. CTGAN specializes in tabular data and uses conditional techniques to handle imbalanced discrete and multi-modal continuous variables. Compared to other tabular generative models such as TGAN [14] or TVAE [13], CTGAN showed the most realistic output to our dataset, preserving the relationship between columns. The data generation process can be done in a serial or parallel manner. Users with limited resources can split the data in chronological order, train the CTGAN model on each split, synthesize samples from each model, and aggregate the result. For each split, the model is trained for 300 epochs.

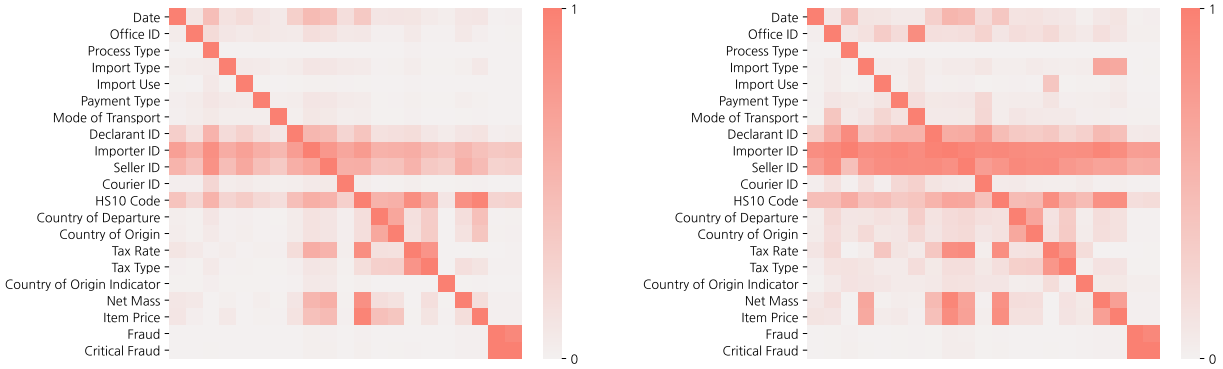
3.3 Maintaining Correlated Attributes

Tabular data have correlated attributes. For example, attributes *HS10 Code—Country of Departure—Country of Origin—Tax Rate—Tax Type* are highly correlated based on customs valuation policies.



(a) Tax Rate **(b) Net Mass** **(c) Critical Fraud**

Figure 2: Representative feature distributions are similar between synthetic data and source data.



(a) Feature correlation in synthetic data **(b) Feature correlation in source data**

Figure 3: Feature correlation in synthetic data and in source data are also similar to each other.

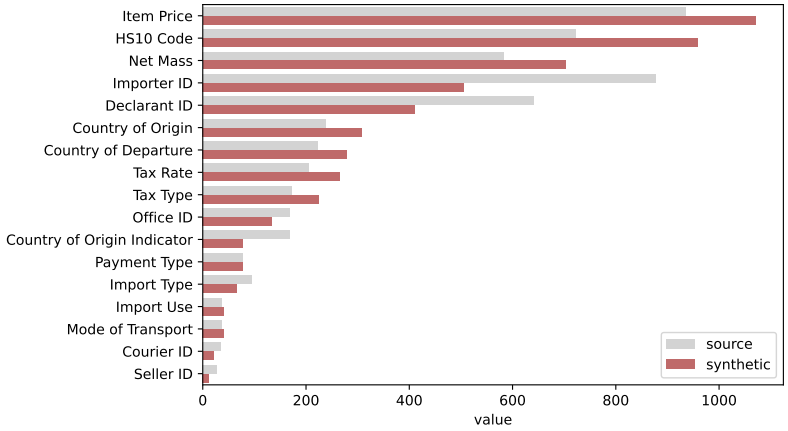


Figure 4: Important features for fraud detection task are also similar between the two datasets.

To make the import declaration data more realistic, a synthesizer should maintain the correlated attributes and their values during the generation process. Although CTGAN is a tabular-specific generative model, the output does not always reflect clear correlations between attributes. To maintain dependencies, we aggregate correlated attributes into a single column and save it temporarily before running the CTGAN model. After running the model, the value is split to have the original format. Another example is *Item Price*, which is an attribute correlated with the *HS10 Code* and the *Net*

Mass of an item. To maintain this relationship, *Item Price* is reconstructed after the data generation step by multiplying *Net Mass* and the unit price of each *HS10 Code*.

4 APPLICATION— FRAUD DETECTION

This section introduces how our dataset can be used as a benchmark for the customs fraud detection problem.

4.1 Background

Smuggling and tax evasion are fatal threats to society and customs administrations prevent those risks by customs control. Due to high trade volume and limited budget, it is difficult to conduct an exhaustive inspection of all items, so customs define a set of rules to screen out high-risk items based on the contents of import declarations. Therefore, establishing an intelligent customs selection or fraud detection system is key to facilitating the customs clearance process [4–6, 8]. By predicting which item is likely-fraud, they can determine the inspection level of each item. The most suspicious items are subjected to physical inspection requiring human labor, so the smarter the algorithm, the more efficiently customs can operate its workforce.

4.2 Using the Data

The fraud detection problem aims to find the patterns behind the features in predicting the target label *Fraud*. Data is split into three pieces. We assign the first 12-month of data to the training set, the following three months into the validation set, and the last three months into the test set. Categorical variables are label-encoded and numerical variables are min-max scaled. We apply various models including Logistic Regression, Decision Tree, Random Forest, AdaBoost in scikit-learn [11] and gradient boosting decision tree (GBDT) models such as LightGBM [3], CatBoost [12], XGBoost [1]. We set the model to predict each record's fraud score ranging from 0 to 1. Among test records, n% of items with the highest fraud score are inspected. Model performance is evaluated by the precision@n%, representing how many inspected items are actually fraud.[3]

4.3 Performance Comparison

Table 3 shows the performance trend of applying various fraud detection algorithms on synthetic and source data. The results are averaged over five runs. Given that customs administration inspect a limited quantity of goods, we considered two inspection rates – 5% and 10%. In both datasets, precision on 5% setting is higher than that of 10%, and the performance of GBDT models such as CatBoost, XGBoost, and LightGBM is higher than the other models. Interestingly, the performance gained by applying an advanced model is distinguishable in synthetic data with a low inspection rate setting. We conclude from these results that the synthetic data can be used as an open benchmark to develop advanced fraud detection algorithms.

[3]The amount of workforce available for physical inspection is usually fixed, so it is important to achieve the best performance within the limited inspection capacity, without changing n. Therefore, precision@n% is a more suitable metric than AUC or f-score.

Table 3: The fraud detection performance in the synthesized data follow a similar trend to the real data.

	Synthetic data		Source data	
Model | Precision	n = 5%	n = 10%	n = 5%	n = 10%
Logistic Regression	0.258±0.016	0.248±0.013	0.348±0.016	0.343±0.014
AdaBoost	0.341±0.012	0.313±0.007	0.463±0.019	0.428±0.020
Decision Tree	0.346±0.014	0.314±0.011	0.424±0.022	0.407±0.014
Random Forest	0.364±0.012	0.339±0.007	0.472±0.015	0.421±0.013
CatBoost	0.621±0.011	0.520±0.006	0.568±0.010	0.499±0.010
XGBoost	0.675±0.011	0.579±0.008	0.539±0.024	0.465±0.019
LightGBM	0.762±0.003	0.641±0.003	0.604±0.013	0.519±0.008

5 DISCUSSION

Area of research: Besides fraud item detection, this import declaration data can be used for solving various data science problems in the customs domain such as HS code classification [7], trade pattern analysis between the key players such as importers, declarants, and offices. We also expect to see user-friendly support in customs e-clearance systems.

Distribution of data: For users' convenience, we create the data of inspected items, thus fully-labeled. However, not all items are inspected in practice. According to each country's inspection standards, many goods are cleared without going through any physical inspection, especially in developed countries with low fraud rates. We can simulate this partially-labeled scenario through post-processing by erasing some data labels.

Degree of fabrication: Anonymizing the data is insufficient to mitigate the potential risk of releasing the data. Adversaries may catch the patterns between the key players and disrupt the trade order even if the declarant code, product classification code, and extraction country code are anonymized. Therefore, we release synthesized data instead.

Generative model: Recently, the diffusion model has been discussed as an alternative way of generating artificial data in the computer vision domain [2]. A diffusion model specialized for tabular data can also be used to create the synthetic data.

6 CONCLUSION

We present the customs import declaration data produced as part of sharing challenging data science problem in customs administration and facilitate the collaboration between customs and data science communities. With careful fabrication strategy, the generated data is fairly similar to the actual data, and can be used as a benchmark for downstream tasks such as fraud detection.

REFERENCES

[1] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *KDD*. 785–794.
[2] Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. *arXiv preprint arXiv:arXiv:2105.05233* (2021).
[3] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *NeurIPS*.
[4] Sundong Kim, Tung duong Mai, Sungwon Han, Sungwon Park, Thi Nguyen, Jaechan So, Karandeep Singh, and Meeyoung Cha. 2022. Active Learning for Human-in-the-loop Customs Inspection. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[5] Seongchan Kim, Sa-Kwang Song, Minhee Cho, and Su-Hyun Shin. 2021. Transaction Pattern Discrimination of Malicious Supply Chain using Tariff-Structured Big Data. *The Journal of the Korea Contents Association* (2021).
[6] Sundong Kim, Yu-Che Tsai, Karandeep Singh, Yeonsoo Choi, Etim Ibok, Cheng-Te Li, and Meeyoung Cha. 2020. DATE: Dual Attentive Tree-Aware Embedding for Customs Fraud Detection. In *KDD*.
[7] Eunji Lee, Sundong Kim, Sihyun Kim, Sungwon Park, Meeyoung Cha, Soyeon Jung, Suyoung Yang, Yeonsoo Choi, Sungdae Ji, Minsoo Song, and Heeja Kim. 2021. Classification of goods using text descriptions with sentences retrieval. In *Korea Artificial Intelligence Conference (KAIA)*.
[8] Tung-Duong Mai, Kien Hoang, Aitolkyn Baigutanova, Gaukhartas Alina, and Sundong Kim. 2021. Customs fraud detection in the presence of concept drift. In *ICDM IncrLearn Workshop*.
[9] Kunio Mikuriya and Thomas Cantens. 2020. If Algorithms Dream of Customs, do Customs Officials Dream of Algorithms? A Manifesto for Data Mobilisation in Customs. *World Customs Journal* 14, 2 (2020).
[10] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. 2016. The Synthetic Data Vault. In *DSAA*. https://sdv.dev/
[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[12] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features support. In *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 6639–6649.
[13] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling Tabular data using Conditional GAN. In *NeurIPS*.
[14] Lei Xu and Kalyan Veeramachaneni. 2018. Synthesizing Tabular Data using Generative Adversarial Networks. *arXiv preprint arXiv:1811.11264* (2018).