Detecting Suspicious Timber Trades

Debanjan Datta,[1] M. Raihanul Islam,[1] Nathan Self,[1] Amelia Meadows,[2] John Simeone,[2,4] Willow Outhwaite,[3] Chen Hin Keong,[3] Amy Smith,[2] Linda Walker,[2] Naren Ramakrishnan[1]

[1]Virginia Tech, Arlington, Virginia, USA, [2]World Wildlife Fund, Washington DC, USA
[3]TRAFFIC, Cambridge, UK, [4]Simeone Consulting LLC, Anchorage, Alaska, USA

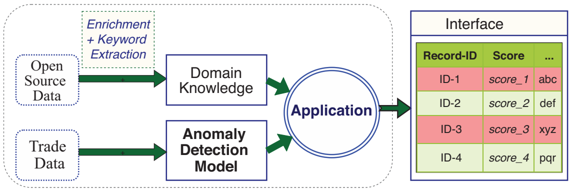
Abstract

Developing algorithms that identify potentially illegal trade shipments is a non-trivial task, exacerbated by the size of shipment data as well as the unavailability of positive training data. In collaboration with conservation organizations, we develop a framework that incorporates machine learning and domain knowledge to tackle this challenge. Modeling the task as anomaly detection, we propose a simple and effective embedding-based anomaly detection approach for categorical data that provides better performance and scalability than the current state-of-art, along with a negative sampling approach that can efficiently train the proposed model. Additionally, we show how our model aids the interpretability of results which is crucial for the task. Domain knowledge, though sparse and scattered across multiple open data sources, is ingested with input of domain experts to create rules that highlight actionable results. The application framework demonstrates the applicability of our proposed approach on real world trade data. An interface combined with the framework presents a complete system that can ingest, detect and aid in the analysis of suspicious timber trades.

1 Introduction

International trade of wood and forest products is a major component of the global trade of renewable natural resources. While there is substantial legal trade involving timber products, illegal logging and trade of illegally sourced wood products threatens ecosystems and livelihoods worldwide, deprives governments of revenues, and is often associated with other activities like corruption and illicit financial flows (Lawson and MacFaul 2010). Timber that are at risk of being illegally harvested or traded from entering the US market are referred to as high risk timber.

Illegal logging is the third largest category of transnational crime with an annual retail value estimated to be between $52 and $157 billion as of May 2017. The United States is the world's largest importer of timber products by value, totalling $51.5 billion in 2017 (22% of all global imports). Wiedenhoeft and others 2019 determined a significant portion of retailers are involved in mislabelling or misrepresentation of timber species. Trade records are considered suspicious when they have a higher likelihood of item mislabelling, fraud or pertain to high risk timber.



Record-ID	Score	...
ID-1	*score_1*	abc
ID-2	*score_2*	def
ID-3	*score_3*	xyz
ID-4	*score_4*	pqr

Figure 1: System overview of the application framework.

For our target users which are comprised of the government agencies, current approaches for detecting suspicious transactions involve experts or analysts inspecting individual trade records (bills of lading). They do not possess tools that are specific to the timber trade or that use machine learning. The size and complexity of trade data suggests an automated approach that incorporates timber-specific domain knowledge and highlights potentially suspicious records which can lead to further investigation or action by authorities. Domain specific data sources about high risk timber species which can be used to find suspicious transactions are crucial but scattered or incomplete and require enrichment and extraction with inputs from domain experts. While post-hoc analysis of a record may reveal whether a record was suspicious, ground truth data are unavailable as of yet to train a model.

Anomaly detection algorithms have been applied for detecting fraudulent activities, with the intuition that they can detect relatively infrequent records that do not conform to usual patterns. While some previous approaches rely on statistical anomalies in terms of weight, volume or declared dollar values, such numerical attributes are not available in all cases, including ours. Unexpected co-occurrences are often present across attributes of transactions and it is assumed that identification of such anomalous patterns can detect suspicious transactions (Das and Schneider 2007). Examples of such anomalies that might signify suspicious transactions include a shipper exporting goods they do not usually trade in, or a consignee importing goods through a port without prior precedence. Thus we formulate the given task as a problem of unsupervised anomaly detection in categorical data, combined with ingestion, enrichment and application of domain knowledge that detects suspicious timber trades (see Figure 1). Such frameworks do not exist currently and there is a critical need for a better decision support system for assessing trade data for potential anomalies.

Our contributions are as follows: (1) An end-to-end application framework that combines our anomaly detection model with targeted domain knowledge with a user interface for detecting potentially suspicious records in trade data. (2) A new embedding based unsupervised anomaly detection approach for categorical datasets. (3) A simple and efficient negative sampling approach for multivariate categorical data is proposed that is used in training our model. (4) Interpretability of model output utilizable by end-users.

2 Related Work

Anomaly detection has been studied in the context of applications such as cybersecurity, fraud, social networks and law enforcement (Aggarwal 2017). Goldstein and Uchida (2016) provide a detailed survey on unsupervised anomaly detection techniques in continuous valued data. Fraud detection based on anomaly detection in trade and financial records have been proposed in works such as Das and Schneider (2007) and Abdallah, Maarof, and Zainal (2016). Transaction fraud detection approaches have also used methods such as actor specific transaction sequences or profiles (Lp and others 2018). These approaches are not readily applicable to our case given the characteristics of trade data.

Unlike graph, sequence or spatial data, categorical data does not have implicit distance measures (Zhang et al. 2015). Some of the approaches for unsupervised anomaly detection in categorical multivariate data are distance or density based, frequency and itemset based, and information theoretic (Taha and Hadi 2019). Many of these approaches have limitations in terms of accuracy and scalability. Proposed in Das and Schneider (2007), *Condition* calculates the conditional probability of a record based on attribute sets of size k and combines them using heuristics that cover the entire set of attributes. Some notable information theoretic approaches are *Krimp* (Vreeken and others 2011; Smets and Vreeken 2011), *Comprex* (Akoglu et al. 2012) etc. While these methods outperform previous methods, they are not scalable. *Comprex*, one of our baselines, has time complexity in order of $\mathcal{O}(t^2)$, where t is total number of entities. Tang et al. (2015) presents an itemset based approach for mining contextual outliers on categorical data, however it also does not scale well.

APE (Chen et al. 2016) is a state-of-the-art embedding based approach. Entities are represented in a common embedding space and pairwise interactions between them are computed to obtain the likelihood of a record. Dependence on pairwise co-occurrence relationships between entities might not capture higher order relationships, and results in training time that is polynomial in terms of the number of entities. Chen et al. (2017) uses autoencoder ensemble for anomaly detection on continuous-valued data. Wang et al. (2018) presents an embedding norm based model for behavioral prediction in citation networks. Works such as Hu et al. (2016) proposed network anomaly detection using representation learning.

3 Proposed Anomaly Detection Method

Embeddings can capture co-occurrence relationships between entities efficiently without explicit enumeration, such as in itemset or conditional frequency based approaches. We propose the ***M**ulti-relation **E**mbedding based **A**nomaly **D**etection (MEAD)* model to detect anomalies in categorical multi-relational data.

Preliminaries

A *domain* is defined as a set of elements sharing a common property, denoted as U_j for $j = 1, 2...l$, where l is the number of domains. A domain U_j consists of a set of *entities*. The i^{th} *entity* belonging to the j^{th} domain is denoted by e_j^i. There is no implicit ordering among entities in a domain. The count of entities in a domain is termed as *arity* of the domain. A tuple of entities, with one entity belonging to each of the l domains forms a *multi-relation* or *record*. It is denoted as r, and the set of all records as R. More formally, $r = \bigcup_{j \in 1...l} \{e_j^r\}$. Here e_j^r is the entity in record r belonging to j^{th} domain. *Context* is defined as the reference group of entities with which an entity occurs, implying an entity can be present in multiple contexts. The context for an entity e_j^r with respect to a record r is the set $\bigcup_{i \neq j} e_i^r$.

Problem statement: Given a dataset R consisting of multivariate categorical data (records) which are assumed to be normal, learn a model that can predict the likelihood of a test record being normal. The test records with likelihood below a threshold are judged anomalous.

Model Architecture

The model architecture consists of a single embedding layer so that all entities belonging to all domains are represented in the same latent space. We define *embedding* as a transformation f_j for the j^{th} domain. Let the embedding of the entity in record r belonging to the j^{th} domain be $f_j(e_j^r)$, denoted is as x_j^r. Each domain has a weight which is denoted as W_j. The Hadamard product between W_j and x_j^r obtains the weighted entity embedding, as shown in Eq. (1). We apply a non-linearity on the square of the Euclidean (L_2) norm of the resulting weighted vector to obtain the likelihood of occurrence, or score, of a record r, as defined in Eq. (2). The use of the hyperbolic tangent function, *tanh*, ensures the scores are between 0 and 1, since the input is always positive. A test record with a high score is deemed normal, whereas records with low scores are considered anomalous. The non-linear activation function (*tanh*) has a faster convergence rate than a standard logistic function (LeCun et al. 2012). We find that the sigmoid function performs only marginally worse.

$$z_j^r = W_j \odot x_j^r; \quad z_r = \left(\|\Sigma_{j=1}^{l} z_j^r\|_2 \right)^2 \tag{1}$$

$$P_\theta(r) = \tanh(z_r) \tag{2}$$

The entity embeddings are trained based on the *context* in a way that can capture co-occurrence relationships between entities. The aim of the model is to calculate the approximate likelihood, or score, of a transaction based on the embeddings of the constituent entities. The intuition for our approach is that the weighted sum of entities corresponding to

Table 1: Dataset names and descriptions of the domain names for each and their respective arities in parenthesis. The training and test set sizes for each test case are reported in last two columns.

Data Set	Domains (Arity)	Train Size	Test Size
Peru Export (PE)	Customs Code (10), Location Code (117), Port of Unlading (349), HTS Code (56, Shipment Destination (78), Shipper ID (577), Transport Method (5)	32108	*TC-1*: 5159 *TC-2*: 5152
China Import (CI)	Admin Region (386), Consignee ID (5270), Country of Sale (122), Province (31), Shipment Origin (127), Trade Type (10), Transport Method (6), HTS Code (74)	81896	*TC-1*: 25572 *TC-2*: 25574
China Export (CE)	Admin Region (456), Country of Sale (193), Province (31), Shipment Destination (193), HTS Code (85), Shipper ID (9227), Trade Type (9), Transport Method (6)	216614	*TC-1*: 77579 *TC-2*: 77579
US Import 1 (US-1)	Carrier (548), Consignee ID (5113), Port of Lading (238), Port of Unlading (64), Shipment Destination (113), Shipment Origin (116), Shipper ID (6193), HTS Code (95)	140382	*TC-1*: 51190 *TC-2*: 51190
US Import 2 (US-2)	Carrier (701), Consignee ID (8960), Port of Lading (286), Port of Unlading (75), HTS Code (97), Shipment Destination (136), Shipment Origin (126), Shipper ID (10661)	238927	*TC-1*: 81238 *TC-2*: 81238

valid transactions should be additive in nature. This is due to the co-occurrence of such entities and their shared contexts. Let θ be the parameters of the model and $P_\theta(r)$ be the probability distribution that estimates the likelihood of a record being normal. Let P_e be the empirical distribution of data. Our goal is to determine a set of embeddings that models the likelihood of a record being normal. To find a distribution P_θ that approximates the empirical distribution P_e such that the distance between them is minimized, we restate our objective as min_θ KL-Divergence$(P_\theta \| P_e)$. Simplifying and removing some constants we get the form shown in Eq. (3).

$$d = -\Sigma_{r \in R} P_e(r). \log(P_\theta(r)) \tag{3}$$

Given the exponentially large possible combinations of entities that can form records, a direct estimation of likelihood of records is computationally intractable. An indirect approximation approach is required and for this purpose negative sampling is chosen. The complete loss function is defined in Eq. (4). While training the model, $P_e(r)$ for normal instances is assumed to have a value of 1.

$$\mathcal{L} = -\Big(\sum_{r \in R} \big(logP_\theta(r) + \sum_{k \in r'} log(\tanh(z_k^{-1})) \big) \Big) + \mathcal{L}_{\mathcal{Z}} \tag{4}$$

$$\mathcal{L}_{\mathcal{Z}} = \sum_{r \in R} \frac{1}{|l|} \sum_{j=1}^{l} (1 - \|z_j^r\|_2)^2 \tag{5}$$

The negative sign on the first collective term is to change the maximization objective to a minimization. The reciprocal of the norm of negative samples is taken before applying *tanh* since the model should be trained such that their scores should be very low. The second part of the the first component in Eq. (4) is the sum of log likelihood of negative samples. Here r' is the set of negative samples corresponding to multi-relation r. The term in Eq. (5) is added to force the representations of the entities to have an L_2 norm close to 1. This helps in achieving interpretability of the model output, as explained in Section 5.

Context Preserving Negative Sampling

Negative sampling is an adaptation of *Noise Contrastive Estimation* (Gutmann and Hyvärinen 2010). Negative samples for records are defined contextually since a combination of entities may be present in certain contexts but absent in others. Generation of negative samples for multi-relations has two major challenges.

First, exhaustive computation of all possible negative samples is intractable given the large size of the datasets. However the negative samples must have adequate variation as well. Second, negative samples should resemble the original record to an extent and not be equivalent to random noise. This is to ensure the model learns co-occurrence of entities in various contexts. Chen et al. (2016) proposes an approach for generating negative samples for record data where one domain's entity is perturbed per negative sample. This does not scale well as number of domains increases.

Thus we propose the the approach shown in Algorithm 1, utilizing random subspace sampling (Barandiaran 1998). By randomly perturbing a part of the record, the relative context of the unperturbed set of entities, as well the entire record, changes. At the same time this also introduces variation among negative samples, although by preserving at least half the entities resemblance to the original record is maintained. This leads to negative samples that accurately train the model.

Algorithm 1: Context preserving negative sampling

```
Data: Training records
Result: Negative samples for each training record
for each training record $r \in R$ do
    for i = 1:k negative samples do
        Select $j \in [1, \lfloor l/2 \rfloor)$ domains randomly;
        for each domain in $j$ do
            Replace original entity with a randomly
              chosen entity (same domain);
```

4 Model Evaluation and Analysis

Data

Real world trade data sets containing partially standardized, individual bills of lading are obtained from Panjiva (2019). We use 5 sets of import and export data from 3 different countries (US, China and Peru) for our evaluation. Records unrelated to timber products are discarded, using item description codes as explained in Section 6. The data is first

Table 2: Comparison of performance in terms of average precision between the baselines and our method.

Case	Method	PE	CI	CE	US-1	US-2
TC-1	Condition	0.565	0.696	0.686	0.588	0.598
	APE	**0.913**	0.915	0.947	0.833	0.894
	MEAD	0.896	**0.949**	**0.949**	**0.971**	**0.970**
TC-2	Condition	0.583	0.739	0.716	0.609	0.615
	APE	0.943	0.922	0.982	0.963	0.956
	MEAD	**0.947**	**0.982**	**0.983**	**0.992**	**0.991**

Table 3: Comparison between average precision of *Comprex* and *MEAD*. Due to the long run-time of *Comprex*, 2000 test samples are used. Average run time on test samples is shown with each method demonstrating the scalability of *MEAD*.

Method		PE	CI	CE	US-1	US-2
Comprex	*TC-1*	0.98	0.98	0.99	0.94	—
	TC-2	0.99	0.99	0.99	0.98	—
	Time(s)	**249**	**425**	**515**	**1142**	**DNF**
MEAD	*TC-1*	0.90	0.95	0.94	0.96	0.97
	TC-2	0.94	0.98	0.98	0.99	0.99
	Time(s)	**0.02**	**0.02**	**0.02**	**0.02**	**0.03**

preprocessed by selecting relevant attributes and removing rows with missing values. Attributes which are superfluous and with entropy of 0 or 1 are removed. Subsets of the entire dataset are used for evaluation, with different sizes to understand model performance with varying number of entities and records. We train the model on part of the data and use the chronologically subsequent segment as test set. Instances present in training data are removed from the test set. To test the performance of our algorithm, we evaluate how well it captures unexpected or unusual co-occurrences of entities which are considered anomalies. Due to absence of labelled data we generate two cases of anomalies to be detected. In *TC-1* a pair of entities from each test set record are randomly perturbed. For *TC-2* we perturb a triplet of entities. Thus there are two test sets with both normal and anomalous records for each dataset, similar to reference (Chen et al. 2016). In generating negative samples for all relevant methods, we ensure negative samples do not overlap with training data. The details of the datasets used are shown in Table 1.

Baselines and Results

The area under the precision-recall curve (average precision) is reported over multiple test runs for each method in Tables 2 and 3. The following baselines are used.

Condition (Das and Schneider 2007) uses conditional probability tests along with heuristics to determine anomalous instances. We set hyper-parameters $\alpha = 0.1$ and $\beta = 0.1$. k is set to 2 for dataset CE and 1 for others.

APE (Chen et al. 2016) uses a common embedding to represent the entities and computes the pairwise interaction between them to calculate the likelihood of a multi-relation (event) occurring.

Comprex (Akoglu et al. 2012) is a compression (Minimum Description Length Principle) based approach as discussed in Section 2. It does not require any parameters. We used the authors original implementation for our experiments.

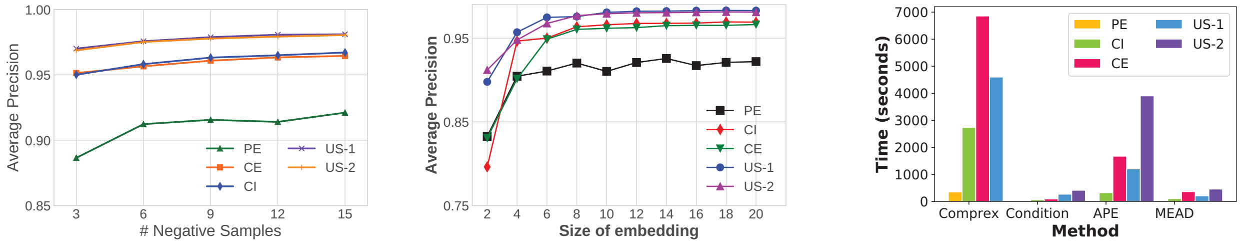
MEAD performs equivalent or better than both *Condition* and *APE* in most cases. While *APE* performs comparably for smaller datasets, our method handles larger datasets better. Comparison with *Comprex* is shown in Table 3, with both the average precision and testing time reported. Due to the long run time of *Comprex* a smaller test set of 2000 records is sampled from the original datasets. While *Comprex* provides better results in some cases, no significant advantage is observed for larger datasets. However the test time of *Comprex* is very high, and we are unable to train the model for the largest dataset (*US-2*) after running for 2 days, making it unsuitable for such use cases, as also noted in (Chen et al. 2016; Taha and Hadi 2019). For numerical stability, a very small number (10^{-8}) is added in logarithmic computations and gradient clipping is used. The *Adam* optimizer is used for training embedding models.

Model Parameters and Analysis

Negative Samples are used in training the model. The intuitive understanding is that more negative samples would lead to a better approximation, thus a more accurate model. The model performance for varying numbers of negative samples per record used in training the model is reported in Figure 2a. A comparatively smaller number of negative samples does not deteriorate performance drastically. Thus *MEAD* can be trained with a small number of negative samples incurring a lower computational cost, especially for a large datasets. **Embedding size** is an crucial hyperparameter, since a larger embedding can encode more information. Average precision as embedding size increases for all datasets, averaged over both test cases, is reported in Figure 2b. Increasing embedding size increases model performance, though it plateaus after a certain point. Also, datasets with a large number of entities benefit from a larger embedding size which coincides with our intuition. **Scalability** is one of the key requirements for a model to be effective in a practical application. The execution times for *MEAD* and baselines are shown in Figure 2c. While *Condition* has a low execution time using $k = 1$, the average precision is inferior. *APE* has greater execution time due to $O(l^2)$ pairwise computations and $k \times l$ negative samples. *Comprex* has the highest computational cost making it unsuitable.

5 Interpretability

Though not adequately addressed by existing methods, interpretability is crucial in terms of usability for end users. As shown in Eq. (5), the model is optimized so that the L_2 norm or magnitude of weighted embeddings vectors for the entities is approximately 1. Thus while vectors have approximately equal magnitude, their direction encodes information about interactions among entities represented by them. It may be imagined that vectors lie on the surface of a hypersphere. The intuition here is that embedding vectors for usually co-occurring entities are similarly oriented. Thus their sum, and hence L_2 norm of their sum will be high. Conse-



(a)

(b)

(c)

Figure 2: (a) Effect of varying the number of negative samples on model performance for each dataset. The average precision value gradually increases as number of negative samples increases. Values are averaged over both test cases. (b) Effect of varying embedding size on model performance. Precision is averaged over both test cases per dataset. (c) Execution time (train and test) for each of the methods on all datasets. *Comprex* time shown for small (2000) test set, excluding US-2.

Table 4: Interpretability: Average hit rate, averaged over both test cases for each of the datasets.

Selection	PE	CI	CE	US-1	US-2
Random	0.37	0.41	0.41	0.41	0.41
Entity score	0.57	0.58	0.61	0.58	0.60

quently, records consisting of co-occurring entities in multiple contexts have similarly aligned vectors and will have a higher score. Conversely if a record contains a set of entities that do not usually co-occur, the vectors representing these entities will not be similarly oriented as the *context*. Thus such a record will be scored low. This idea is shown in Figure 3. As shown in Eq. (1), z_j^r is the weighted embedding vector of the entity of domain j in multi-relation r. Let M_r be the mean of the weighted vectors. The mean is taken since it is a scaled sum of the embeddings that determine the final score. Let the entity score be denoted as q_j^r, which is the scalar projection of u_j^r on M_r, shown in Eq. (6).

$$M_r = \frac{1}{|j|}\Sigma_j z_j^r \; ; \quad q_j^r = \frac{(z_j^r)^T M_r}{||M_r||_2} \qquad (6)$$

The entity score is meant as an indicator of to what extent an entity is interesting in the context and can provide the user suggestions towards understanding the anomaly. The mean of the entity scores is taken, denoted as $\overline{q^r}$. Entities in the record, whose scores are lower than $\overline{q^r}$, $q_j^r < \overline{q^r}$ are highlighted as interesting. Let there be p perturbations in the anomaly instance and let our approach suggest q domains to investigate. Hit rate (HR) is calculated as $|(p \cap q)|/|p|$. Approximately 50% improvement in pointing out entities to investigate over a random approach is achieved, shown in Table 4. While the approach does not find the exact cause, it provides the user an effective guidance towards analysing an anomaly.

6 Application Framework

Domain Knowledge Integration

Integration of domain knowledge is crucial to the creation of an effective framework for detecting suspicious timber trades. Data from environmental conservation oriented sources like International Union for the Conservation of Nature (IUCN 2019), CITES (UNEP 2019), and World Wildlife Fund (Walker 2015) contain scientific names, common names, and region information. These are ingested and processed. Botanical terms (family, genus, species) and keywords pertaining to relevant high risk timber are extracted using labels internal to these data sources and inputs from our partnering domain experts. They are processed and collated to obtain the set of high risk timber specific flora with common names, genus and species for each. This is a challenging task as data is scattered and nomenclatures are often incomplete or have multiple conflicting versions.

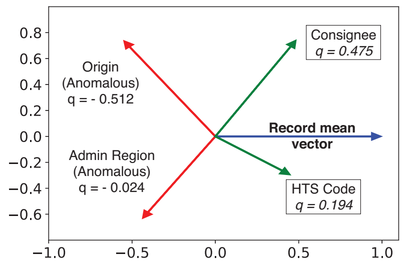
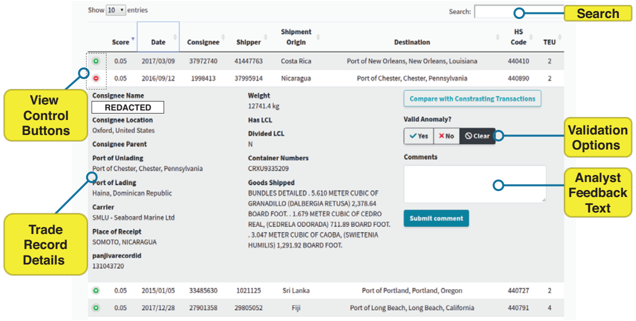


Figure 3: A 2-D projection showing contextually anomalous entities are divergent from the record mean, whereas the normal entity is aligned. Record ID:202859875 , dataset CI

Harmonized System (HS) codes and Harmonized Tariff Schedule (HTS) codes describe the type of goods traded. Their text descriptions contain some scientific (mostly genus level) and common names of plants (Chan et al. 2015). These have a hierarchical structure with a staggered granularity among the first two, four and six digits. The first six digits are standardized by World Customs Organization, and four additional country specific digits add further granularity. We use the six digit HTS codes as the trade data utilized for this work has the last four digits redacted (Panjiva 2019). HTS code data consisting of codes and their respective descriptions at all resolutions are collected, cleaned and processed to obtain associated keywords at the resolution of six digits. The HTS codes are utilized in multiple stages. Firstly,



Show 10 ▾ entries   Search:

	Score	Date	Consignee	Shipper	Shipment Origin	Destination	HS Code	TEU
⊕	0.05	2017/03/09	37972740	41447763	Costa Rica	Port of New Orleans, New Orleans, Louisiana	440410	2
⊖	0.05	2016/09/12	1998413	37995914	Nicaragua	Port of Chester, Chester, Pennsylvania	440890	2

Consignee Name
REDACTED
Consignee Location
Oxford, United States
Consignee Parent
Port of Unlading
Port of Chester, Chester, Pennsylvania
Port of Lading
Haina, Dominican Republic
Carrier
SMLU - Seaboard Marine Ltd
Place of Receipt
SOMOTO, NICARAGUA
panjivarecordid
131043720

Weight
12741.4 kg
Has LCL
Divided LCL
N
Container Numbers
CRXU9335209
Goods Shipped
BUNDLES DETAILED . 5.610 METER CUBIC OF GRANADILLO (DALBERGIA RETUSA) 2,378.64 BOARD FOOT. . 1.679 METER CUBIC OF CEDRO REAL, (CEDRELA ODORADA) 711.89 BOARD FOOT. . 3.047 METER CUBIC OF CAOBA, (SWIETENIA HUMILIS) 1,291.92 BOARD FOOT.

Valid Anomaly?
Comments

	Score	Date	Consignee	Shipper	Shipment Origin	Destination	HS Code	TEU
⊕	0.05	2015/01/05	33485630	1021125	Sri Lanka	Port of Portland, Portland, Oregon	440727	2
⊕	0.05	2017/12/28	27901358	29805052	Fiji	Port of Long Beach, Long Beach, California	440791	4

Figure 4: User interface of application framework with collapsed and expanded views.

trade records are selected that match domain expert specified two and four digit HTS codes (e.g. 44, 9303) which are known to contain timber products. After the anomaly detection model (MEAD) is run and the records are scored, specific human defined filters based on six digit HTS codes are applied to highlight actionable records.

Using regular expressions and n-gram based keyword matches on text descriptions associated with six digit HTS codes that may contain known high risk timber from collated domain specific sources are extracted. Furthermore, HTS codes covered by legislation like the Lacey Act and country-specific logging and export bans are also obtained (Forest-Trends 2017, USDA 2017). Thus if a record has HTS code matching any one of these human defined filters, the analyst can further investigate the record. It is important to note that while certain HTS codes may contain high risk species they may correspond to such a large number species and be present in so many trades that a simple rule-set based matching is neither analyzable nor actionable by analysts.

Application Interface

A user interface is developed for analysts to observe, analyze and provide feedback on results. A table with two sub-views is used to display transactions, sorted by scores so as to show more suspicious records first. A collapsed view displays the most important domains for each record, which have been selected according to preference of domain experts. This provides flexibility without clutter, given the large number of domains. A further expanded view displays all the domains to present the complete information, allowing for in-depth analysis by end users. In this view the analyst can also provide feedback through dedicated buttons and a text box for comments. Given an anomaly may be differently interpreted by an analyst compared to data mining practitioner and ground truth is unavailable, user inputs can help in inter-user collaboration and subsequent refinement of system design and performance. Search and sort functions are provided in the interface, along with pagination tools to enhance usability. All development has been done using Python 3 and uses libraries such as Tensorflow, Django, spaCy and NLTK. We plan to containerize the framework for deployment by our target users. This comprises of ensuring compatibility with target infrastructure, adding capacity for dynamic data ingestion in specified formats, automated retraining of the machine learning model, and formatting output to exact specifications.

ID: **114457154** (1st percentile)

Carrier	Consignee	Lading Port	Unlading Port	Dest.	Origin	Shipper	HTS Code
0.20	-0.44	0.22	0.69	-0.01	0.33	-0.11	-0.30

ID: **117387903** (5th percentile)

Carrier	Consignee	Lading Port	Unlading Port	Dest.	Origin	Shipper	HTS Code
-0.13	-0.07	-0.15	0.35	0.26	0.24	0.05	0.15

Figure 5: Case study analysis of suspicious records at 1^{st} and 5^{th} percentiles. Interesting domains and corresponding entity scores (q_j^r) are highlighted, demonstrating interpetability of output.

7 Case Study

To demonstrate the effectiveness of our proposed framework, a case study is presented here. *Due to confidentiality reasons, we are unable to list specific names or identifying information of countries, shipments, and ports.*

For the case study, *MEAD* is trained on US Import data for 6 months of the year 2015 and data from the next 3 months is taken as the test set. The records in the test set are sorted in non-decreasing order by their scores. Subsequently HTS code based human defined filters are applied to obtain user relevant records. Anomalies are expected to be few among the data and have low scores, and thus we randomly select and investigate records with scores at different percentiles.

The first record considered (ID 114457154) has a score in the 1^{st} percentile. The record and its associated entity scores (q_j) are shown in Figure 5 with interesting entities highlighted in red. There is no preexisting record of a transaction between this consignee and this shipper. The consignee also has never previously imported items with the HTS code 940161, which can contain endangered timber. Further, this shipper has no previous record of shipping to this destination and neither the shipper nor consignee has had any previous shipments through this carrier. Thus there are multiple unexpected co-occurrences present in the record, which is interpreted as anomalous and is therefore deemed potentially suspicious. Here the entity scores successfully highlight the domains that constitute the unusual co-occurrences. The next record we analyze (ID 117387903), is in the 5^{th} percentile. We note that both the shipper and the consignee have previous transactions involving the HTS code 440792. However, it is found that there are no previous transactions involving this consignee and this carrier. Furthermore, this carrier has never previously transported items with this HTS code. The shipper has not previously shipped through either this carrier or this lading port. Thus there are multiple unexpected co-occurrences which makes this record potentially suspicious. The record and its associated entity scores are shown in Figure 5. Similar to the previously analyzed case, the interesting entities are highlighted by our method.

The last record we present (ID 116654151) has a score in the 90^{th} percentile (not shown for brevity). The corresponding HTS code 440710 matches the list of curated HTS codes that can contain high risk timber. However the lowest frequency of entity co-occurrence among entities of the record is 5, between the consignee and the shipper. The next lowest co-occurrence frequency is 10, between consignee and destination. So this record does not present an instance of unexpected trading pattern, with other entity combinations co-occurring with higher frequency. This record has been aptly scored high by our model, reflecting a higher chance of being normal. The above examples, where suspicious records are correctly scored low and normal ones are scored higher, demonstrate the overall effectiveness of our method in finding potentially suspicious trade records.

8 Conclusion and Future Work

Detection of potentially suspicious and possibly illegal timber trade, which bears ecological and economic detriments, is a compelling yet challenging problem. In our framework we combine machine learning and domain knowledge to propose a solution to this task. The proposed embedding-based anomaly detection approach achieves better accuracy and scalability than the state-of-the-art in our experiments on trade data. Case studies demonstrate how our framework can effectively identify potentially suspicious trades in real large-scale data. The framework is designed with an interface that focuses on functionality and flexibility. The input provided by domain experts and users is important for future research that can include semi-supervised or active learning-based approaches, and can lead to improvements in model interpretability. Ablation studies, that is, comparing performance across control groups of analysts with and without the framework for the given task, would constitute a more formal evaluation. The key metrics would include time taken to accomplish tasks that identify suspicious trades and throughput efficiency, precision and recall. Our target users are interested in automating their workflow at present and our method performs well, as demonstrated.

Refining the overall system through such an in-situ evaluation process to fine tune the interaction between users, interface and algorithm is part of planned future work. There is room for research into improving our method and framework such as taking into account concept drift in trade records, dealing with suspicious trade instances which are non-anomalous and customizing the user interface further as per requirements of analysts. Thus our current work motivates a continuing research direction.

References

Abdallah, A.; Maarof, M. A.; and Zainal, A. 2016. Fraud detection system: A survey. *Journal of Network and Computer Applications* 90–113.

Aggarwal, C. C. 2017. An introduction to outlier analysis. In *Outlier analysis*. Springer. 1–34.

Akoglu, L.; Tong, H.; Vreeken, J.; and Faloutsos, C. 2012. Fast and reliable anomaly detection in categorical data. In *CIKM*, 415–424.

Barandiaran, I. 1998. The random subspace method for constructing decision forests. *IEEE TPAMI* 20(8):1–22.

Chan, H.-K.; Zhang, H.; Yang, F.; and Fischer, G. 2015. Improve customs systems to monitor global wildlife trade. *Science* 348(6232):291–292.

Chen, T.; Tang, L.-A.; Sun, Y.; Chen, Z.; and Zhang, K. 2016. Entity embedding-based anomaly detection for heterogeneous categorical events. In *IJCAI*, 1396–1403.

Chen, J.; Sathe, S.; Aggarwal, C.; and Turaga, D. 2017. Outlier detection with autoencoder ensembles. In *SDM*, 90–98.

Das, K., and Schneider, J. 2007. Detecting anomalous records in categorical datasets. In *KDD*, 220–229.

Forest-Trends. 2017. Known forest product export bans. https://www.forest-trends.org/known-log-export-bans. Last accessed: 2019-11-01.

Goldstein, M., and Uchida, S. 2016. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one* 11(4).

Gutmann, M., and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, volume 9, 297–304.

Hu, R.; Aggarwal, C. C.; Ma, S.; and Huai, J. 2016. An embedding approach to anomaly detection. In *ICDE*, 385–396.

IUCN. 2019. The IUCN Red List of Threatened Species. https://www.iucnredlist.org. Last accessed: 2019-11-01.

Lawson, S., and MacFaul, L. 2010. *Illegal logging and related trade: Indicators of the global response*. Chatham House London.

LeCun, Y. A.; Bottou, L.; Orr, G. B.; and Müller, K.-R. 2012. Efficient BackProp. In *Neural networks: Tricks of the trade*. 9–48.

Lp, X., et al. 2018. Transaction fraud detection using gru-centered sandwich-structured model. In *CSCWD*, 467–472.

Panjiva. 2019. Panjiva trade data. https://panjiva.com. Last accessed: 2019-11-01.

Smets, K., and Vreeken, J. 2011. The odd one out: Identifying and characterising anomalies. In *SDM*, 804–815.

Taha, A., and Hadi, A. S. 2019. Anomaly detection methods for categorical data: A review. *ACM Computing Surveys*.

Tang, G.; Pei, J.; Bailey, J.; and Dong, G. 2015. Mining multidimensional contextual outliers from categorical relational data. *Intell. Data Anal.* 19(5):1171–1192.

UNEP. 2019. The Species+ Website. Nairobi, Kenya. https://www.speciesplus.net. Last accessed: 2019-11-01.

USDA. 2017. Lacey Act Declaration Requirement. https://www.aphis.usda.gov/aphis/ourfocus/planthealth/import-information/sa_lacey_act. Last accessed: 2019-11-01.

Vreeken, J., et al. 2011. Krimp: mining itemsets that compress. *Data Min. Knowl. Disc.* 23(1):169–214.

Walker, L. 2015. WWF's Global Forest and Trade Network: Country Profiles. Technical report, World Wildlife Fund.

Wang, D.; Jiang, M.; Zeng, Q.; et al. 2018. Multi-type itemset embedding for learning behavior success. In *KDD*, 2397–2406.

Wiedenhoeft, A. C., et al. 2019. Fraud and misrepresentation in retail forest products exceeds US forensic wood science capacity. *PloS one* 14(7).

Zhang, K.; Wang, Q.; Chen, Z.; et al. 2015. From categorical to numerical: Multiple transitive distance learning and embedding. In *SDM*, 46–54.