A Framework to Assess Risk of Illicit Trades Using Bayesian Belief Networks

Rashid Anzoom, Rakesh Nagi, Chrysafis Vogiatzis

A Framework To Assess Risk of Illicit Trades Using Bayesian Belief Networks

Rashid Anzoom[1[0000−0003−4699−6379]], Rakesh Nagi[1[0000−0003−4022−6277]], and Chrysafis Vogiatzis[1[0000−0003−0787−9380]]

University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
Department of Industrial and Enterprise Systems Engineering
{ranzoom2,nagi,chrys}@illinois.edu

Abstract. Recent years have seen the initiatives against illicit trades gain significant traction at both national and global levels. A crucial component in this fight is correct assessment of the risks posed by different trades across different regions. To aid in this cause, we provide a risk prediction framework based on Bayesian Belief Networks. It involves the development of a causal model incorporating variables related to the rise/decline of the illicit trade volume. The influence of these variables are determined by training on available data that are allowed to update over time. Implementation on a sample case study shows relatively low prediction accuracy of our model. Factors constraining its performance are analyzed and possible ways to avert them are discussed. We expect this framework to act as a decision support tool to the policymakers and strengthen them in the fight against illicit trades.

Keywords: Illicit Trade · Risk Assessment · Bayesian Belief Networks.

1 Introduction

The sustained growth and diversification of illicit trades remain a substantial threat worldwide [1]. Recent statistics report the annual turnover of these trades to be in the order of 2.2 trillion US dollars, which corresponds to approximately 8–15% of the global GDP [11]. Efforts to control/disrupt illicit trade are heavily constrained by the scarcity of resources (e.g., workforce, finance), emphasizing the importance of allocating them efficiently. Before planning for disrupting illicit trade, though, we need reasonable estimates of the proliferation risks of different trades in distinct regions of the world.

One could base this estimation process on the available statistics; however, such an approach is risky since data on illicit supply chains are rare and suffer from multiple shortcomings (including incompleteness, unclear boundary specification, low dynamics) [1]. An alternate strategy involves discerning factors that affect the rise or decline of these trades, as well as their degree of influence which we assume to be probabilistic. Available information on these variables will then lead to the estimation of risk. In this paper, we introduce a framework to conduct this assessment in a systematic manner. In particular, we employ Bayesian Belief Networks (BBN) modeling, which is widely known for dealing with probabilistic inference in complex systems [18].

Prior to discussion of the proposed approach, we review previous works related to our research problem. Existing literature on this topic is mostly qualitative in nature: identifying specific and general risk factors for different trades [1]. A few quantitative models have been proposed to predict individual trade risk/volume. For instance, growth of narcotics market was modeled as a function of a small number of variables (e.g., utility, risk, enforcement level, economic status) [2, 5]. Researchers have also employed BBN models in some of these works. An expert-driven model is presented at [8] to predict the occurrence of poaching event of rhinos. Data-driven models aimed to predict deforestation, which is related to illegal logging [20, 12]. Alternate methodologies used here include artificial neural network and Gaussian process. Apart from these, BBN has also been used to detect activities in illicit trades (e.g., prediction of fraudulent food products) [21]. Using separate prediction models for each trade category might not be convenient for the policymakers. Perhaps with that in mind, The Economist Intelligence Unit [22] developed a country-wise risk index to indicate the prevalence risk of illicit trades. However, the index is linear in nature and defined by expert opinion. Furthermore, it does not distinguish different trade categories and focuses on the regional factors only. Thus, there is need to develop a model that can predict the risk of individual trade categories at different regions as well as provide an aggregated risk score for a particular area or trade. Our framework aims to build a customized model with these features.

The remainder of this paper is organized as follows. Section 2 describes the steps in developing and implementing the framework. Section 3 presents a case study for demonstrating our framework. Our discussion concludes in Section 4 through a critical review of the approach and a brief discussion of future avenues.

2 Risk assessment framework

The framework to assess the risk of illicit trades consists of five steps as shown in Figure 1. The following subsections discuss these steps sequentially.

2.1 Identification of factors related to illicit trades

The discussion on the causes behind the inception and continuation of illicit trades can be found in several places of the literature. Some of the established factors include poor socio-economic conditions, operational risk or lack thereof, the potential return on investment, etc. We divide these factors into two sets: region-specific and trade-specific. The Economist Intelligence Unit (EIU) considered 20 features of the first category to predict regional suitability for illicit trades [22]. The second set, on the other hand, denotes the expected supply chain performance in a business. We propose to incorporate the drivers of the

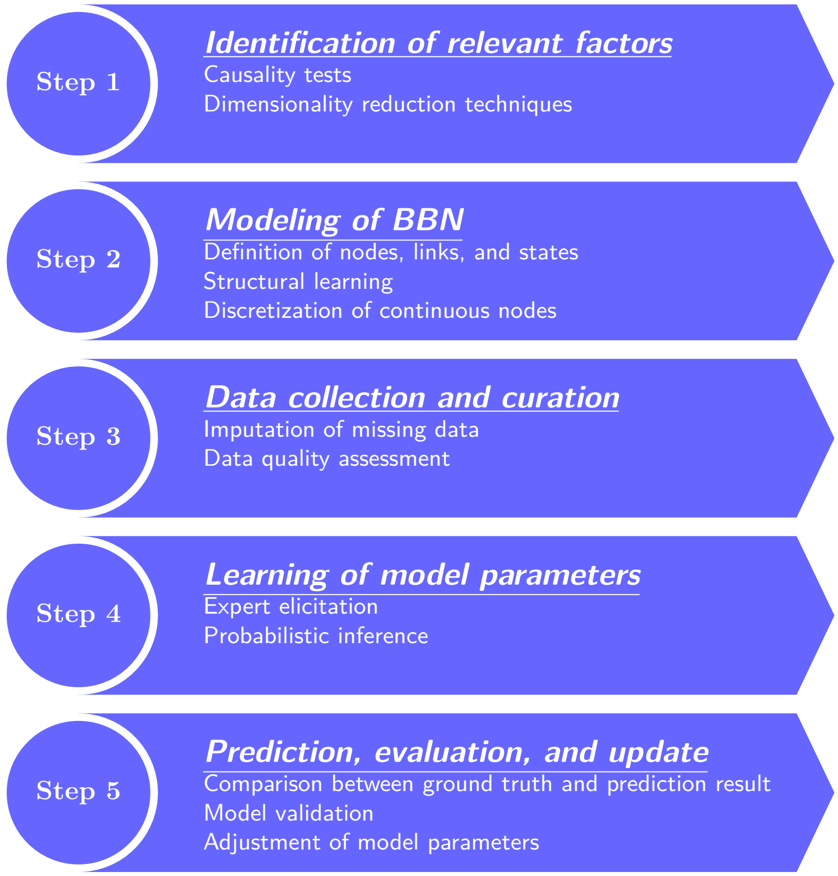

Fig. 1. The five main steps for our illicit trades risk assessment framework.

illicit supply chain in this set (e.g., ease of production/acquisition, storage, concealment and transportation, access to resources, profitability, money laundering, corruption). Beyond this, one can always incorporate additional variables through causality tests, i.e., statistical hypothesis testing. Ordiano [16] recently demonstrated another approach using node embedding and clustering. For customization, it is also possible to filter out some variables through dimensionality reduction techniques (see [23] for details).

2.2 Modeling of a Bayesian belief network

Following the selection of variables, we move forward to modeling. Our adopted approach, BBN, is a directed acyclic graphical model of causal relationships based on probability theory [3]. In the constructed graph, nodes denote the variables of interest; while edges represent the causal relationship (dependence) between them. In our case, we can consider five sets of nodes: region-specific factors (A), trade-specific factors (T), trade statistics (S), data reliability (D), and trade risk (R). Trade statistics denote the types of data available regarding illicit trades (e.g., instances reported, transaction volume, number of arrests), while data reliability indicates the confidence in the data at hand. The trade risk is predicted as a probabilistic function of all these variables. In terms of equation, this can be written as follows:

$$P\left(S|A,T,R,D\right) = P(A)P(T)P(R|A,T)P(D)P(S|R,D). \quad (1)$$

In a simpler model, nodes within a set are considered conditionally independent of each other. For a better representation of reality, one can conduct multivariate regression analyses or causality tests among the variables. Given data availability, it is also possible to learn the model structure using several algorithms [19]. Two other salient node attributes in this model are the type and state of nodes. The model supports both discrete and continuous nodes, although the former is preferred. Continuous nodes have to go through discretization by some interval points. The methods for such discretization can be user-specified or algorithmic/data-driven (supervised or unsupervised) [4]. Once we specify all these issues, our model is ready to learn the parameters.

2.3 Data collection and curation

As mentioned earlier, the deceptive nature of illicit trades makes it difficult to collect reliable data on them. A discussion on the existing data sources is provided in [1]. Among the five mentioned sources, organizational databases remain the most useful ones for our analysis since they follow a standard format and are regularly updated. Still, it cannot be said that the available statistics accurately represent reality.

This is why it is important to include in our model information regarding the quality and the reliability of the used data. The four major dimensions of data quality are accuracy, completeness, consistency, and timeliness [17]. In the BBN, the set of nodes representing the data reliability aspect represents the state of these quality dimensions. In terms of trade statistics, one can use different types of data as nodes or apply fusion at the information level to obtain an aggregated measure.

2.4 Learning the model parameters

In this step, we train our model with the existing dataset to learn the parameters, i.e., to learn the conditional probabilities (CP). While we intend to derive these probabilities straight from the data, it is not unusual for some information to be limited in amount or to be simply unavailable. In such cases, researchers have often used expert opinions for conditional probability elicitation [14]. Considering this, we implement an inference method that is adaptable to such scenarios, including the combination of expert opinion and evidential information, as well as incomplete or small dataset. Further information regarding these algorithms can be found at [7] and [13].

2.5 Prediction, evaluation, and model update

Given the parameters and prior probabilities of the variables would provide the risk of a specific trade in a particular region as per Equation (1). If the ground truth data (actual trade risk/statistic) becomes available, one can compare it with the predicted results through different metrics (e.g., confusion matrix, k-fold cross-validation, spherical payoff) [10].

Most of the databases we use are updated on an annual basis. Incorporating these new pieces of evidence will update the model regularly, improving its adaptability. It is even possible to assign variable weights to cases as well as unlearn previous instances [15].

Finally, should the prediction error become significantly high, it becomes necessary to scrutinize the model. In this case, techniques like sensitivity analysis, cross-validation, and scenario analysis can provide insight regarding the model [6].

3 Case study

To illustrate the framework introduced in Section 2, we present a sample case study to compute the proliferation risk of illicit trades for different countries.

Step 1 We begin with the decision to include 4 trade-specific and 14 region (country)-specific factors influencing illicit trades. The region-specific factors are divided into four categories (trade transparency, supply and demand, governmental policy, and customs environment) as described in [22]). As a consequence, we introduce a node for each of them in the BBN. Two more nodes are introduced to represent the country environment and ease of trade, which influence the node *trade risk*. For data reliability, we only consider one node (completeness) since the age for all data is the same. Finally, we have nodes representing two types of trade statistics (trade volume and number of cases), taking the tally of nodes to 28. he selection of these variables was based on data availability, although information regarding the trade-specific factors was not found. For the sake of model completeness, we assign values to them based on our judgment, e.g., transportation of narcotics is considered to be easy because of its size while transportation of arms is considered difficult.

Step 2 With the selection of the variables complete, we move to build the model. Most of the variables are represented by scores and thus continuous in nature. Availability of data allows a reasonable discretization of continuous nodes. In this model, we employ the equal frequency strategy, which assumes the quartiles as cutpoints. Considering the minimum and maximum possible value of the factors, up to two more cutpoints can be included. The interval between these cut-points represents the states of the discretized nodes. Links between the nodes are assumed to be predetermined, so we do not run any causality test or structural learning. The complete structure of the causal model is visible in Figure 2. For ease of discernment, the node-sets are color-coded and a legend is provided.

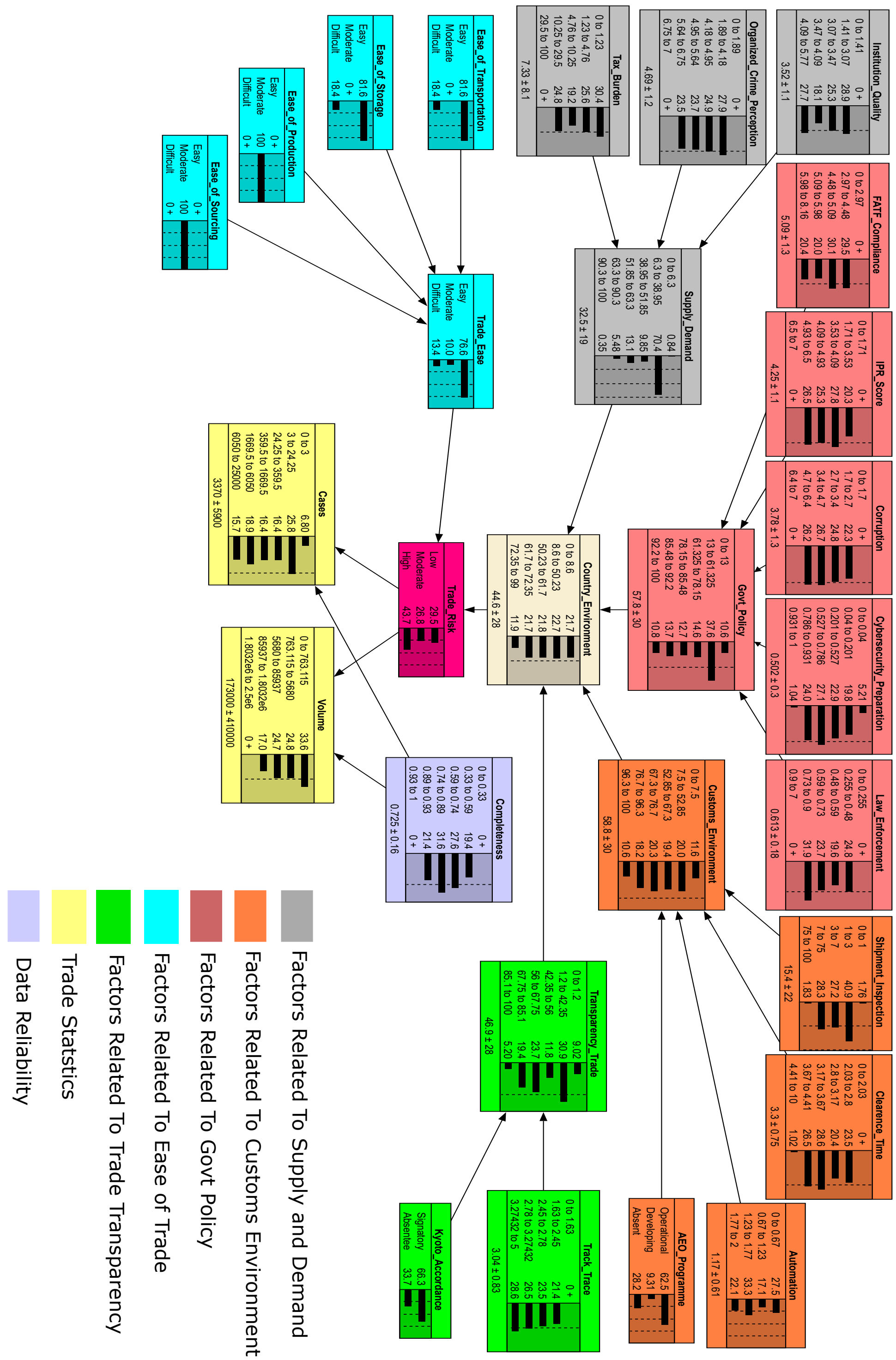

Fig. 2. The BBN model for the case study, built using NETICA.

Step 3 Our next step involves collection and preprocessing of relevant data. Information related to the factors are found on 15 separate databases, the majority of which are listed in [22]. Besides that, information on law enforcement capability, institutional quality, and corruption is extracted from the World Internal Security and Police Index[1], the GCI (Global Competitiveness Index) database, and the World Economic Forum Global Competitiveness Report[2]. On the other hand, the UNODC (United Nations Office on Drugs and Crime) database[3] provides statistics on illicit trades. No data is found on trade-specific factors, as mentioned earlier. All of the sources used fall under the category of open organizational databases. However, heterogeneity exists in their structure and the volume of data they provide. For example, UNODC provides 27 year-long data regarding drug seizure, while EIU provides tracking competence for only 2 years. Furthermore, each database holds additional information besides the desired ones. For inference of the conditional relationships between the factors, one would need data for all attributes in one single database. We achieve this by collecting the latest available records (including statistics of two trades: narcotics and arms trade) across 160 countries and merging them into a country database. We also compute the completeness and timeliness of the data. The attribute with the highest completeness was the cybersecurity index (95.63%), while the lowest was the supply-demand (45%). The lowest, highest and median tuple (countrywise) completeness was 33%, 92.6%, and 59%, respectively. The overall relationship completeness was 72.31%.

Step 4 For learning the conditional probabilities, we feed this data into NETICA[4], a software specializing in BBN modeling and analysis. Netica uses three methods to learn the conditional probabilities: count, expectation maximization (EM), and gradient descent method. The last two are used for inferring from an incomplete database, which applies to our case. The available data set was divided into two sets (learning and testing) for validation. The learning set is used to train the model, while the testing set is used for evaluating our prediction. For comparison, we train the model using both EM and the gradient descent method. Once trained, the model assumes values for the conditional probabilities. An example is shown in Figure 3. Here, we show the influence of tracking capability and the Kyoto Addendum on the trade transparency score.

Step 5 Once the training is complete, our focus shifts to assessing the accuracy of prediction. Since the actual proliferation risk is not known, the test is done on the trade statistics instead. Results show EM to be more accurate (38.46%) than gradient descent (28.85%). The scores for quadratic loss, logarithmic loss, and spherical payoff are 0.825, 3.66, and 0.457. The accuracy is still quite low for a prediction model, but considering the incompleteness of data, simpler model structure, and the inherent difficulty of assessing illicit trades, we

[1] http://www.ipsa-police.org/
[2] http://reports.weforum.org/global-competitiveness-index-2017-2018/
[3] https://dataunodc.un.org/
[4] https://www.norsys.com/netica.html

Track_Trace	Kyoto_Accordance	0 to 1.2	1.2 to 42.35	42.35 to 56	56 to 67.75	67.75 to 85.1	85.1 to 100
0 to 1.63	Signatory	16.667	16.667	16.667	16.667	16.667	16.667
0 to 1.63	Absentee	16.667	16.667	16.667	16.667	16.667	16.667
1.63 to 2.45	Signatory	1.39e-4	1.15e-4	99.999	1.48e-4	1.31e-4	1.38e-4
1.63 to 2.45	Absentee	1.40e-4	99.999	1.30e-4	1.36e-4	1.31e-4	1.40e-4
2.45 to 2.78	Signatory	7.50e-5	71.871	7.34e-5	8.50e-5	28.129	7.56e-5
2.45 to 2.78	Absentee	2.18e-4	99.999	2.13e-4	2.20e-4	1.97e-4	2.20e-4
2.78 to 3.27432	Signatory	4.36e-5	15.829	36.901	42.089	5.182	4.57e-5
2.78 to 3.27432	Absentee	1.35e-4	14.285	71.43	14.284	1.33e-4	1.36e-4
3.27432 to 5	Signatory	3.21e-5	3.431	13.656	27.548	51.899	3.466
3.27432 to 5	Absentee	3.33e-4	33.333	33.333	33.333	3.33e-4	3.33e-4

Fig. 3. Conditional Probability of Trade Transparency Score Given Track and Trace Performance and Kyoto Accordance status (Display from NETICA).

consider this value to be reasonable. The omission or imputation of incomplete tuples is expected to improve this accuracy.

Sensitivity analysis measures the degree to which variation in posterior probability distribution of the target node (risk index) is explained by other variables, i.e., how influential different nodes are in predicting risk index [9]. In this case (discrete node), the measure is entropy reduction. Table 1 shows part of our results, listing the individual factors with higher influence. As expected, trade volume and the number of cases have the greatest influence on trade risk. The next two in the list are ease of storage and transportation, which explains the higher spread of narcotics trade over arms in general. The last factor, cybersecurity preparation, is a bit surprising since the transactions are mostly done physically. This might be an issue worth investigating in the future. Once these results are validated, one can predict the risk with greater confidence. We have left the trade risk here as a categorical variable. That said, it is also possible to consider it as an index of continuous nature. The same methodology can be applied for different scales of regions (continent, country, state/district), given, of course, that data for different regions are made available.

Table 1. Sensitivity of Trade Risk To Other Node Findings.

Node	**Mutual info (entropy)**	**Percent**	**Variance of belief**
Trade volume	0.33	21.3	0.429
Cases	0.082	5.26	0.012
Ease of storage	0.019	1.24	0.004
Ease of transportation	0.019	1.24	0.004
Track and trace	0.009	0.601	0.001
Cybersecurity preparation	0.006	0.419	0.01

4 Conclusion

The ability to predict trade proliferation risk is expected to have great value in the fight against the illicit economy. However, it requires handling a multitude of challenges. This paper discusses the main ideas behind building such a prediction model, and demonstrates the challenges of building one through a case study. The first issue requiring attention is the scarcity and limited availability of data since it has significant impact on the prediction performance. None of the trade-specific factors in the model are clearly defined in the literature. Development of measures for them would be a good idea for future research. For data reliability assessment, timeliness of data merits inclusion. However, the measure of volatility (duration of data validity) in the case of illicit supply chain needs to be set *a priori*. Caution should also be exercised while choosing the method for discretizing continuous variables. In the equal-frequency method, updating the parameters would alter the probability distributions to some extent but the quartiles remain the same, thus affecting the expected value. Finally, if data updated annually become sufficiently available, one should consider detecting dynamics both within and between different variables over time. Dynamic Bayesian networks should be a good choice to carry out this work. These recommendations would be implemented in our future works to overcome the aforementioned limitations and build a workable model with higher prediction accuracy and greater insights.

References

1. Anzoom, R., Nagi, R., Vogiatzis, C.: A review of research in illicit supply-chain networks and new directions to thwart them. IISE Transactions (2021). https://doi.org/10.1080/24725854.2021.1939466
2. Baveja, A., Jamil, M., Kushary, D.: A sequential model for cracking down on street markets for illicit drugs. Socio-Economic Planning Sciences **38**(1), 7–41 (2004)
3. Ben-Gal, I.: Bayesian networks. In: Encyclopedia of Statistics in Quality and Reliability, vol. 1. Wiley & Sons (2008)
4. Beuzen, T., Marshall, L., Splinter, K.D.: A comparison of methods for discretizing continuous variables in Bayesian networks. Environmental Modelling & Software **108**, 61–66 (2018)
5. Caulkins, J.P., Padman, R.: Interdiction's impact on the structure and behavior of the export-import sector for illicit drugs. Zeitschrift für Operations Research **37**(2), 207–224 (1993)
6. Chen, S.H., Pollino, C.A.: Good practice in Bayesian network modelling. Environmental Modelling & Software **37**, 134–145 (2012)
7. Ji, Z., Xia, Q., Meng, G.: A review of parameter learning methods in Bayesian network. In: Huang, D.S., Han, K. (eds.) Advanced Intelligent Computing Theories and Applications. pp. 3–12. Springer International Publishing (2015)
8. Koen, H., de Villiers, J., Roodt, H., de Waal, A.: An expert-driven causal model of the rhino poaching problem. Ecological Modelling **347**, 29–39 (2017)
9. Li, C., Mahadevan, S.: Sensitivity analysis of a Bayesian network. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering **4**(1) (2018)
10. Marcot, B.G.: Metrics for evaluating performance and uncertainty of Bayesian network models. Ecological Modelling **230**, 50–62 (2012)
11. Mashiri, E., Sebele-Mpofu, F.Y.: Illicit trade, economic growth and the role of customs: A literature review. World Customs Journal **9**(2), 38–50 (2015)

12. Mayfield, H., Smith, C., Gallagher, M., Hockings, M.: Use of freely available datasets and machine learning methods in predicting deforestation. Environmental Modelling & Software **87**, 17–28 (2017)
13. Mkrtchyan, L., Podofillini, L., Dang, V.N.: Bayesian belief networks for human reliability analysis: A review of applications and gaps. Reliability Engineering & System Safety **139**, 1–16 (2015)
14. Mkrtchyan, L., Podofillini, L., Dang, V.N.: Methods for building conditional probability tables of Bayesian belief networks from limited judgment: An evaluation for human reliability application. Reliability Engineering & System Safety **151**, 93–112 (2016)
15. Nielsen, T.D., Jensen, F.V.: Bayesian Networks and Decision Graphs. Springer New York (2007)
16. Ordiano, J.Á.G., Finn, L., Winterlich, A., Moloney, G., Simske, S.: A method for estimating driving factors of illicit trade using node embeddings and clustering. In: Mexican Conference on Pattern Recognition. pp. 231–241. Springer (2020)
17. Pipino, L.L., Lee, Y.W., Wang, R.Y.: Data quality assessment. Communications of the ACM **45**(4), 211–218 (2002)
18. Pourret, O., Naïm, P., Marcot, B.: Bayesian Networks: A Practical Guide to Applications. John Wiley & Sons (2008)
19. Scanagatta, M., Salmerón, A., Stella, F.: A survey on Bayesian network structure learning from data. Progress in Artificial Intelligence **8**, 425–439 (2019)
20. Silva, A.C., Fonseca, L.M., Körting, T.S., Escada, M.I.S.: A spatio-temporal Bayesian network approach for deforestation prediction in an Amazon rainforest expansion frontier. Spatial Statistics **35**, 100393 (2020)
21. Soon, J.M.: Application of Bayesian network modelling to predict food fraud products from china. Food Control **114**, 107232 (2020)
22. The Economist Intelligence Unit Limited: The global illicit trade environment index (2018)
23. Van Der Maaten, L., Postma, E., Van den Herik, J.: Dimensionality reduction: A comparative review. Journal of Machine Learning Research **10**(66-71), 13 (2009)