MULTIVARIATE FREIGHT DATA GENERATION FOR ASSESSING CUSTOMS' RISK EVALUATION TOOLS

Farzad Kamrani[(a)], Pontus Hörling[(b)], Thomas Jansson[(c)], Pontus Svenson[(d)]

[(a),(b),(c),(d)]Department of Decision Support Systems
Division of Information and Aeronautical Systems
Swedish Defence Research Agency
SE-164 90 Stockholm, SWEDEN

[(a)]kamrani@kth.se, [(b)]hoerling@foi.se, [(c)]thomas.jansson@foi.se
[(d)]pontus.svenson@foi.se

ABSTRACT

Nowadays, enormous volumes of goods are transported all over the world in containers, mostly in container ships. It is a formidable task to ensure that no illegal activities such as smuggling of prohibited goods that violate customs rules and regulations occur. Many indications on potential smuggling can be found just by scrutinizing all documentation that follows these containers, and the goods within them. In the EU's container security project (CONTAIN), one among many important tasks has been to present a prototype document analysis system for customs, called CustAware, to find these indications. The standardized way to submit necessary information to customs today is via Entry Summary Declarations (ENS) messages, normally in electronic form. We have developed a simulator (the ENS-Simulator) that generates a high flow of realistic customs-oriented data contained in such ENS messages. By adding a much smaller amount of ENS messages that could indicate peculiar shipments the data is used to train CustAware to find the needles (indications on strange activities) in the haystack (all normal ENS messages). In this paper, we describe the process of generating a high rate of normal messages in a realistic way, to represent a typical message inflow at a large customs risk assessment and management centre.

Keywords: container traffic, multivariate data generation

1. INTRODUCTION

In today's global world, enormous amounts of goods are transported along supply chains between producers and consumers. Of particular importance is the shipping container traffic on the oceans and between sea and inland ports, accounting for about 70% of all cargo transport. Inland ports often work as shunting and reloading sites between different means of transport such as trucks, barges and trains. To ensure that all steps from sender (”consignor”) to receiver (”consignee”) function smoothly; supply chain security therefore is a subject that continuously attracts more attention. The challenge is to use the available data and information in order to detect and prevent threats to the supply chain.

There is a lot of data available in logistics that could potentially be used for security purposes. Different stakeholders have different needs of control of the container and goods flow. Business partners must exchange commercial and logistic information, also known as Business to Business (B2B) information, in order to guarantee reliable, economically profitable and efficient goods transfer as well as secure transfer of payment. What is basically important is to satisfy the needs of the consignors and consignees of the goods. There are freight forwarders, port operators, shipping lines, truck, barge and train companies and different kinds of brokers involved in the process with the common goal to satisfy those needs.

Customs' needs are largely different from business partners' needs. Customs' task is to facilitate trade and collect customs duties, as well as to control the flow of goods so no illegal or hazardous goods enter national borders. The information customs needs is partly requested from the business side, known as Business to Authority (B2A) information, and has some overlap with B2B information. They also need other types of information, such that can be provided by intelligence services like customs' own confidential sources, nationally and internationally, police, military and coast guard.

Threats to the supply chain of relevance for Europe can be divided into different categories:

- Threats towards Europe through the supply chain. This encompasses smuggling of illegal and counterfeit goods, pirate copies, and smuggling of dangerous substances and weapons in order to perform a terrorist attack inside Europe.
- Threats towards the supply chain itself. This could be theft and intrusions into ports and terminals or terrorist attacks against supply chain infrastructure.

- Natural disasters and accidents. This category is characterized by the lack of an antagonistic opponent.

In order to meet the challenges posed to Europe's supply chain, the European Commission (EC) included a large portfolio of supply chain security projects in the EU's Seventh Framework Programme for Research (FP7 2014) agenda. In this paper, we describe parts of the CONTAIN project, which started in October 2011 and had its final demonstrations in the first quarter of 2015. CONTAIN deals with several aspects of container security, from advanced sensors for positioning and localization to shared infrastructure and message formats for information sharing to decision support systems (DSS). Addressing the first of the threats above, CONTAIN has developed the CustAware system (Brynielsson, Westman, and Svenson 2015), which aims to provide customs inspectors with a DSS that enables them to determine, which containers to inspect manually and/or using sensors (*scanning*) based on all available information. There is a strong need for systems like this, since resource constraints put a limit on the number of containers that are inspected, especially in large ports where it is only possible to inspect a few percent of all containers that pass through. Choosing *which* containers to inspect (*targeting*) is one of the most important tasks for customs.

CustAware will enable customs inspectors to work more dynamically with the risk profiles that they are currently using, and shows how they could update the models continuously based on feedback from inspections and police investigations. In order to test this decision support tool, there is a need for realistic data. However, it is normally difficult to get real B2B (Business to Business) or B2A (Business to Authorities) data for evaluation purposes. B2B data is often commercially proprietary and only available for collaborating companies. B2A data can most often not be shared by customs to other authorities or companies for confidentiality reasons. Businesses must be able to rely on customs keeping their data secure in order to be willing to share their data with them. Customs themselves must be guaranteed that the data and methodology they use for finding indicators for illegal trade are not revealed.

Hence, in order to test CustAware and other tools developed in the CONTAIN project, we need a simulator that can be used to provide data that is similar to real B2A data. In this paper, we describe the simulator developed for this purpose.

An important part of the CONTAIN project is the evaluation of results. The technological advances developed in CONTAIN presented in the final demonstrations are used as the basis for an assessment and evaluation procedure, following the procedure developed by (Asp, Carling, Hunstad, Johansson, Johansson, Nilsson, and Waern 2008).

The present paper contributes both to the development of CustAware and its evaluation. In order to develop the methods and procedures for updating the risk profiles, it is necessary to have realistic data to work with. The outputs of the simulator can also be used to evaluate the effectiveness of the procedures for dynamic risk profile adaptation that are developed in CONTAIN.

2. RELATED WORK

Modelling and simulation (M&S) based methods have been used in marine environments such as harbours in order to support decision makers in operational procedures. Tremori, Massei, Madeo, and Reverberi (2013) provide an overview on a combined approach using M&S and data fusion techniques to analyse complex maritime scenarios involving asymmetric threats. Longo (2010) uses simulation and design of experiments techniques to develop analytical models that express the operational efficiency as function of critical parameters of the inspection process.

As mentioned in the introduction, EU is currently funding a large number of projects addressing supply chain security. In this section, we briefly describe some of these projects and how the work presented here complements them.

SAFEPOST, EUROSKY and CASSANDRA address the first risk mentioned above (threats towards Europe using the supply chain). SAFEPOST deals with postal security and is developing an information-sharing platform and advanced sensors that will enable faster and more accurate detection of dangerous substances in parcels. The EUROSKY project deals with similar matters as CONTAIN, but for air cargo transport instead of shipping containers, while CASSANDRA also focuses on shipping containers. In CASSANDRA, the main emphasis is on the data pipeline concepts, which will enable customs to use business information in a better way.

The second risk (threats towards the supply chain itself) is also the subject of a number of projects, amongst others SUPPORT, ARENA, PROTECTRAIL, and TASS. The subject of SUPPORT is port security and in particular perimeter security for ports using advanced sensors and information fusion systems. PROTECTRAIL and TASS are devoted to railroad and airport security, respectively, while ARENA deals with securing transports in motion, studying both ship and truck security.

Finally, the EC is also looking into ways of handling natural disasters and accidents, for example in the Demonstration Programme Driver. The Demonstration Programme CORE also aims to collect the results from all previous supply chain security projects and show how everything fits together.

In some of these projects, simulators have been used to construct synthetic data to be used to test the developed tools. However, none of them has developed a simulator for generating freight data. To our knowledge, the present work is the first that deals with generating multivariate freight data to enable realistic evaluation of supply chain security systems.

3. CONTAINER SECURITY DATA AND INFORMATION

Different types of information are generated before and during container transport. For keeping track of physical status of the container and the goods therein, electronic door seals, as well as, e.g., light, temperature, humidity and CO2 sensors can be placed on or within containers to detect when the container door is opened, or to detect anomalous changes of the physical parameters measured. Alarms or logs from such detectors can be used to indicate abnormal events from the container, e.g., cooling failure for a container with such functionality, or attempts to break into and smuggle items, animals or people. Furthermore, positioning devices can be used to keep track of a container's location. One of the motivations for this is to log the positions and times during land transport for logistics as well as security purposes. Unjustified stops or deviations from planned truck routes can indicate suspicious activities. Information messages are also routinely generated at certain events; from first load to final unload, during fully normal logistic handling of a container along its route. This is done to log its transport status, and to transfer necessary documentation on its contents and planned final route. Adherence to national and European level work laws can also be verified by checking such log information.

3.1. Information for Customs

Since the CONTAIN project is oriented towards customs' needs, we have focused on how to simulate the requested B2A information exchange in the container handling process, more specifically that requested by customs. Today, this B2A information is most often submitted as so-called Entry Summary Declaration (ENS) messages (European Customs Information Portal 2013); a pre-arrival information sent to the customs authority for all goods entering the EU, according to EU legislation (Taxation and Customs Union - European Commission 2006).

Within the EU, an ENS should be lodged at the first port of entry into the EU by the carrier or by an agent acting for the carrier. Normally, the ENS for containerized goods should be formally lodged at least 24 hours before the goods is loaded at the first port of departure, which is often many days or several weeks before it is expected to reach its final port of destination. The ship might be visiting several transit ports in between, which in many cases are not known whilst creating the ENS-document. In such case the ENS declarant is responsible for sending updated ENS information to customs. In any case, the time span between lodging of the ENS and the call of the container ship at the first port of entry in Europe can be used by customs to reason about what goods will probably arrive with the ship, and what actions to take. In some situations, this can even lead to a so-called "Do Not Load" response before the ship leaves the port, in which case the container cannot be loaded on board at all. Thus, customs needs an effective DSS for finding containers, which have a higher than average probability of carrying unmanifested or prohibited material. Especially, having in mind the enormous volumes of goods that can be transported on a single large-size container ship (a single ship can carry up to 18 270 TEU (ten-foot equivalent units), i.e., 18 270 half-size or 9 136 full-size containers).

3.2. Implementation of the Entry Summary Declaration

Each member state within the EU has to implement an Import Control System (ICS) according to a specific architecture (European Customs Information Portal 2013). However, front-end systems for data entry differ between nations. Large shippers of course interface directly to the customs information system and submit bulk data.

In some cases some of the data attributes are allowed to be placed at different levels; at the declaration item of goods level, at the declaration header level, and at the conveyance report level. Moreover, different member states have different sanction possibilities regarding the quality of the submitted ENS document, meaning that some countries simply reject the transport if not certain quality level of submitted data is achieved, while other countries have no legal support for doing so. Altogether, these conditions have led to different ICS-systems, with different ENS document structures having different levels of quality.

4. GENERATING CRS

Within several earlier freight-oriented EU projects, the so-called Common Framework (CF) information model was developed (e-Freight 2013a), see Figure 1.

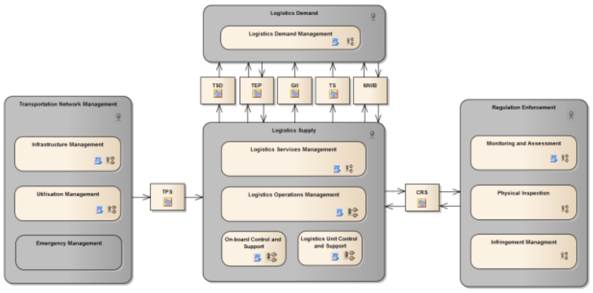
Logistics Demand
Logistics Demand Management
TSD
TEP
GII
TS
MWB
Transportation Network Management
Infrastructure Management
Utilisation Management
Emergency Management
TPS
Logistics Supply
Logistics Services Management
Logistics Operations Management
On-board Control and Support
Logistics Unit Control and Support
CRS
Regulation Enforcement
Monitoring and Assessment
Physical Inspection
Infringement Managment

Figure 1: A high-level picture of the Common Framework. CRS forms the connection between "Regulation Enforcement" and "Logistics Supply". Picture from the e-Freight project.

CF extends the OASIS Universal Business Language (UBL) model intended for standardized B2B information transactions in XML (OASIS). The extensions are mainly to provide more interoperability between logistics stakeholders and to enable the expression of B2A data. The latter (B2A) subset of CF, called the Common Reporting Schema (CRS), is "a single, standardized, electronic reporting document which includes all the information fields which are necessary and sufficient for reporting to authorities in all member states and across all modes" (e-Freight

2013b). Parts of CF have been tested (e.g., in Gothenburg port), but the overall framework is still not in operation. However, being intended as a pure electronic way to exchange information, it will probably be recommended within the EU countries. The CRS format is implemented as an XML Schema, containing all necessary fields for mapping ENS data requirements. Table 1 demonstrates a suggested mapping between the ENS data requirements to corresponding fields within the CRS. The ENS simulator to be described in the following produces a Comma Separated Values (CSV) list of simulated text entries, such as consignor name, address, goods commodity code, etc., that must fill the corresponding fields in a valid ENS, or as in our case, CRS message. CRS messages are required to be serializable and exchanged in some kind of self-describing format, such as JSON or XML reflecting its semantic (tagged) content. In the CRS generator, written in Java, we have chosen to parse the simulated CSV-formatted ENS message and from this build up a Java CRS object with fields that reflect the message, followed by serialization of this object to XML using the Java XML Binding (JAXB) framework. This automatically structures the ENS fields of the CRS message in the proper hierarchical format of such messages. The message is then regarded as complete to submit to a hypothetical customs office, in our case an instance of CustAware that parses the messages and analyses their content on the fly for indicators on malicious goods or activities. The submission uses the same web service REST communication paradigm (Fielding and Taylor 2000) that is used by CustAware, and that is also used by the common platform in CONTAIN.

Table 1: Mapping from ENS to CRS fields.

ENS	CRS tag
Transport document number	<cbc:ID>
Declaration date and signature	<crs:SubmissionDate>
Person lodging the summary declaration	<crs:ReportingParty>
Carrier	<cac:CarrierParty>
Notify party	<cac:NotifyParty>
	<crs:Consignment>
Commercial reference number	<cbc:ID>
Number of items	<cbc:TotalGoodsItemQuantity>
Gross mass	<cbc:GrossWeightMeasure>
Place of loading	<cac:PlannedPickupTransportEvent>
Place of unloading code	<cac:PlannedDeliveryTransportEvent>
Consignee	<cac:ConsigneeParty>
Country(ies) of routing codes	<cac:TransitCountry>
First place of arrival code	<cac:FirstArrivalPortLocation>
Transport charges method of payment code	<cac:PaymentTerms>
	<cac:TransportHandlingUnit>
Container number	<cbc:ID>
	<cac:TransportEquipment>
Seals information and seals identity	<cac:TransportEquipmentSeal>
	<cac:GoodsItem>
Goods description	<cbc:Description>
Type of packages (code), number of packages, shipping marks	<cac:Item>
Goods item number	<cbc:ID>
UN Dangerous Goods code	<cac:HazardousItem>
	<crs:MainCarriageShipmentStage>
Mode of transport code	<cbc:TransportModeCode>
Date and time of arrival at first place of arrival in the EU	<crs:EstimatedArrivalTransportEvent>
	<crs:TransportMeans>
Identity of active means of transport crossing the border	<crs:MaritimeTransport>
Nationality of active means of transport crossing the border	<cbc:RegistrationNationalityID>
Conveyance reference number	<cac:DocumentReference>

The CRS tags above are defined with the following XML namespaces:
cac=”urn:oasis:names:specification:ubl:schema:xsd:CommonAggregateComponents-2”
cbc=”urn:oasis:names:specification:ubl:schema:xsd:CommonBasicComponents-2”
crs=”urn:eu:specification:efreight:schema:xsd:CommonReportingSchemaExtensions-1.0”	

5. SIMULATING CONTAINER SECURITY DATA

Since our focus is on information to customs, we simulate ENS data by synthesizing a large ”bulk” stream of realistic messages, which are continuously received by customs in a real situation. In practice, this stream can be voluminous; perhaps several thousands of messages per second to a single national customs centre. This enormous inflow must to a major extent be analysed on-the-fly, which requires the analysis to be at least partly automatized. Another task in the CONTAIN project is to develop tools for customs to implement their already existing knowledge into software on how to find suspicious indicators and patterns in the ENS data they receive, such as that recently implemented in CustAware. Verifying and validating such a tool will be very important, from message parsing to the ability of the analysis of message content on different levels of complexity.

5.1. Introduction to Problem Setting

When simulating information on goods transported in containers, as well as on the containers, we must in some consistent way be able to put together information such as typical data on shipping line, port of first departure, transit ports, final destination port, consignor, consignee, goods type, points in time for lodging ENS and for call at first port of entry in the EU so it mimics some real goods status message, e.g. B2B status messages sent between business partners, or B2A messages as ENS to customs.

One natural way to generate ENS data is to model the entire system as a many-to-many interlinked logistics process with actors as agents. Goods are requested by consignees in destination countries, in which ports act like sinks for shipped goods. Sources are departing ports in countries where consignors export the goods. In between, we have the entire logistic process. Shipping lines follow specific routes and containers are filled with requested goods at the exit ports. This combined trade and logistics simulation would then produce ENS messages when a ship leaves its first port of departure with all containers and goods within them. At transit ports, one could simulate container off-loading and loading onto another ship as well as the movement of goods between offloaded containers. Since this method generates the data according to the definition of the process, it ensures that the variates are correctly correlated.

Although it is theoretically possible to develop such a simulation system, the modelling required for this would be too time-consuming and too difficult to validate for our needs: simulation of ENS messages. A more pragmatic approach is to use real data to generate simulated input data. A naive method would be to create a table where columns are different ENS data types containing valid ENS data values extracted from real data (e.g. consignor, consignee, shipping line, container ID, goods type) and randomly sample from table columns and combine the results to form a message. However, this process ignores the correlations among data and could produce combinations that are unrealistic. One needs some kind of restrictions or conditions that guides the process, for instance who tends to ship what goods to whom. Incorporation of this knowledge is necessary to avoid generation of unrealistic combinations of, e.g., consignors and goods such as fruit dealers that ship petrol or electronics. The key requirement in the generation of multivariate samples is the need to ensure an appropriate correlation structure among the components of the multivariate vector (Cheng 1988).

The simulated ENS bulk stream should have a profile very much like real messages of this type. This should of course include errors that might appear due to simple carelessness and misunderstandings of how to correctly fill in the information in an ENS message. This can be accomplished by using real messages as a ground for the simulation, which will be described in Section 6. This simulated data will then mimic the stream of standard messages that constitutes the “normal picture” of incoming goods and containers. The process is sketched in Figure 2.

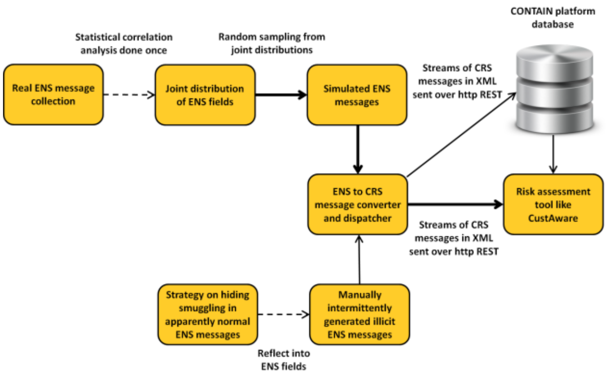
CONTAIN platform
database
Statistical correlation
analysis done once
Random sampling from
joint distributions
Real ENS message collection
Joint distribution of ENS fields
Simulated ENS messages
Streams of CRS
messages in XML
sent over http REST
ENS to CRS message converter and dispatcher
Risk assessment tool like CustAware
Streams of CRS
messages in XML
sent over http REST
Strategy on hiding smuggling in apparently normal ENS messages
Manually intermittently generated illicit ENS messages
Reflect into
ENS fields

Figure 2: The entire ENS generation to analysis chain in one picture. Full line thickness: Message flow volume. Dashed lines: Strategies for message simulation.

6. REAL DATA AS BASE FOR SIMULATIONS

Several types of real data exist that could be used as “templates” for generating a ground for simulated goods data messages. This is typically B2B transactional data generated along the goods logistic chain or B2A data, such as ENS messages sent to customs. What we ideally need for the simulation of such messages is the statistical joint distribution between all different data contents in a large set of real messages. This means that the number of messages in the sample data must be large enough, and the frequencies of unique occurrences of its entries, must not be too small in order to get high enough statistical confidence for different combinations of entries in the messages. In practice, it is not so easy to find this data. Annually compiled data from national statistical institutions on trade patterns tend to be rather heavily aggregated and therefore not suitable for generating statistical distributions on a fine-grained level. Non-aggregated raw-data is often more or less classified since it might reveal business-sensitive information. Therefore, it is of great value to acquire raw-data with a fine level of granularity. We have achieved this by obtaining some real data from a customs agency. In total, we have received around 130,000 ENS messages that were anonymized in order to not reveal classified contents and replaced potentially sensitive fields in the messages like names and addresses by hash codes or unique aliases.

6.1. Generating Univariate Random Variables from Empirical Data

There are several advantages associated with generating random variates from empirical distributions, perhaps chief among them the fact that an empirical distribution by definition is valid and there is no need for validation. However, there also may be problems which should be considered when sampling from empirical distributions. One problem is that if the data is not sufficiently large, it may not be representative of the complete probability distribution. Outliers and rare data may also cause problems. The procedure for variate generation from a discrete empirical distribution is the following:

1. Arrange the data X in some arbitrarily order $X=\{(x_1,m_1),\dots,(x_n,m_n)\}$, where m_i is the number of occurrence of item x_i.
2. Determine $\{(x_1,F(x_1)),\dots,(x_n,F(x_n))\}$, where $F(x_i)=\sum_{j=1}^{i}\frac{m_j}{m}$ and $m=\sum_{i=1}^{n}m_i$.
3. Generate u from $U(0,1)$.
4. Find the interval $(F(x_i)\le u\le F(x_{i+1}))$.
5. Return x_i.

6.2. Generating a Multivariate Distribution from Empirical Data

It is generally more difficult to model dependent data using empirical data, since it requires a larger amount of data to capture the structure of the mutual dependence. Below we discuss four approaches to this problem.

6.2.1. Independent variables

If the components of a random vector $X=(x_1,x_2,\dots,x_n)$ are independent, that is $F(x_1,x_2,\dots,x_n)=F(x_1)F(x_2)\dots F(x_n)$, one can generate X_i from the marginal distributions $F_i(\cdot)$ independently and combine them to produce X . However, this method cannot be used to generate dependent variates. In general, the marginal distributions contain far less information about the structure of the data than required to generate dependent variates. It is worth to recall that even knowledge about the correlation matrix, which is a measure of linear coupling between variables, does not provide sufficient information about the underlying distribution of data (in the same sense that expected value and variance in general do not capture all information about a random variable). This can be illustrated, considering samples generated from three completely different bivariate distributions, having the same correlation matrix (see Figure 3).

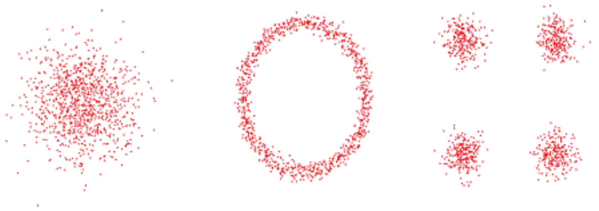

Figure 3: Three different bivariate distributions, having the same correlation matrix (Wikipedia 2014).

6.2.2. Using the complete joint distribution

The joint probability density function of a set of stochastic variables gives the probability that a particular outcome of one variable is combined with certain outcomes of other variables. Assuming that the number of variables is small and the joint distribution is known, that is, for each item $(x_1, x_2, \ldots, x_n)_j$ in the random vector $X = (X_1, X_2, \ldots, X_n)$, the number of occurrences of that item, m_j is known, one can use the procedure described in Section 6.1 to generate samples from $(X_1, X_2, \ldots, X_n)$. Since the generated variates have the same frequency as the empirical data, the dependence between variables is preserved. One obvious drawback of this method is the exponential growth of the size of the probability table by increasing the number of variables. If each discrete variable X_i can take on d_i different values, then the size of the probability table is $\prod_{i=1}^{n} d_i$. However, if the variables are strongly dependent, that is if the values of some of the variables to some extent determines the values of the others, then the probability of many of entries in the table will be 0, which results in a sparser probability table.

6.2.3. Conditional sampling

Another equivalent approach is conditional sampling, which uses the Bayes chain rule. This approach exploits causal dependence between the variables to express a joint distribution as a product of conditional probabilities, for example see (Russell and Norvig 2010).

While the full joint distribution gives random samples of all variables concurrently, conditional sampling provides a more step-wise process based on a causality reasoning, where conditional probabilities are built from assessments of the dependence of variables. This assessment can sometimes be derived from the available data. The conditional sampling procedure is straight-forward; the joint distribution function $F(x_1, x_2, \ldots x_n)$ is used to determine all marginal distributions $F_i(\cdot)$ of X_i for $i = 1,2, \ldots, n$ and conditional distributions $F_k(\cdot \mid X_1, X_2, \ldots, X_{k-1})$ of X_k given $X_1, X_2, \ldots, X_{k-1}$ for $k = 2,3, \ldots n$. By a tedious but straightforward manipulation of definitions of marginal and conditional distributions, one can show that the following procedure generates multivariate samples with given joint distribution $F(x_1, x_2, \ldots, x_n)$:

- Generate X_1 from the marginal distribution $F_1(\cdot)$
- Generate X_2 from $F_2(\cdot \mid X_1)$
- Generate X_3 from $F_3(\cdot \mid X_1, X_2)$
- $\ldots$
- Generate X_n from $F_n(\cdot \mid X_1, X_2, \ldots, X_{n-1})$
- Return $X = (X_1, X_2, \ldots, X_n)$

If it is possible to partition X into two sets of independent variables Y and Z such that

$$F_X(X) = F_Y(y) F_Z(z), \quad (1)$$

for all $x \in X$, $y \in Y$ and $z \in Z$, $x = y \cup z$, one can use the above procedure to independently generate these two set of variates and combine them to produce the vector X.

Moreover, if there exist Y, Z and W, such that

$$F_{Y,Z}(y, z \mid W) = F_Y(y \mid W) F_Z(z \mid W), \quad (2)$$

that is Y and Z are conditionally independent given W, the above algorithm can be used to generate W. Then given W, both Y and Z are generated from $F_Y(y \mid W)$ and $F_Z(z \mid W)$, respectively and vector X is produced as the union of Y, Z and W.

Since one can transform between using joint probability distributions on one hand and conditional sampling and using the Bayes chain rule on the other hand, it might be a matter of taste how it is implemented in practice; joint, conditional, or a mixture of those like so-called Bayesian Networks (Russell and Norvig 2010).

6.2.4. Copulas

Consider a continuous random vector $(X_1, X_2, \ldots, X_n)$, where marginal distributions are denoted by $F_i(x)$.

Then,

$$(U_1, U_2, \ldots, U_n) = (F_1(X_1), F_2(X_2), \ldots, F_n(X_n))$$

is a random vector with uniform margins. The joint cumulative distribution function (CDF) of $(U_1, U_2, \ldots, U_n)$, that is

$$C(u_1, u_2, \ldots, u_n) = P[U_1 \leq u_1, U_2 \leq u_2, \ldots, U_n \leq u_n]$$

is said to be the copula for $(X_1, X_2, \ldots, X_n)$ (Embrechts, McNeil, and Straumann 1999).

A nice property of the copula is that it separates the dependence structure between the components of the random vector and the marginal distribution functions. In order to generate a random vector $(X_1, X_2, \ldots, X_n)$, one can generate a sample $(U_1, U_2, \ldots, U_n)$ from the copula distribution and then construct $(X_1, X_2, \ldots, X_n) = (F_1^{-1}(U_1), F_2^{-1}(U_2), \ldots, F_n^{-1}(n))$, where $F_i^{-1}(U_i)$ is the inverse of the cumulative distribution function for margins.

However, the problem to generate variates is not completely resolved, since the copula $C(u_1, u_2, \ldots, u_n)$ depends on the joint distribution, which is usually unknown. In that case, the copula is approximated by some known distribution. One used approximation is the Gaussian copulas with the correlation matrix Σ:

$$C_{\Sigma}^{Gauss}(u) = \Phi_{\Sigma}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_n)),$$

where Φ_{Σ} and Φ^{-1} are the CDF of a multi-Gaussian distribution with mean zero and covariance Σ and the inverse CDF of a standard Gaussian, respectively. The use of Gaussian copula can be justified by the fact that the distribution is the sum of a large number of i.i.d. random variables with finite variance so the central limit theorem applies. The correlation matrix Σ can be estimated from statistical data.

Another approach is to learn the correct copula from sample data. This could be done using Genetic Programming or a similar technique that gradually refines a function based on comparison with real data.

6.3. Generating ENS Records Using Empirical Data

We have developed an application in Java that generates sample ENS data that is converted on-the-fly to CRS messages as was described in Section 4. Here, the main concern has been that the generated data is convincing and does not contain impossible or improbable items, as a consequence of systematic errors or shortcomings in the input data modelling process. A clear example of an impossible data item is existence of two different ENS records that indicate that a specific container is simultaneously in two different vessels. An example of improbable data item is a record indicating that Sweden exports bananas to Ecuador. Without a careful examination of the input data and a clear strategy for generating variates it is difficult to avoid these kinds of deficiencies in the synthetic data. However, it should be clarified that, it is accepted or even desired that the generated data contains the same types of abnormality as those found in the empirical data, for instance spelling errors or incomplete data.

Each ENS record consists of 141 fields (columns in Excel file), most of which are discrete-valued (e.g. name, address, country of consignee and consignor, type and code of goods, vessel and container ID). However, the records also contain some continuous-valued fields such as the weight of goods and arrival time. This means, theoretically that we require generating a highly-correlated random vector with 141 components. In order to reduce the number of components, several data pre-processing measures (i.e. cleaning, integration and reduction) have been taken:

- all fields that are empty or very rarely used are removed from the table,
- fields containing irrelevant information for the input modelling, such as internal ID values for Swedish Customs are ignored,
- redundant fields like country name where the information is already given by country code are removed,
- some fields containing relevant information that seem not to affect the result of the risk calculation algorithms are removed,
- by making assumptions like the uniqueness of the name of companies in each country, some fields such as the address of companies are determined by other fields (name and country) and are removed.

These steps reduce the number of fields to 55, which is considerably less than the original 141 fields, however, still too large to calculate the frequency of unique records. Therefore, a core of 26 highly correlated fields is distinguished and their frequency is used as the joint distribution to generate the corresponding features as described in Section 6.2.2.

Although the remaining 29 fields are in different ways correlated to these 26 fields, they are either dependent on one or a small number of fields in this subset, or can be treated as independent without compromising the trustworthiness of the generated data. For instance, it is less critical if the amount of import of some goods by different consignors does not follow the real distribution. Hence, it is possible to generate the remaining fields conditioned on one or a small number of fields. As described in Section 6.2.3, when it is possible to partition random variables into independent subsets, or conditionally independent subsets, equation (1) respectively (2) can be used to randomly generate these subsets of variables and combining them afterwards.

Most of the goods that enter European ports are transported in containers. The ID number of the carrying container (BIC), which is an ISO standard, is one of the entries in ENS messages, and an important one for associating a container with which goods is transported in it. Container IDs are sampled uniformly from the list of container IDs existing in the data set (nearly 18,000 unique IDs) and are assigned to the vessels sorted by the arrival date. In order to avoid assigning a container simultaneously to different vessels, they are marked as occupied once they are drawn and are not added back to the pool of containers until a fixed time period has elapsed.

6.4. Injecting "Dangerous Containers"

Above, we described a process to simulate an in some sense "normal" inflow of ENS messages. This requires that real messages, which are the ground of the synthetic data, contain few enough anomalous messages that are even in reality indicators of illicit activities. However, how customs exactly finds those messages is often secret, so we have not been able to remove them. But pretending that all messages in the common ground are fully normal, we also want to intermittently "inject" a much smaller amount of information that could indicate malicious activities; mainly smuggling of illegal goods. In the end of the message flow chain, it is then up to the analysing functionality (customs officers, Artificial Intelligence tools, or a combination thereof) to find the needles (messages indicating malicious activities) in the haystack (normal messages) (Brynielsson, Westman, and Svenson 2015). Thus, it is important to make sure that injection of "dangerous containers" is separate from risk-assessment or decision support system that has the task to find them.

The details of the CustAware tool is out of the scope of this paper. However, initial tests conducted in non-operational environments for demonstration purposes has been promising. The CustAware is be able to use different models to produce risk indicators (risk values) and support the user to decide which containers or consignments to inspect. The models can be either rule based (blacklists, whitelist, etc.) or learn from data. Both unsupervised learning (anomaly detection) and supervised learning (outcomes from previous inspections) can be used to create the risk models.

7. DISCUSSION

We have described a methodology for simulating ENS-like messages for customs. It is preliminary used within the CONTAIN project to test systems developed to handle ENS data in the CRS format. One important aspect is to evaluate the working process of updating and refining the risk profiles in the CustAware tool for efficient identification of potential illegal activities such as smuggling. It can also be used to test customs' systems for receiving such messages, as well as the ability of those systems to detect indications on ongoing attempts for smuggling or other illicit behaviour.

ACKNOWLEDGMENTS

This research has been funded by the R&D Programme of the Swedish Armed Forces and by the European Commission under Grant Agreement no. 261679 (CONTAIN). We are grateful to customs officials from several countries for enlightening discussions and to Tom Cane of BMT for discussions on the Common Framework.

REFERENCES

ARENA. Architecture for the recognition of threats to mobile assets using networks of multiple affordable sensors. Available from: https://www.informationsystems.foi.se/~arena-fp7 [accessed 13 April 2015].

Asp B., Carling C., Hunstad A., Johansson B., Johansson P., Nilsson J., Waern A., 2008. COAT-Användarguide (Communications Assessment - User Guide). Technical report. Swedish Defence Research Agency FOI. Linköping, Sweden.

Brynielsson J., Westman T., Svenson P., 2015. Decision support for customs' container risk management. Manuscript in preparation.

BIC. Bureau International des Containers et du Transport Intermodal (BIC). Available from: http://www.bic-code.org [accessed 12 April 2015].

CASSANDRA. Common assessment and analysis of risk in global supply chains. Available from: http://www.cassandra-project.eu [accessed 13 April 2015].

Cheng, R. C. H., 1988. Random variate generation. In J. Banks ed. Handbook of simulation: principles, methodology, advances, applications and practice. A Wiley-Interscience publication, 139–172. New York: John Wiley & Sons Inc.

CONTAIN. Container security advanced information networking. Available from: http://www.containproject.com [accessed 13 April 2015].

CORE. Consistently optimised resilient ecosystem implementation of the secure transit of goods through global supply chain system. Available from: http://www.coreproject.eu [accessed 13 April 2015].

e-Freight, 2013a. Deliverable no. D1.3b: e-Freight Framework Information Models. Available from: http://www.efreightproject.eu [accessed 12 April 2015].

e-Freight, 2013b. Deliverable no. D6.5: e-Freight Policy Recommendations. Available from: http://www.efreightproject.eu [accessed 12 April 2015].

Embrechts P., McNeil A., and Straumann D., 1999. Correlation and dependence in risk management: properties and pitfalls.

European Customs Information Portal, 2013. Entry Summary Declarations (ENS): Consolidated FAQs. Available from: http://ec.europa.eu/ecip/documents/procedures/import_faq_en.pdf [accessed 12 April 2015].

EUROSKY. Developing a Single European Secure Air-cargo Space. Available from: http://www.euroskyproject.eu [accessed 13 April 2015].

Fielding R. T., Taylor R. N., 2000. Principled Design of the modern web architecture. Proceedings of the 22Nd International Conference on Software Engineering, pp 407-416. June 5-7, Limerick, Ireland.

FP7, 2014. EU's Seventh Framework Programme for Research. Available from: http://ec.europa.eu/research/fp7 [accessed 13 April 2015].

Longo F., 2010. Design and integration of the containers inspection activities in the container terminal operations. International Journal of Production Economics, vol. 125(2), pp. 272-283.

OASIS. OASIS Universal Business Language (UBL). Available from: https://www.oasis-open.org/committees/ubl [accessed 12 April 2015].

PROTECTRAIL. The railway-industry partnership for integrated security of rail transport. Available from: http://www.protectrail.eu [accessed 13 April 2015].

Russell S. J., Norvig P., 2010. Artificial intelligence – a modern approach (3. Internat. ed.). Pearson Education.

SAFEPOST. Reuse and development of Security Knowledge assets for International Postal supply chains. Available from: http://www.safepostproject.eu [accessed 13 April 2015].

SUPPORT. Security upgrade for ports. Available from: http://www.supportproject.info [accessed 13 April 2015].

TASS. Total airport security system. Available from: http://www.tass-project.eu [accessed 13 April 2015].

Taxation and Customs Union - European Commission, 2006. Data requirements for entry and exit summary declarations and for simplified procedures. Available from: http://ec.europa.eu/taxation_customs/index_en.htm [accessed 12 April 2015].

Tremori, A., Massei, M., Madeo, F. and Reverberi, A. (2013). Interoperable simulation for asymmetric threats in maritime scenarios: a case based on virtual simulation and intelligent agents. International Journal of Simulation and Process Modelling, Volume 8, Issue 2-3, pp. 160-167.

Wikipedia 2014. Correlation and Dependence. Available from: http://en.wikipedia.org/wiki/Correlation_and_dependence [accessed 12 April 2015].

AUTHORS BIOGRAPHY

Farzad Kamrani is a Scientist at Swedish Defence Research Agency (FOI) in Stockholm, Sweden. He holds a M.S. and a PhD in Computer Science from the University of Gothenburg and KTH Royal Institute of Technology, respectively. His research interest includes modelling and simulation. His email address is kamrani@kth.se.

Pontus Hörling is Senior Scientist at FOI in Stockholm. He holds a M.S. and a PhD in Physics from, respectively, the University of Uppsala and KTH Royal Institute of Technology. His research mainly deals with data- and information fusion and related simulation. His email address is hoerling@foi.se.

Thomas Jansson is Senior Scientist Assistant at FOI in Stockholm. He holds a M.S. in Computer Science from the Technische Universität München, Germany. His email address is thomas.jansson@foi.se.

Pontus Svenson is Deputy Research Director at FOI in Stockholm. He has a PhD in Theoretical physics from Chalmers University of Technology. His research interests include decision support systems, information fusion and machine learning. His email address is pontus.svenson@foi.se.