Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

MULTIVARIATE DATA GENERATION FOR CUSTOMS RISK EVALUATION TOOL

Farzad Kamrani, Pontus Hörling,
Thomas Jansson, Pontus Svenson

Department of Decision Support Systems
Swedish Defence Research Agency (FOI)
SE-164 90 Stockholm, SWEDEN

ABSTRACT

Today, vast volumes of goods are transported all over the world in containers. Customs authorities are charged with detecting smuggling and can find indications of this by screening documentation on containers that is by law provided by shippers and carriers. In the Contain project, Decision Support Systems for customs to do this risk profiling are developed. In order to test these systems, we have developed a tool (ENS-simulator) to provide simulated input to the profiling tool. In this paper, we present the simulation tool and describe the method used for generating a high rate of messages in a realistic way, to represent a typical message inflow at a large customs risk assessment center.

1 INTRODUCTION

The CustAware system (Brynielsson, Westman, and Svenson 2014) is part of the EU Contain project, which aims to provide customs inspectors with a decision support system (DSS) that enables them to determine which containers to inspect manually or by using sensors (*scanning*) based on all available information. Choosing *which* containers to inspect (*targeting*) is one of the most important tasks for customs, especially in large ports where only a few percent of all containers are inspected. In order to test the CustAware system, 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, since the data is sensitive.

Today, the B2A information exchange in the container handling process to customs 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 must be filed at the first port of entry into the EU by the carrier (or its agent).

Within several earlier freight-oriented EU projects, the so-called Common Framework (CF) information model was developed (e-Freight 2013). CF extends the OASIS Universal Business Language (UBL) model intended for standardized B2B information transactions in XML (OASIS 2013). The extensions provide more interoperability between logistics stakeholders and enable the expression of B2A data. Within CF, the Common Reporting Schema (CRS) is implemented as an XML Schema, containing all necessary fields for mapping ENS data requirements.

2 SIMULATING CONTAINER SECURITY DATA

The *ENS-simulator* produces a Comma Separated Values (CSV) list of simulated text entries, such as consignor name, address, goods commodity code, 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 a self-describing format, such as JSON or XML reflecting its semantic (tagged) content. In the Contain experiments, the message is then submitted to an instance of CustAware that parses the messages and analyzes their content on the fly for indicators of illegal goods or activities.

To ensure reliable tests, it is important to preserve correlations in the data (e.g. port of first departure, transit ports, final destination port, consignor, consignee, goods type) and produce combinations that are valid. By invalid combinations we mean e.g., consignors and goods such as fruit dealers that ship petrol or electronics. This can be accomplished by using real messages as a ground for the simulation.

3 GENERATING ENS RECORDS USING EMPIRICAL DATA

There are several advantages in using empirical distributions to generate random variates, perhaps chief among them the fact that an empirical distribution by definition is valid and there is no need for validation. Despite difficulties to obtain data, we have acquired real data with a fine level of granularity, from a customs agency, which contains more than $100,000$ ENS messages. The data has been anonymized in order to not reveal classified contents.

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 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. However, if the variables are strongly dependent, the probability of many of entries in the table will be 0, which results in a more sparse probability table. This method ensures 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 modeling process.

Each ENS record consists of 141 fields, most of which are discrete-valued (e.g. name, address, vessel). However, the records also contain some continuous-valued fields (e.g. weight). After data preprocessing (i.e. cleaning, integration and reduction), the number of fields is reduced to 55, from which a subset of 26 highly correlated fields is distinguished. Frequencies of occurrence of records in this subset yields the joint distribution of the data, which is used to generate the corresponding fields. The remaining 29 fields are generated either as independent variables or dependent on a small number of fields.

Assuming that all generated messages are normal, we also need to intermittently "inject" a much smaller amount of information that could indicate malicious activities; mainly smuggling of illegal goods. This message flow, is then used to evaluate whether the CustAware is an adequate decision support system and can help the customs officers to find the needles (messages indicating malicious activities) in the haystack (normal messages).

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).

REFERENCES

Joel Brynielsson and Tommy Westman and Pontus Svenson 2014. "Decision Support for Customs' Container Risk Management". Manuscript in preparation.

e-Freight 2013. "Deliverable 1.3b: e-Freight Framework Information Models". Project Report. http://www.efreightproject.eu/uploadfiles/e-FreightD1.3be-FreightFramework.pdf.

European Customs Information Portal 2013. "Entry Summary Declarations (ENS): Consolidated FAQs". Accessed Apr. 9, 2014. http://ec.europa.eu/ecip/documents/procedures/import_faq_en.pdf.

OASIS 2013. "OASIS Universal Business Language (UBL)". Accessed Apr. 9, 2014. https://www.oasis-open.org/committees/ubl.

Taxation and Customs Union - European Commission 2006. "Data requirements for entry and exit summary declarations and for simplified procedures". Accessed Apr. 9, 2014. http://ec.europa.eu/taxation_customs/index_en.htm.