

Detecting Shipping Fraud in Global Supply Chains using Probabilistic Trajectory Classification

Ron Triepels¹ and Hennie Daniels^{1,2}

¹*CentER, Tilburg University, Warandelaan 2, Tilburg, The Netherlands*

²*Rotterdam School of Management, Erasmus University, Burg.Oudlaan 50, Rotterdam, The Netherlands*

1 INTRODUCTION

Advances in transportation technologies and the liberalization of trade restrictions have transformed trade into a global phenomenon during the last two decades (Taylor, 2002). While the economic impact of global trade has been widely studied, it is only recently that researchers in the field of supply chain management have begun to investigate the negative impact of globalization on supply chain visibility. The inability to track goods in transit involves risks, particularly in terms of the customs brokerage practices of freight forwarders.

Restrictions in international trade are fading away due to preferred trade agreements between countries and trade organizations. Firms within preferred trading areas can exploit differences in wages and knowledge across countries. This allows them to acquire difficult-to-obtain goods at competitive prices. However, the increasing diversity of firms, products, and shipping options makes global supply chains considerably complex. Trade is initiated between buyers and sellers whose identity and trade behavior is relatively unknown to the freight forwarder. Also, global supply chains affect a lot of intermediate shipping companies that operate in highly competitive markets. Shippers are afraid to reveal too much information and be put out of business by competitors.

At the same time, innovation and technology have significantly reduced transportation costs (Hummels, 2007). One of the most noteworthy innovations in logistics is containerization. Containerization refers to a system of containers designed to move goods efficiently across all modes of transport. Goods at the beginning of the supply chain are clustered in containers based on their ownership or destination. The standardized design allows shippers to quickly exchange containers at terminals without having to open them. Although this greatly improves efficiency, it also negatively affects supply chain visibility because what is being shipped in a given container is no longer visible from the outside.

Risks for the freight forwarder mainly arise when goods marked for import must be declared to the customs authorities of the importing country. Several intermediate shippers may be involved in a single transportation who only share a bill of lading to attest to their provided services. The problem here is that there is no guarantee that the goods listed on these documents are actually inside the containers. Opening every container marked for clearance is problematic because this would imply major operational costs. Instead, freight forwarders must solely rely on externally created shipping documentation and declare goods they usually do not even see (Hesketh, 2010).

The rise of information technology provides opportunities to improve the customs brokerage practices of freight forwarders (Gordhan, 2007). On one hand, information technology allows more detailed supply chain data to be recorded. Radio frequency identification (RFID) has received significant attention in the supply chain sector. The small size of the RFID tags and their low production costs makes them useful in tracking and tracing international cargo flows (Chang et al., 2011) and reducing delays at customs clearance locations (Hsu et al., 2009). On the other hand, information technology provides the tools to share data among supply chain participants. Several technologies have been proposed to connect shippers and freight forwarders, such as electronic data interchange (EDI) (Murphy and Daley, 1999). This project investigates how freight forwarders can use data mining to extract knowledge from their extensive supply chain repositories and use it to fight shipping fraud.

2 STATE OF THE ART

Shipping fraud is committed in many different ways and on different scales, ranging from local cargo theft to international smuggling. Either way, evidence of a fraud scheme must be covered in the corresponding

shipping documentation. This is known as document fraud or misrepresentation. Misrepresentation is the act of manipulating facts in contracts or agreements with the intent to benefit by commercial gain (Hill and Hill, 2009). The tendency of fraudsters to manipulate documentation elicits an interesting and challenging question. To what extent does shipping documentation correspond to reality? Researchers have studied how different data mining techniques can help freight forwarders answer this question.

2.1 Classification

Classification is the problem of finding a good model that assigns an observation to the right class based on a set of observations with known classes.

Classification can be used to determine whether goods on a declaration are correctly classified (Filho and Wainer, 2007). The researchers constructed a hierarchical Bayesian model consisting of four features: type of goods, consignee, country of origin, and destination country. Associations between the features are learned from a large sample of correctly and incorrectly classified goods. It turns out that the proposed model achieves high prediction accuracy despite the high cardinality of the features.

A related study takes a broader approach and constructed a classifier based on association rules to predict the risk level of a declaration (Yaqin and Yuming, 2010). In this study, the classification model is built by performing association rule mining on a large set of features. In addition to the model by Filho and Wainer, this study also considers features such as weights and prices. The risk level of a declaration is determined by the number and nature of association rules that match. Freight forwarders can use these risk levels to pay more attention to declarations that have a high probability of being fraudulent.

2.2 Regression

Regression is the problem of finding a good function that captures the underlying relationship between a set of variables.

Logistic regression has been applied to predict the extent to which a declaration involves smuggling (Hua et al., 2006). Fitting a good regression function is difficult due to the inhomogeneous nature of customs data. The study uses a two-step cluster method to divide customs data into seven approximately homogeneous clusters. Similarity between declarations in these clusters is based on features such as the prices and weights of goods. For each cluster, a logistic regression function is fitted with a set of variables that

are correlated with smuggling. Finally, a decision rule is defined to combine the predictions of the individual regression functions. Performance tests show that the decision rule significantly improves the efficiency of customs inspections.

2.3 Anomaly Detection

Anomaly detection is the problem of identifying combinations of features that significantly deviate from a statistical norm.

Brazil has built an anomaly detection system for their customs systems in conjunction with universities in the country (Digiampietri et al., 2008). The system uses Markov chains and n-grams to search for product descriptions on a repository of historic traded goods. If a satisfactory match is found, historic data about the product and its properties are retrieved and highlighted in a set of diagrams. These diagrams show the statistical distribution of combinations of product properties. In this way, customs agents can visually inspect how much goods marked for clearance deviate from what is expected.

Anomaly detection suffers from two major drawbacks. First, it can be hard to pinpoint anomalies from diagrams alone, as the decision boundary is defined by a subjective interpretation. There is a general tendency to overrate anomalies, which results in a overload of suspicious cases. Second, anomalies cannot be compared based on their severity. This makes it difficult to make a deliberate choice between the cost to undertake a detailed investigation and the savings that can potentially be achieved.

Ranking can be used to prioritize anti-fraud investigation (Kopustinskas and Arsenis, 2012). The study proposes a way to calculate a numerical ranking for price outliers in trade data. A suitable method is to multiply the expected loss in duties by the probability that a declaration is fraudulent. Expected loss is estimated by multiplying the traded quantity and difference in unit price, whereas the fraud probability is estimated by the p-value of statistical tests. The ranking measure allows freight forwarders to identify declarations for further investigation, which can achieve high savings.

3 RESEARCH PROBLEM

Shipping fraud has mainly been investigated by analyzing product specifications and aggregated geographical data. Research shows that particular types of shipping fraud can be detected when looking at the statistical properties of declarations. However,

these approaches mainly neglect how goods actually find their way to the consignee. The trajectory of a shipping container presumably reflects hidden patterns that are characteristic of certain types of fraud. The current literature lacks a well-grounded explanation of how fraud influences shipping trajectories and of how we can detect it from supply chain data.

International trade is moving towards vertical specialization in which each country produces particular goods for the stages of a production sequence (Hummels et al., 2001). At the same time, firms seek ways to optimize the logistics between countries based on economic considerations, such as price, flexibility, and service level (Tongzon, 2009; Chang et al., 2011). Vertical specialization and shipping optimization is expected to create distinct patterns that reflect how countries trade and under what conditions goods are being transported. Deviations from what are expected to be normal shipping trajectories can therefore point to possible cases of fraud.

3.1 Trajectory Classification

Trajectory classification is the problem of predicting the class of moving objects based on historic trajectories (Lee et al., 2008). This problem has recently received increased attention in the data mining field due to its potential applications. Trajectory classification is applied in video surveillance to detect anomalous behavior (Owens and Hunter, 2000) and air pollution measurement to study the pollution of air masses approaching a given site (Riccio et al., 2007). Despite the increased attention, the application of trajectory classification in global supply chains has for the most part been unexplored.

The classification of moving objects can be probabilistic or non-probabilistic. A probabilistic classifier uses a generative model to determine the most likely class by calculating a probability distribution over all classes. In contrast, a non-probabilistic classifier predicts a class without specifying the uncertainty by which the prediction is made. The use of a probabilistic classifier has the advantage that it allows the incorporation of decision theory and the making of automated decisions based on utility functions. Decision theory is vital for fraud detection systems to prioritize anti-fraud investigations. This leads to the following research question.

Research Problem. *How can freight forwarders apply probabilistic trajectory classification to predict shipping fraud in global supply chains?*

We identified four sub-problems that need to be studied to apply trajectory classification to supply

chain data. This includes the generation of discriminating patterns, the support of multidimensional data, the reliability of model parameters, and the choice of the probabilistic classifier.

3.2 Discriminating Patterns

An important process in the construction of an effective classifier is the generation of discriminating patterns. Discriminating patterns are combinations of features that occur with disproportionate frequency in some classes (Fang et al., 2011). Including them in a classifier has proven to significantly increase predictive power. However the features of a shipping trajectory are different from traditional features. First, locations of a trajectory occur in a logical order. Moving goods to a given location can be discriminatory, but this premise may not hold for trajectories in the reverse direction. Second, trajectories can become quite large in size and can involve many different locations. This is expected to raise some performance issues. Research is needed to effectively find discriminating patterns in global supply chains.

Sub-problem 1. *How can discriminating patterns in shipping trajectories be effectively generated?*

3.3 Multidimensional Data

Trajectory classification models in the literature are of one dimension, that is, they only model the direction in which an object is moving. While this is sufficient in most applications, the movement of goods in a supply chain is influenced by many more dimensions or shipping concepts. Take the International Commercial terms or Incoterms as an example. Incoterms were introduced to clearly communicate how tasks, costs, and risks are allocated between the supply chain participants. Because each supply chain participant has a different stake in the shipment, different terms likely imply different trajectories. An important question is therefore how such additional dimensions can be added to the classification of a trajectory.

Sub-problem 2. *How can trajectories be classified based on multidimensional trajectory data?*

3.4 Reliable Parameters

The parameters of a probabilistic classifier need to be estimated from data. This is a difficult task because samples of real life data often include rare events and are loosely controlled. Bayesian inference sufficiently estimates these rare events by incorporating prior knowledge and updating a distribution when evidence is gleaned from a sample. Research argues that

Bayesian inference supports the most optimal way of decision making (Braithwaite et al., 2007). However, the problem that follows is how much, and from what source, expert knowledge is needed to estimate the model parameters in a reliable way. This problem needs to be further explored to make classification effective in a practical environment.

Sub-problem 3. *How can Bayesian inference be applied to estimate the parameters of the trajectory classifier in a reliable way?*

3.5 Generalizability

The literature provides many techniques to perform probabilistic classification, including naive Bayes and tree-augmented Bayes (Friedman et al., 1997). The choice of a classifier depends greatly on the nature of the data from which the classification task is learned. Naive Bayes suffers from the strong independence assumption between features. This assumption can partly be avoided by using a classifier such as tree-augmented Bayes. However, performing model averaging to find the best tree-like model remains difficult (Cerquides and De Mántaras, 2003). It is important to investigate the strengths and weaknesses of different probabilistic classifiers and their generalizability to typical supply chain data.

Sub-problem 4. *Which type of probabilistic classifier generalizes well to supply chain data?*

4 OUTLINE OF OBJECTIVES

The primary goal of this research project is to build a framework for freight forwarders to construct effective classifiers for the purpose of predicting fraudulent shipping trajectories. Secondary goals are to:

1. Create a general description of the customs brokerage processes of freight forwarders, including tasks, actors, and information flows.
2. Identify the relationship between shipping trajectories and different types of fraud.
3. Define a general process that describes how trajectory classification can be set up, including:
 - (a) A strategy to effectively generate discriminating shipping trajectories.
 - (b) A method to support multidimensional trajectory features.
 - (c) A strategy to estimate the model parameters in a reliable way.
 - (d) An overview of different classifiers and their expected performance.

4. Define a set of design principles that freight forwarders can use to implement probabilistic trajectory classification in practice.

5 METHODOLOGY

During this research project, a designable artifact will be created that follows the design process proposed by the design science paradigm. Design science is defined as a research process that involves the creation and evaluation of information technology (IT) artifacts designed to solve a specific organizational problem (von Alan et al., 2004).

Our project is positioned between the environment of freight forwarding and the knowledge base of the data mining community. Freight forwarding practices are combined with established techniques in data mining to come up with a solution to the growing visibility problems in global supply chains. The designable artifact will be a data mining framework that aims to improve the decision making in customs brokerage by analyzing historic shipment trajectories. To safeguard validity and effectiveness, we design the framework according to a modified version of the design science research model (Peffer et al., 2006).

Research activities are divided into four design stages; see Figure 1. In the first stage, we explore the problem that occurs in the domain of interest and define a corresponding objective for the research project. Both topics have already been discussed in previous sections. Therefore, the remainder of this section elaborates on the other stages of the research project.

5.1 Requirements

In the second stage, the requirements that need to be satisfied by the data mining framework are defined by performing a requirements analysis. The requirements analysis consists of a field study and a literature review.

A field study will be performed at a large international freight forwarder to explore its customs brokerage practices and corresponding information provisioning. Customs brokerage has become a complex process with the establishment of numerous international trading policies and shipping concepts. Therefore, this study is also conducted to make us familiar with this domain knowledge. Emphasis will be put on the way that shipping fraud is committed in practice and how fraud characterizes itself in the data of a shipment and its declaration. Data for the field study will

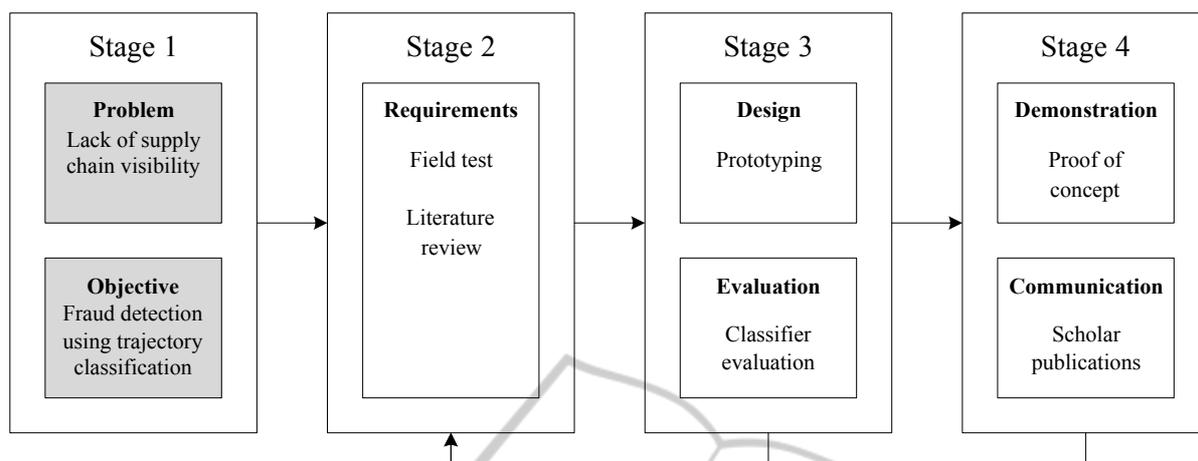


Figure 1: The four stages of the research project (stage one is discussed in this paper and is therefore highlighted). The template for this model is adopted from (Peffer et al., 2006).

mainly be collected through interviews with managers and customs agents.

Existing solutions proposed in the data mining literature will be explored by an extensive literature review. The literature review will be conducted to gain an overview of existing solutions proposed to fight shipping fraud. Papers dedicated to this topic are hard to find and are scattered in various disciplines, including data mining, customs, finance, and operations research. Therefore, a comprehensive bibliography search will be performed across various libraries and journals. The literature search will only select conference papers and journal papers. Each publication will be classified according to the six main tasks of data mining (Fayyad et al., 1996), along with the advantages and disadvantages of the chosen technique. Solutions are compared, and the review will elaborate on the extent to which these solutions are able to improve supply chain visibility in practice. Insights will be combined with the conclusions of the field study to formulate the design requirements.

5.2 Design

In the third stage, the design of the data mining framework takes place. Design activities include the construction of classifiers using prototyping and the evaluation of these classifiers by conducting performance evaluations.

Prototyping is the process of developing a set of trial versions of an artifact to clarify requirements and reveal critical design decisions (Gordon and Biegan, 1995). The technique has become increasingly popular in the field of software development because it allows the building of systems in a short span of time with feedback from end-users. Prototypes will

be built by taking samples from the databases of the freight forwarder. These samples will initially be relatively small and concentrated in specific trade lanes but will gradually increase in size and scope when the project progresses. Each sample will be stored in a Microsoft SQL server instance to explore the data and perform data cleaning and transformation. Data is then loaded into statistical package R to construct different classifiers.

5.3 Evaluation

Each classifier that is built as a prototype will be tested on its performance, that is, its ability to detect shipping fraud. A classifier performs well if it scores high on all of the following three quality indicators:

1. Accuracy - is the ability to correctly predict class labels on previously unseen data. Accuracy will be tested by separating data into a training set and a test set and measuring the number of correctly predicted classes in the test set.
2. Scalability - is the ability to deal with large data sets within an acceptable amount of time. Scalability will be tested by increasing the sample size and scope while measuring the effect on the accuracy and time complexity.
3. Robustness - is the ability to make correct predictions based on loosely controlled data. Robustness will be tested by decreasing the amount of prior information used during parameter estimation and measuring the effect on the accuracy.

The performance of all classifiers in a design iteration are compared to each other. If there exists a classifier that performs satisfactory, i.e. it scores at least high on accuracy, then a proof of concept will be

built. Otherwise, the design requirements are revised in close corporation with the freight forwarder.

5.4 Demonstration

In stage four the results of and insights about the design activities are reported to the freight forwarder and the data mining community.

The best performing classifier will be used to build a proof of concept. A proof of concept will be built to demonstrate the main features of the classifier to the freight forwarder and to verify that it has real-world application. The demonstration will show how the classifier can be used during the customs brokerage processes, how employees can operate the classifier, and what savings the freight forwarder can expect to achieve. If the freight forwarder accepts the proof of concept, then the data mining technique and the proof of concept will be reported to the data mining community via scholarly publication. Otherwise, the design requirements will be revised and a new proof of concept will be built.

5.5 Communication

Important findings in the design of an effective classifier will be reported back to the data mining community through scholar publication. The research project is likely to produce at least the following papers:

1. A paper on the motivation and possibilities to use supply chain trajectories to identify different types of shipping fraud.
2. A paper that describes the data mining framework to apply trajectory classification in practice.
3. A paper with a case study that describes a proof of concept applied at the freight forwarder.

The second paper may contain too much detail for a single publication. Therefore, the paper is likely to be divided into several smaller papers that address one or two sub-problems and one that provides a general overview of the framework.

6 STAGE OF THE RESEARCH

The first two stages of the research project are almost completed. An international freight forwarder, whose identity will remain unknown for privacy reasons, has been contracted for the project. Several interviews have been performed with customs managers and customs agents at different offices of the freight forwarder. Based on these interviews, the customs

brokerage processes are described in detail. Also, the literature review is completed from which the most important publications are included in this paper. Insights from the interviews and the literature review are used to define the first draft of the design requirements.

Regarding the third stage, samples are taken from the shipping system and customs system of the freight forwarder. Neither information system is currently integrated, so some effort was needed to correctly link shipments and declarations in a new database.

The data is used to construct a Bayesian network that models the locations of a cargo trajectory together with a set of important declaration features (Liu et al., 2014). Predictive reasoning is applied to investigate the change in probabilities of features when evidence of a trajectory is inserted into the network. Based on the evidence inserted, the network predicts the type of goods, the use of preferential documents, and the customs duties to be paid. The initial prototype demonstrates the ability to detect human errors and potential cases of fraud by investigating the trajectory of a shipment. However, we found the trajectory of a shipment cannot be modeled on a sufficient granularity in a Bayesian network. It is expected that more detailed locations of the trajectory need to be included in the classification model to achieve sufficient performance for real-life application.

Currently, we are investigating how cargo trajectories can be modeled in more detail using discriminating sequential patterns. We use the three-step mining approach proposed by (Cheng et al., 2007) to build a classification model. First, sequential mining is performed to find frequent combinations of locations in the cargo trajectories. Second, feature selection is performed to select the most discriminating patterns. Third, the data set is transformed and a classifier is built based on the selected features. The main advantage of this approach is that trajectories can be of any size and can contain repeating parts of sequences. Different classifiers and strategies to find discriminating trajectories are currently being investigated.

7 EXPECTED OUTCOMES

The expected outcome of this research project is a framework for freight forwarders to construct effective trajectory classifiers for the purpose of detecting fraud in international cargo flows. Freight forwarding practices are integrated with established techniques in the field of data mining. The resulting framework is likely to impact both freight forwarding practices and techniques for trajectory classification.

On one hand, freight forwarders struggle with the extensive amount of data they have to deal with during their operations. The framework guides freight forwarders through the process of setting up data mining initiatives and shows how such initiatives can improve customs brokerage practices. Implementations of the framework are likely to increase the effectiveness of customs compliance. On the other hand, the literature on trajectory classification lacks techniques to analyze global supply chains. Supply chain data has the characteristics of big data and is therefore difficult to analyze. The framework is expected to contribute a new application of trajectory classification to the data mining literature and show how it can effectively be used vis-à-vis on supply chain data.

REFERENCES

- Braithwaite, R. S., Roberts, M. S., and Justice, A. C. (2007). Incorporating quality of evidence into decision analytic modeling. *Annals of internal medicine*, 146(2):133–141.
- Cerquides, J. and De Mántaras, R. L. (2003). Tractable bayesian learning of tree augmented naïve bayes models. In *ICML*, pages 75–82.
- Chang, Y. S., Son, M. G., and Oh, C. H. (2011). Design and implementation of rfid based air-cargo monitoring system. *Advanced Engineering Informatics*, 25(1):41–52.
- Cheng, H., Yan, X., Han, J., and Hsu, C.-W. (2007). Discriminative frequent pattern analysis for effective classification. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 716–725. IEEE.
- Digiampietri, L. A., Roman, N. T., Meira, L. A., Ferreira, C. D., Kondo, A. A., Constantino, E. R., Rezende, R. C., Brandao, B. C., Ribeiro, H. S., Carolino, P. K., et al. (2008). Uses of artificial intelligence in the brazilian customs fraud detection system. In *Proceedings of the 2008 international conference on digital government research*, pages 181–187. Digital Government Society of North America.
- Fang, G., Wang, W., Oatley, B., Van Ness, B., Steinbach, M., and Kumar, V. (2011). Characterizing discriminative patterns. *arXiv preprint arXiv:1102.4104*.
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37.
- Filho, J. and Wainer, J. (2007). Using a hierarchical bayesian model to handle high cardinality attributes with relevant interactions in a classification problem. In *IJCAI*, pages 2504–2509.
- Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, 29(2-3):131–163.
- Gordhan, P. (2007). Customs in the 21st century. *World Customs Journal*, 1(1):49–54.
- Gordon, V. S. and Bieman, J. M. (1995). Rapid prototyping: lessons learned. *IEEE software*, 12(1):85–95.
- Hesketh, D. (2010). Weaknesses in the supply chain: who packed the box. *World Customs Journal*, 4(2):3–20.
- Hill and Hill (2009). *Nolo's Plain-English Law Dictionary*. Nolo.
- Hsu, C.-I., Shih, H.-H., and Wang, W.-C. (2009). Applying rfid to reduce delay in import cargo customs clearance process. *Computers & Industrial Engineering*, 57(2):506–519.
- Hua, Z., Li, S., and Tao, Z. (2006). A rule-based risk decision-making approach and its application in china's customs inspection decision. *Journal of the Operational Research Society*, 57(11):1313–1322.
- Hummels, D. (2007). Transportation costs and international trade in the second era of globalization. *The Journal of Economic Perspectives*, pages 131–154.
- Hummels, D., Ishii, J., and Yi, K.-M. (2001). The nature and growth of vertical specialization in world trade. *Journal of international Economics*, 54(1):75–96.
- Kopustinskas, V. and Arsenis, S. (2012). Risk analysis approaches to rank outliers in trade data. In *Advanced Statistical Methods for the Analysis of Large Data Sets*, pages 137–144. Springer.
- Lee, J.-G., Han, J., Li, X., and Gonzalez, H. (2008). Traiclass: trajectory classification using hierarchical region-based and trajectory-based clustering. *Proceedings of the VLDB Endowment*, 1(1):1081–1094.
- Liu, L., Daniels, H., and Triepels, R. (2014). Auditing data reliability in international logistics: An application of bayesian networks. In *Proceedings of the 16th International Conference on Enterprise Information Systems (ICEIS 2014)*.
- Murphy, P. R. and Daley, J. M. (1999). Edi benefits and barriers: comparing international freight forwarders and their customers. *International Journal of Physical Distribution & Logistics Management*, 29(3):207–217.
- Owens, J. and Hunter, A. (2000). Application of the self-organising map to trajectory classification. In *Visual Surveillance, 2000. Proceedings. Third IEEE International Workshop on*, pages 77–83. IEEE.
- Peffer, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V., and Bragge, J. (2006). The design science research process: a model for producing and presenting information systems research. In *Proceedings of the first international conference on design science research in information systems and technology (DESRIST 2006)*, pages 83–106.
- Riccio, A., Giunta, G., and Chianese, E. (2007). The application of a trajectory classification procedure to interpret air pollution measurements in the urban area of naples (southern italy). *Science of the Total Environment*, 376(1):198–214.
- Taylor, A. M. (2002). Globalization, trade, and development: Some lessons from history. Technical report, National Bureau of Economic Research.
- Tongzon, J. L. (2009). Port choice and freight forwarders. *Transportation Research Part E: Logistics and Transportation Review*, 45(1):186–195.

- von Alan, R. H., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 28(1):75–105.
- Yaqin, W. and Yuming, S. (2010). Classification model based on association rules in customs risk management application. In *Intelligent System Design and Engineering Application (ISDEA), 2010 International Conference on*, volume 1, pages 436–439. IEEE.

