Designing a new methodology for customs risk models

Alwyn Hoffman, Sonja Grater, Willem C Venter, Juanita Maree and David Liebenberg

Abstract

Effective risk management is a prerequisite to find an acceptable balance between the objectives of a customs operation and the streamlined flow of goods. The customs operations in many developing countries are characterised by high levels of physical inspections, with resulting disruption of trade flows, but with little positive impact for the regional economy. Most developed economies have moved towards customs risk management models based on the analysis of rich datasets that can be used to accurately determine the risk represented by a cargo consignment without physically stopping it. The use of such models can result in reduced physical inspections without increasing the risk to Customs of either losing income or allowing the influx of illegal contraband. It, therefore, represents a more optimal compromise between the interests of customs and those of trade, reducing the economic cost to the region and making the region more attractive to global economic partners. In this paper we develop a rigorous methodology that utilises electronic data transacted between Customs and trade to characterise the risk attributes of cargo consignments and then extract a model that can be applied in real time to minimise disruption of trade flows while reducing Customs risks to levels that are below set thresholds. This paper builds on previous work of Laporte (2011) and others but extends their results by developing a more detailed methodology to quantify the impact of a variety of input factors and demonstrating how an optimal set of inputs can be selected to arrive at an effective risk management model.

1. Background and introduction

In recent years many security problems, such as the 9/11 terrorist attack in America, have resulted in some customs administrations changing their priorities to include higher levels of risk profiling of importers to increase security and safeguard citizens (Manners-Bell, 2017, p. 34). Two common themes in cargo security programs have been to (1) gather more detailed information on which to gauge transaction risks, and (2) move the point of compliance further upstream along the supply chain and away from the point of entry (Prokop, 2017, p. 16). The success of supply chain movements globally depends greatly on the ability of customs to achieve an effective balance between trade facilitation and regulatory intervention, and not to place more restrictions in place (Creskoff, 2016, p. 319). Customs risk management, the screening of data for risk profiling, and the protection of the society must not increase disruptions to the flow of cargo in a supply chain through repetitive stops and inspections (Manners-Bell, 2016, p. 128).

Against this background, there exists very little empirical evidence that customs authorities are using well-designed statistical systems to identify possible high-risk or illegal transactions. Many of the existing systems only combine simple criteria, such as the importer code, the origin of the goods and the applicable tax regime. However, very few empirical or statistical models can be found in the literature.

Therefore, there is a need for more studies to develop statistical techniques that can be used by customs authorities to more efficiently target declarations that should be inspected.

A few studies, such as Laporte (2011), Davaa and Namsrai (2015) and Komarov (2016), designed possible methodologies to manage Customs risk. They explain the use of input factors, such as importer, freight agent, HS classification, country of origin, customs broker, transport mode, provenance and customs regime, in regression-type analysis to distinguish between high – and low-risk transactions. However, no alternative methodologies exist in the literature.

The limited number of historical studies motivates further exploration of a non-intrusive analytics-based approach to customs risk management. No systematic method has been designed to determine which input factors to consider and which should be selected for inclusion in a customs risk management model. The cited references, furthermore, provided no quantified indication of the relationships between the various input factors (e.g. customs office) and the operational customs performance (such as time delays). Such an analysis will indicate which areas of the overall customs operation suffers the most from inefficiencies, and if specific types of cargo or specific entities seem to be unfairly targeted.

In a previous paper (Hoffman et al., 2018), we characterised the effectiveness of an existing customs operation in terms of the contribution of specific input factors on outcomes, like time duration and the probability of finding an infraction, as well as the efficiency of current procedures used to select cargo consignments for inspections. Our finding was that the level of accuracy of the current risk engine appears to be quite low, with the result that most inspections do not produce infractions and do not lead to amendments of rejections, while adding substantial time delays to the operation. This paper aims to continue our previous work by utilising input–outcome relationships for customs processes currently applied in South Africa in order to quantify the capability of each input factor to contribute towards a more effective risk model. We furthermore extract risk models using a variety of empirical modelling techniques, including decision trees, linear regression and neural networks, and compare the capabilities of these techniques to accurately capture input–output relationships. The paper uses transaction-level trade data that was obtained from South African freight forwarders for the period September 2014 to September 2016.

The primary focus of this paper is the development of a rigorous methodology that provides more insight into the underlying relationships between inputs and outcomes than the approaches that were previously reported on (Laporte, 2011; Davaa & Namsrai, 2015; Komarov, 2016). The following specific research questions will be addressed:

1. What is the quantified capability of each identified input factor to predict customs outcomes before these have occurred?
2. Which combination of input factors and outcomes will provide the best risk-prediction capability?
3. Which empirical modelling technique can capture input–output relationships the most effectively?
4. To what extent could the current South African Customs decision-making process be improved?
5. What methodology, in general, should be followed for customs risk management?

This paper aims to contribute towards the creation of new knowledge about the use of empirical modelling techniques to improve customs risk management, in the process providing more accurate solutions to address the challenge an optimal balance between customs risk versus efficient trade flows.

The paper is structured as follows: section 2 describes the available data set, while section 3 describes the set of statistical modelling techniques that were used. Section 4 provides an overview of the methodology used to extract input–output models. Section 5 provides results and findings, while section 6 discusses the results. In section 7 we conclude with recommendations for customs operations and suggest future research work.

2. Description of the data

The authors of this study obtained trade transaction data from several freight forwarders in South Africa, in accordance with an agreement between the North-West University (NWU) and the South African Association of Freight Forwarders (SAAFF). The data that was used in this study represents EDI transactions exchanged between South African Revenue Services customs division (hereafter referred to as SARS Customs) and consignors of goods imported into South Africa from September 2014 to September 2016. The available dataset includes approximately 3.5 million transactions over the given time period.

For each transaction the following information was obtained:

- times and dates when electronic declarations were submitted by consignors and received by SARS
- name of the customs office where declarations were submitted
- HS chapter describing the nature of the cargo
- customs value of the cargo
- mode of transport through which goods entered into South Africa
- Customs Procedure Codes (CPC) reflecting the reason why goods were imported into South Africa
- countries of origin, export and import (some goods may be in transit via South Africa en route to a final destination elsewhere)
- codified identity of the entity submitting the customs declaration (preserving the anonymity of the declarants)
- detailed set of customs response codes communicated to the declarant for each transaction, together with the time and date for each code.

The customs response codes represent the customs outcomes that we would like to be able to predict (e.g. a decision to stop and inspect a consignment or finding an infraction). The corresponding date/time information indicates what the impact was on the flow of cargo in terms of time delays. The remaining data represents input factors that could reasonably be expected to impact the outcomes.

Table 1 provides a summary of the input factors and the level of detail that was included in the data. For the purpose of this study we decided to only focus on the use of categorical input factors; for this reason, customs value (that represents a continuous input value factor) was excluded from the set. Due to the large number of HS classifications we decided to narrow it down to a smaller number of cargo categories, based on HS chapter (i.e. the first 2 digits of the HS code). The naming convention used for this reduced number of HS categories is described in Table 2.

Table 1: Input factors reflecting customs declaration processes

Input factor	Number of categories	Examples
Import/Export	7	Imports, Ex-bond, In transit
Customs Office	36	Durban, Cape Town, etc.
CPC Code	31	10, 11, 12, etc.
Previous CPC Code	23	00, 14, 20, etc.
Country of Origin	237	GB, CN, GE, etc.
Country of Export	222	GB, CN, GE, etc.
Country of Import	197	ZA, ZM, ZW, etc.
Transport Code	9	Ocean, Road, Rail, etc.
Consignors	310	#0, #1, #7, etc.
HS chapter	18	Animal, Chemical, etc.

Table 2: Description of reduced HS chapter codes

Cargo category	HS chapter value range
Animal	HS chapter <= 5
Vegetable	5 < HS chapter <= 15
Food	15 < HS chapter <= 24
Mineral	24 < HS chapter <= 27
Chemical	27 < HS chapter <= 38
Plastic	38 < HS chapter <= 40
Hide	40 < HS chapter <= 43
Wood	43 < HS chapter <= 49
Textile	49 < HS chapter <= 63
Footwear	63 < HS chapter <= 67
Stone & Glass	67 < HS chapter <= 71
Metal	71 < HS chapter <= 83
Machinery	83 < HS chapter <= 85
Transport	85 < HS chapter <= 89
Miscellaneous	89 < HS chapter <= 97
Service	97 < HS chapter <= 99
Other	99 < HS chapter

Table 3 provides a summary of the customs response codes that can be received for any given transaction, and that indicate the actions by customs for that specific transaction.

Table 3: Customs response codes

Customs response code	Description
1	Release
2	Stop for physical inspection at unpack depot or X-ray scanning
4	Refer to other governmental agency (OGA)
6	Reject declaration
13	Supporting documents required
26	Request adjustment to declaration
27	Accept
31	Request additional supporting documents
33	Supporting documents received
36	Booked for physical inspection

3. Statistical techniques used to characterise input–output relationships

To develop a model that can predict customs outcomes from data that is available before any intrusive action has occurred, it is necessary to evaluate the usefulness of the available input factors that can help explain various characteristics of the expected outcomes, such as the level of risk attached to a consignment. We shall use the term 'explanatory variables' for these input factors. The quality of an empirical model derived from a set of available data (normally called the training set) depends on a number of factors; most important of these is whether the desired outputs can in fact be derived or inferred from the available inputs. In a case like the development of a customs risk model one would expect the outcomes to depend on a variety of input factors; some of these will hopefully be contained in the available set of explanatory variables.

The usefulness of such an empirical model is not reflected by how well it can fit input–output relationships present in a given training set, but whether it can learn the ability to generalise: once trained on a given set of data it must also be able to provide useful responses when fed with previously unseen data that was not used during the training process. The reason for this requirement is obvious: if a customs risk model has already been trained on historical data, its practical application will involve feeding it with new data for which the outcomes are still unknown. Such new data was not yet available when the model was trained; the real test for the model is thus how well it generalises out of the training sample.

The capability of an empirical model to explain input–output relationships inside the training set tends to increase with increasing model complexity; however, it tends to lose its ability to generalise if it is allowed to become too complex (Bishop, 1995). Model complexity largely depends on the number of free parameters (also called degrees of freedom) present in the model and of which the values are optimised during the training process; as the number of inputs increase the number of free parameters and model complexity will also tend to increase. The best models are normally those that achieve an acceptable

modelling accuracy over the training set using the smallest number of inputs factors and degrees of freedom. It is, therefore, important to reduce the set of candidate input factors as far as possible before the model is trained to help prevent overtraining and subsequent bad generalisation capability. For this purpose, each potential explanatory variable should be assessed separately and in combination with others for its ability to correctly predict the desired outputs before it is considered for inclusion into a model. We will use analysis of variation (ANOVA) as well as linear correlation analysis to evaluate the respective explanatory variables before we proceed to extract empirical models.

3.1 Analysis of variation

In Table 4 we show the results of an ANOVA applied to the different explanatory variables. We investigated the extent to which these variables can explain the following customs outcomes:

- the duration of the customs process
- the likelihood
  - of a request for additional documentation
  - of being stopped
  - of being stopped and inspected
  - of an infraction to be found.

The F-statistic measures the ratio of between-class-variations (e.g. variation of infraction probability between different customs offices) and within-class-variations (e.g. variations of infraction probability over time for the same customs office). A high F-value indicates, on average, larger differences between categories than within categories, implying that the respective category variable does have the ability to help explain the respective outcome values.

Table 4 shows that, for all of the outcomes, at least some of the explanatory variables have significance. Duration and request for additional documents produce, on average, larger F-statistics, indicating that they may be easier to predict using the available explanatory variables. For all of the outcomes, the F-values are sufficiently large to justify proceeding with the development of prediction models based on the available set of explanatory variables.

Table 4: F-statistics extracted through one-way ANOVA

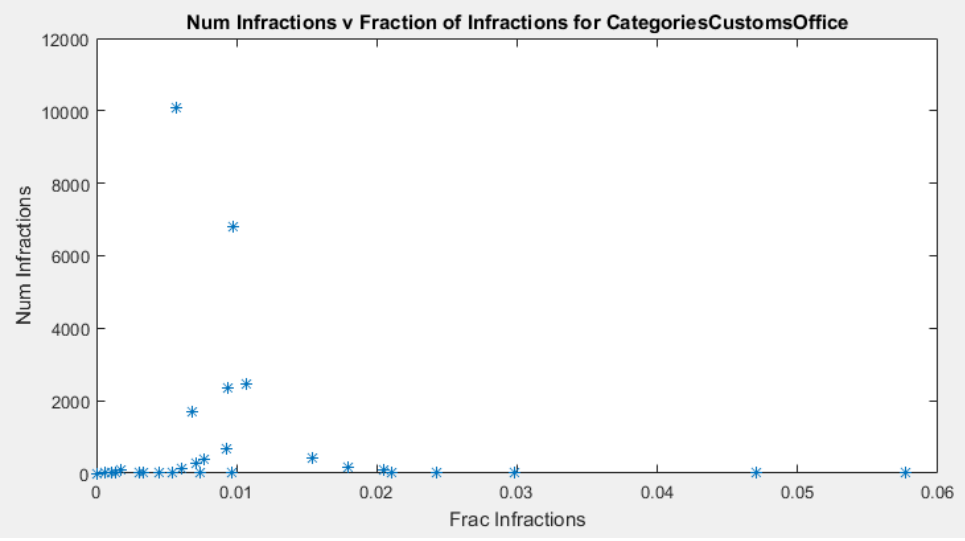
Input Factor	Duration	Req_Add_Docs	Stopped	Stopped_ Inspected	Infraction
CustomsOffice	1150.8	1356.2	43.6	168.8	82.2
HSChapter	892.1	8934.3	32.6	114.1	105.8
TransportCode	2105.4	4533.6	204.1	731.9	499.1
SPSource	206.7	809.5	84.2	234.2	126.5
CpcCode	423.4	352.7	2.6	10.6	93.2
PreviousCpc	118.4	834.1	4.2	22.6	75.8
CountryOfOrigin	50.1	329.2	4.9	13.3	17.8
CountryExport	59.9	352.0	4.8	18.1	17.7
CountryImport	25.4	29.7	1.8	2.0	3.7

3.2 Investigating the classification ability of individual explanatory variables

In this article our primary focus is on the prediction of infraction probability for cargo consignments. In order to determine the difference in the incidence level of infractions within different input categories, we calculated the fraction of infractions present within each category, using each of the selected explanatory variables as the basis for categorisation. Should a specific category be characterised by a high historical infraction rate, membership of that category can be used by a customs risk engine as the basis for selecting future consignments falling within the same category. Should historical behaviour continue into the future, this should allow the risk engine to achieve higher than random success rates.

The degree of success that can be achieved with such an approach will, however, also depend on the number of infractions present within each category. If some categories can easily be identified as being associated with high infraction rates, but these categories only represent a small fraction of the overall number on infractions, the degree of success that can be achieved by selecting such categories will be limited. This is illustrated in Figure 1 for customs offices, which shows the number of infractions against fraction of infractions for each of the 36 customs offices. It can be seen that some customs offices display a much higher than average incidence of infractions, in some cases more than 5 per cent compared to a population average of about 0.6 per cent. It is, however, also clear that all of the customs offices with much higher than average infraction rates also contain only a very small fraction of the total number of infractions present in the entire population. The two largest customs offices, Durban and ORT International Airport, display by far the largest number of infractions, but their averages are not very different from the population average (largely because between them they dominate population averages). Using the customs offices with high infraction rates as predictors will, therefore, make some contribution but will only allow a limited fraction of all infractions to be correctly selected.

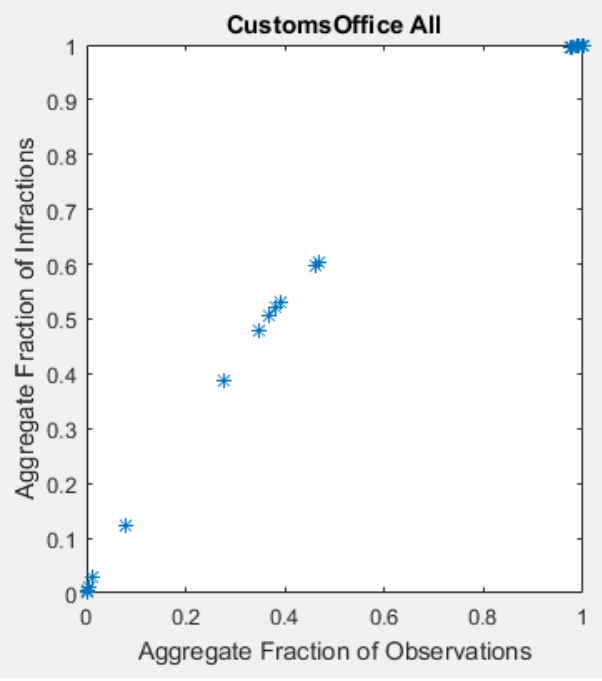
Figure 1: Number of infractions vs fraction of infractions for different customs offices



The accumulative effect of what was described above is shown below in Figure 2. This graph was constructed by selecting consignments for inspection, starting with consignments falling into those categories that historically displayed the largest infraction levels, and gradually adding more categories until all have been selected. If all categories contained the same infraction rates, the resulting graph of aggregate fraction of infractions selected versus aggregate fraction of all observations selected would be

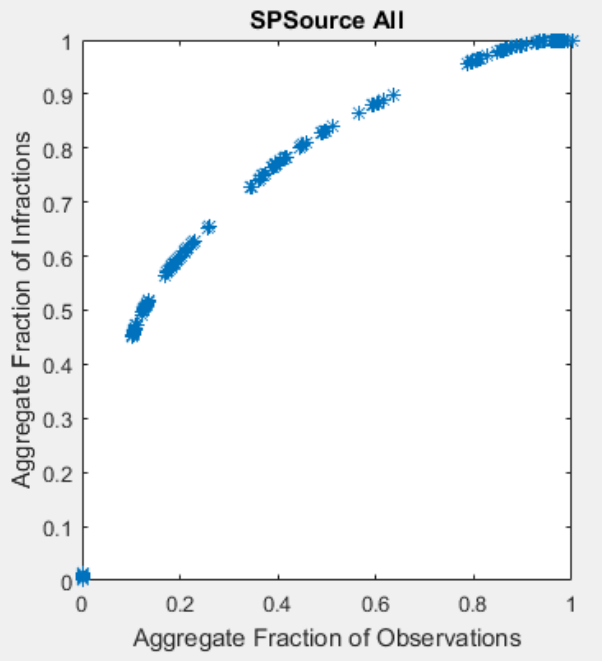
a straight line with gradient 1 from (0,0) to (1,1). In this case the gradient of the graph is initially steeper than 1, as the categories with the higher incidence rates are selected first. As categories with lower incidence rates are added, the graph becomes less steep. The effect of the large fraction of infractions present in the two largest customs offices is clearly illustrated in the graph: when these customs offices are added to those already selected, big jumps occur in the graph. If these two large customs offices are omitted from the criteria for inspections, then only a relatively small fraction of infractions are present in the selected group; once these two categories are added, most infractions are included but then the selected group contains almost all observations. This clearly illustrates why using only one explanatory variable in a risk classifier will lead to minimal success.

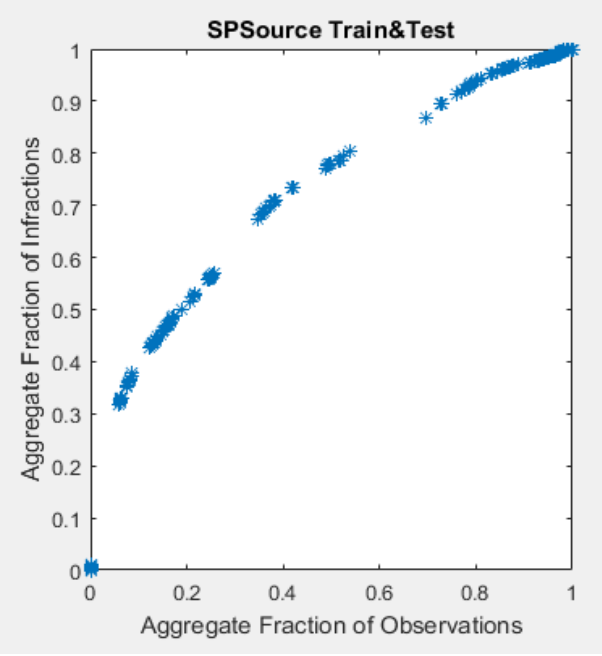
Figure 2: Aggregate fraction of infractions found vs aggregate fraction of observations selected based on customs offices



In Figure 3 we show similar results but using consignor identity as category variable. It is clear that this variable has a superior ability compared to customs office to separate consignments with a high infraction incidence from the rest of the population, as the graphs start off with steeper slopes. We also illustrate the importance of evaluating selection criteria out-of-sample to obtain a realistic indication of its ability to generalise outside of a training set. The graph on the left-hand side shows the aggregate fractions when categories are selected after the category averages were determined using all available samples. This is obviously not a realistic approach, as in practice the categories must be characterised based on historical samples only before a rule is applied to new samples. We applied the correct approach in the graph on the right-hand side, where the available data was first divided into a training and test set of equal size, and where the testing data follows after the training data in calendar time, which will always be the case in practice. It can be seen that the selection ability of this variable was slightly reduced when applying training set averages as selection criteria to test set data.

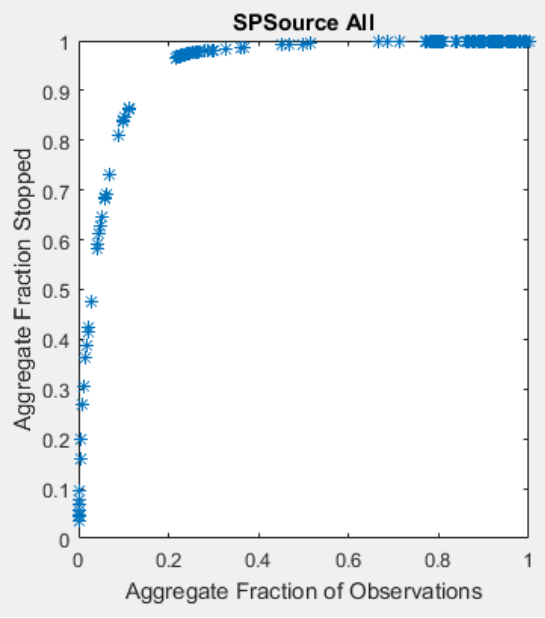
Figure 3: Aggregate fraction of infractions found vs aggregate fraction of observations selected based on consignors

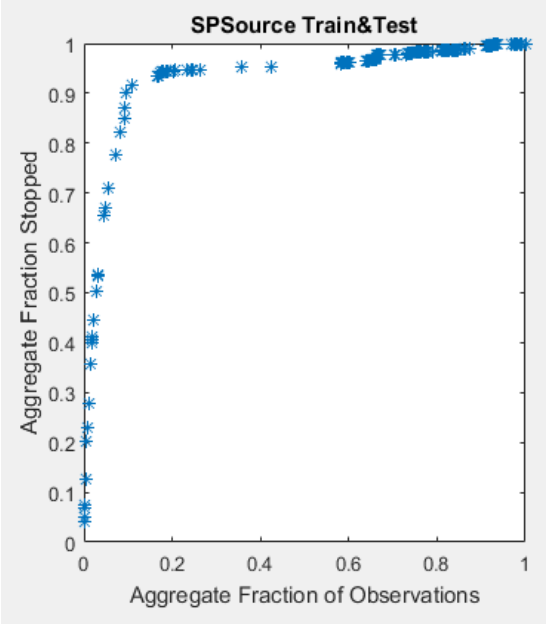




In Figure 4 we show similar results to demonstrate the ability of Consignor identity to correctly predict the incidence of consignments being stopped by customs. It can be seen that a much higher selection accuracy for stopped consignments will be achieved when selecting a given fraction of total observations, compared to the selection accuracy for infractions, as the graph has a much steeper gradient initially and a much lower gradient once the 'elbow' has been passed. This indicates that when customs select consignments for stops, the identity of the consignor seems to play a very important role. The success of this strategy to actually find infractions is not quite as high, as indicated by the results in Figure 3.

Figure 4: Aggregate fraction of stopped consignments found vs aggregate fraction of observations selected based on consignors





3.3 Correlation analysis between input factors and outcomes

The previous section described the ability of different input factors to predict whether consignments will be associated with specific eventualities. In order to perform a direct comparison between the impacts of different input factors on the various outcomes we implemented a linear correlation analysis. This was done in the following manner:

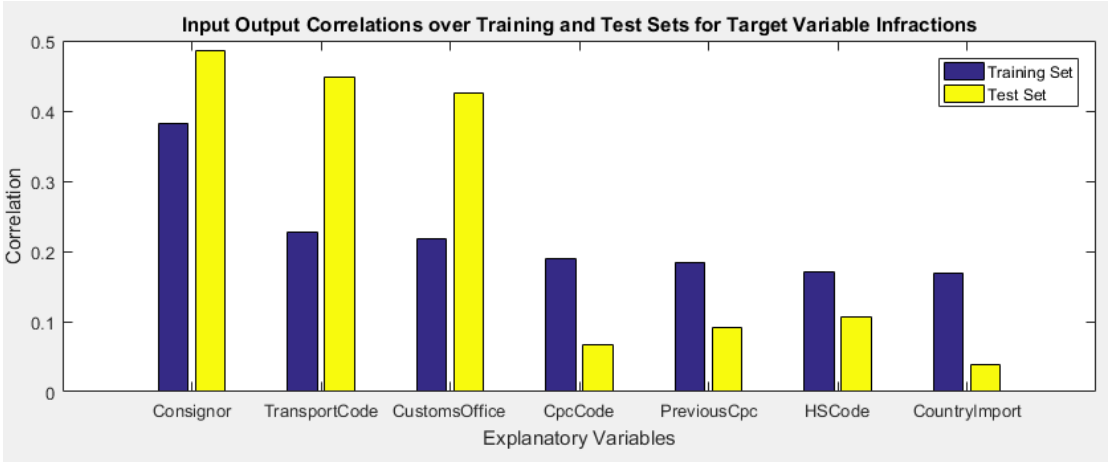
1. For each category of each input factor (e.g. Durban or Cape Town for category customs office) the accumulated average value of each outcome (e.g. fraction of infractions) was calculated as function of time as from the start of the observation period up to the end of each specific month. These averages represent the behavioural characteristics of the respective categories up to that point in time.

2. For each new observation falling in a subsequent month its category memberships were determined (e.g. customs office: Durban; HS chapter: textiles; transport mode: maritime). That observation was then allocated the accumulated averages for the categories to which it belongs as determined at the end of the previous month. Each observation thus inherits the attributes of the categories that it belongs to; as these attributes are continuous variables, it can be used as basis for a correlation analysis with the observed outcomes.

For each new observation and each input factor (e.g. customs office) the Pearson correlation was calculated between the incidence of a specific eventuality for that specific consignment (e.g. whether it contains an infraction) and the historical average for the same eventuality within that specific category (e.g. Durban). Should historical behaviour within that category continue, a positive correlation value would be obtained; if the respective categorisation has no impact on that eventuality then a close to zero correlation will be obtained.

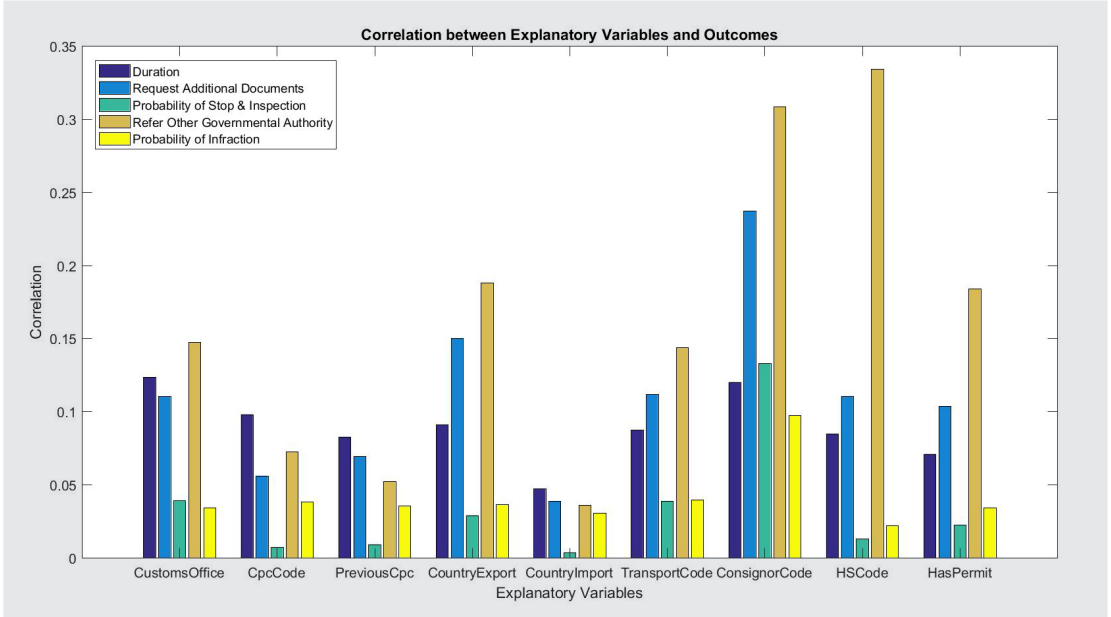
In a previous article (Hoffman et al., 2018) we performed correlation analysis on the same training set and found relatively weak relationships (correlation coefficients of below 5%) between explanatory variables and outcomes. This was partly caused by the fact that we are dealing with a very unbalanced dataset: only about 0.6 per cent of all infraction outcomes are true, resulting in the relationship between this small fraction of infractions and category membership to be obscured by the noise levels in the 99.2 per cent data with false outcomes. We can overcome this limitation by using a weighted dataset where the observations with true outcomes are multiplied to obtain a balanced set with an equal number of true and false outcomes. These results are displayed in Figure 5. The difference is apparent as several correlation coefficients are above 20 per cent. We performed the correlation calculations separately for a training set (first 50% of all weighted observations) and test set (last 50% of all weighted observations), to investigate the degree to which the relationships between inputs and outcomes remain constant over time. The explanatory variables are ranked based on size of correlation coefficient in the training set. While the ranking of explanatory variables based on correlation coefficient is more or less retained in the test set, there are considerable differences with respect to some categories, for example, CPC code and country of import. This emphasises the fact that a model extracted from the training set will not always perform as well over the test set.

Figure 5: Weighted correlation between probability for infraction and 7 explanatory variables



When similar correlations are calculated with respect to other customs outcomes, the same input factors tend to dominate but correlations are significantly larger, as can be seen in Figure 6. This indicates that, while consignor identity is used by Customs as a primary determinant for stops, inspections and requests for additional documents, this strategy meets only with limited success in terms of infractions that are found. It is, therefore, possible that some consignors are unjustifiably discriminated against, with significant implications in terms of overall time delays experienced. The results based on correlation analysis also support the results obtained in the previous section using average historical incidence of eventualities, as consignor is once again clearly the most useful explanatory variable.

Figure 6: Correlation between explanatory variables and different outcomes



4. Methodology for extracting input–output models from the data

In order to allow a direct comparison between our results and those that were previously published, we firstly implement methods based on individual inputs or simple combinations of inputs, similar to the approach as published by Laporte (2011). Secondly, we implement both linear regression and logistic regression models, also similar to the approach of Laporte. We then proceed to implement a neural network-based model as well as a decision-tree model to verify if these more sophisticated techniques can improve upon the performance of simpler techniques.

Each model produces an output, called a risk score, which is compared against a risk threshold; should the score exceed the threshold it is assumed that the event being predicted has occurred—such an observation will typically be selected for inspection. By applying the same threshold to all risk scores generated by a particular model it can be determined what fraction of all observations will be selected for inspection by that model; by comparing the predicted outcomes with true outcomes it can also be determined what the success rate was (i.e. the fraction of total events, e.g. infractions, that were present in the selected set). The Laporte study arbitrarily selected a number of threshold values ranging from 0.01 to 0.5 to create a range of score intervals. We decided to identify the specific threshold values to obtain specific success rates; the required success rates were set at 50 per cent, 80 per cent, 90 per cent and 95 per cent. Different models can then be directly compared in terms of the fraction of observations that must be selected to achieve a desired 'hit rate'.

The description of model extraction as reported by Laporte creates the impression that models were extracted from the entire population after which the models were applied to the same data to determine model performance, as no separate training and test set results were reported in that reference. Such an approach can, of course, not be used in a practical customs risk engine, as a model derived from historical data can thereafter only be applied to new data. We, therefore, extracted models both from the entire datasets (to allow comparison against the results of Laporte) as well as from a training set that represented the 50 per cent of data that occurred first, with model evaluation done on the remainder of the data. It can then be observed to what extent the performance of a model deteriorates or is maintained out of the training sample.

4.1 Simplistic models

We followed Laporte's definition of simplistic models as closely as possible. First each of the explanatory variables as extracted in section 3.3 are evaluated separately, using its value (historical infraction incidence rate) as risk score. Then three combined models are implemented, using the country of origin, HS chapter and consignor as inputs:

- using a simple average over the three inputs as risk score
- using a weighted average (with weights 0.5, 0.3 and 0.2) over the three inputs as risk score
- using the maximum of the three inputs as risk score.

4.2 Econometric models

The same explanatory variables that were used for the correlation analysis in 3.3 above were also used as inputs for econometric models. We experimented with different numbers of input factors, selecting inputs based on their linear correlations with the outcome being modelled. In addition to linear regression and logistic regression we also extracted a neural network model. Neural network models will have benefits over linear regression models should there be significant nonlinear aspects in the input–output relationships (Bishop, 1995). A typical example would be where the incidence of an eventuality starts to increase more rapidly once a specific threshold value in some input factor is exceeded.

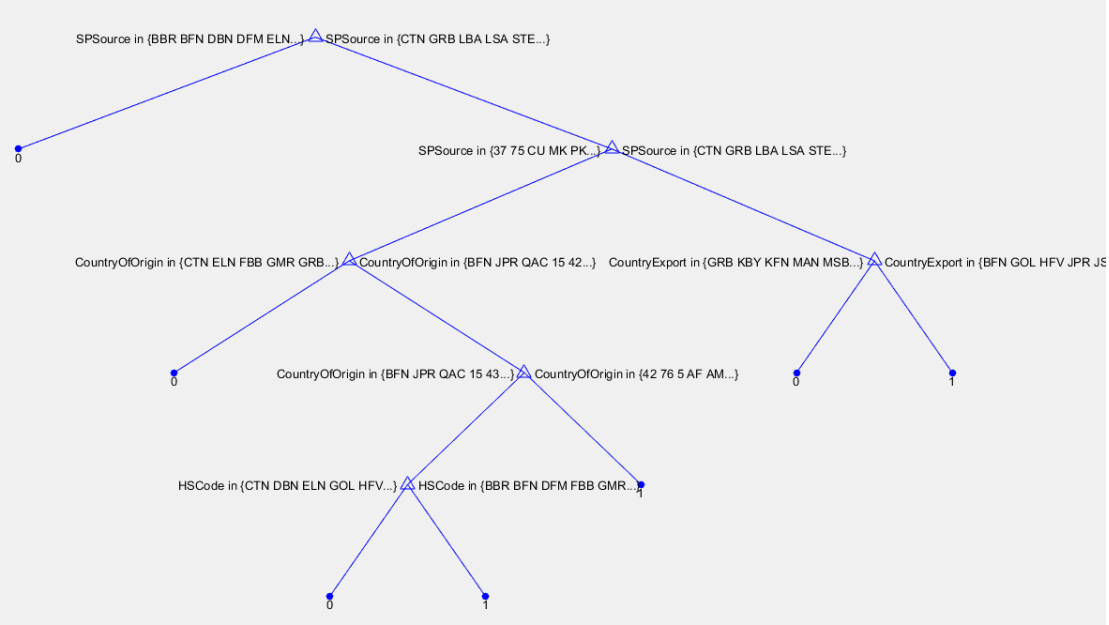
To investigate the presence of such behaviour we extracted feed-forward multilayer perception networks that use the same sets of inputs as the regression models. We used a single hidden layer of which the size is adjusted based on the number of inputs, with sigmoidal transfer functions in the hidden layer and linear transfer functions in the output layer. As neural networks have a tendency to overtrain, we added the use of a validation set that made up 10 per cent of the training set values; training was terminated at the point where the modelling error on the validation set started to increase. Both the input and hidden layers were restricted to not more than ten nodes each; the overall number of degrees of freedom in the models would therefore be less than 120. Given the size of the population (about 3.5 million observations), it was not necessary to apply specific regularisation techniques as the large training set combined with the use of the validation set would help prevent the model from being over fitted to specific training samples.

As mentioned earlier, the target data is highly unbalanced—for our dataset only about 0.6 per cent of cases contained infractions and only a small percentage was selected for scrutiny. If such an unbalanced dataset is directly used for training a regression or neural model, the optimisation technique that determine optimal values for the model parameters will tend to be dominated by to the large number of samples with outcome '0' (e.g. no infraction); this may result in the optimisation process getting stuck in a trivial solution where all observations produce an outcome of '0', providing a high accuracy measured over all outcomes (as e.g. 99.2% of all outcomes contain no infraction) but of no practical value. In order to overcome this problem, we used weighted training sets, by multiplying each of the true outcomes with a factor that equals the ratio between the initial number of true and false outcomes. As a result, both the training and test sets contained about 3.5 million observations, with approximately 50 per cent false and true outcomes in each set.

4.3 Classification trees using original inputs

For our selected problem domain, most of the available input data is inherently categorical in nature (e.g. a cargo consignment can only come from one specific country of origin and belong to a specific HS chapter). For the previous set of modelling techniques, we translated these categorical values into continuous ones as these types of models perform best when fed with continuous input data. It is, however, also possible to directly use the categorical input values in their original format; the obvious model type would then be decision trees. A decision tree consists of many branches, and at each branch point one or more categorical variables are used to make a decision regarding which way to branch. Once the probability of a specific outcome is high enough the branch terminates in an outcome (e.g. an infraction occurred or did not occur). If the outcomes are also categorical then the technique is called classification trees. An example of a simple classification tree is shown in Figure 7.

Figure 7: Example of a classification tree



To investigate if classification trees can resemble the customs decision-making and infraction incidence process more accurately than models using continuous inputs, we trained classification trees that take any number of the original categorical variables as inputs. In the training of decision trees, care must be taken not to overtrain, as a sufficiently large number of branching points will allow any number of unique inputs observations to be correctly classified. Overtraining is prevented by limiting the maximum number of levels in the tree: a tree with only one level will classify all observations as either ones or zeros; with two levels at least one rule will be applied to separate the two classes; as the number of levels are increased more conditions can be defined to use as basis for more refined classification.

To determine at which level the tree must be terminated, we used a validation set defined similarly as in the case of neural network training sets. An initial tree was extracted using no limit on tree level, and the performance of the tree was tested on the validation set. The number of tree levels was then gradually reduced until the classification performance for the validation set reached its maximum value. All of these trees were then also applied to the test set to determine how well the selected tree performs on the test set.

5. Results and findings

Each of the modelling techniques described above were applied to the available data to predict both the probability of specific customs decisions (e.g. to stop and inspect) and the probability of finding an infraction. It must be appreciated that the process that generated the available dataset was not the actual process that generated the incidence of infractions, but only the process as applied by the respective customs authority. It is known that the relevant customs authority in this case applied a set of rules and procedures to generate decisions to stop, inspect, etc.; once such a decision was implemented an infraction that was potentially present in the respective consignment may or may not have been found. Any deficiency in this process, both to correctly stop or scrutinise risky consignments and to find infractions that are present, will be directly reflected in the data. No empirical modelling technique will be able to correct such a customs deficiency, as an empirical model can at best be as good as the data from which it is derived.

It must, therefore, be stated that the models that were extracted mimic the customs process rather than the true incidence of infractions and should therefore be evaluated primarily on that basis. This is supported by the results in section 3, where it could be seen that there are stronger correlations between input factors and the incidence of specific customs decisions than between the same input factors and the incidence of infractions.

5.1 Simplistic models

When applying simplistic models as described in 4.1, the results as displayed in Table 5 were obtained. For direct comparison purposes we limited the use of these models to prediction of infractions, as Laporte only applied these techniques to finding infractions and not to predict customs decisions. For each required selection accuracy (50%, 80%, 90%, 95%) it was determined what fraction of observations would have to be inspected to achieve the corresponding selection accuracy. We repeated this process for two cases:

- by extracting the threshold values from the entire set and then applying the thresholds to the entire set
- by extracting the thresholds from a training set only and then applying them separately to the training and test sets—this would be the more correct approach.

Table 5: Classification results for infraction using simplistic models

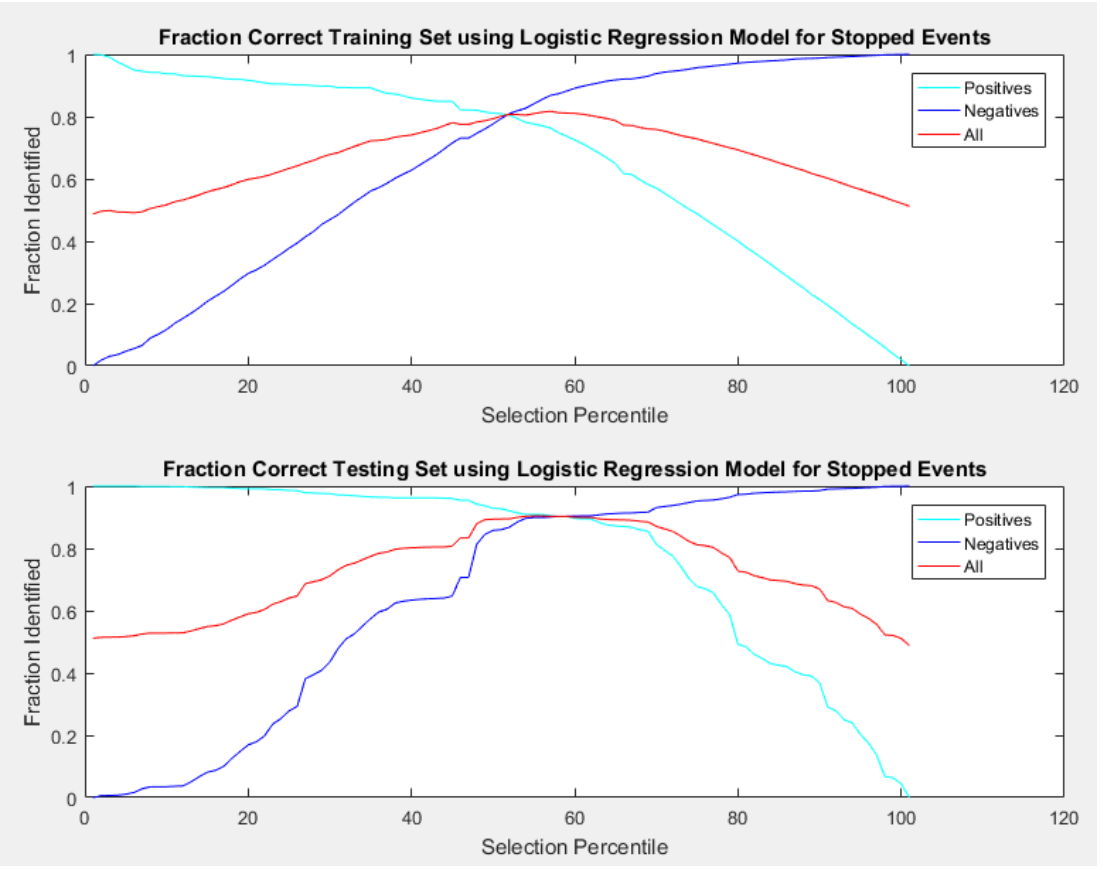
Fraction Infractions Found	Customs Office	CPC Code	Country Export	Country Import	Consignor	HS chapter	Country Origin	Simple Average	Weighted Average	Maximum
All										
0.50	0.37	0.94	0.41	0.97	0.17	0.72	0.43	0.14	0.13	0.22
0.80	0.98	0.94	0.82	0.97	0.49	0.85	0.72	0.67	0.58	0.58
0.90	0.98	0.94	0.82	0.97	0.87	1.00	0.87	0.86	0.78	0.90
0.95	0.98	0.94	0.97	0.97	0.87	1.00	1.00	0.92	0.90	0.90
Train										
0.50	0.39	0.97	0.43	0.97	0.14	0.74	0.43	0.17	0.18	0.15
0.80	0.98	0.97	0.79	0.97	0.41	0.85	0.72	0.58	0.46	0.56
0.90	0.98	0.97	0.97	0.97	0.77	1.00	0.86	0.77	0.66	0.79
0.95	0.98	0.97	0.97	0.97	0.77	1.00	1.00	0.88	0.84	0.92
Test										
0.50	0.97	0.97	0.42	0.97	0.25	0.75	0.42	0.20	0.27	0.23
0.80	0.97	0.97	0.78	0.97	0.76	0.86	0.85	0.75	0.63	0.62
0.90	0.97	0.97	0.96	0.97	0.76	1.00	0.90	0.87	0.83	0.78
0.95	0.97	0.97	0.96	0.97	0.84	1.00	1.00	0.93	0.92	0.91

The quality of the selection measures is reflected by how small a fraction of observations needs to be inspected in order to achieve a specified selection accuracy. As could be expected based on the results of section 3, consignor performed the best amongst the individual inputs. Consignor on its own is only marginally outperformed by the combination techniques, and only for some selection accuracies. It can also be observed that when the model extraction technique is correctly applied, using separate training and test sets, the performance is significantly worse compared to the approach where the model is applied to the same dataset from which it was derived.

5.2 Econometric models

The econometric models accept up to 9 input variables. Outcome accuracies are calculated for both the training and test sets and for both positive and negative outcomes. The impact of gradually changing the threshold level to separate positive and negative outcomes is illustrated in Figure 8, with separate graphics displaying the results for training and test sets, and a separate graph provided for positive, negative and all outcomes. Starting off with a zero-threshold level, all observations are classified as positive, with a resulting 50 per cent accuracy: all positive target values are classified correctly, and all negative target values are classified incorrectly. As this threshold is increased more observations will fall into the negative category; if the model possesses any ability to correctly classify observations then the classification accuracy for the category 'all' should increase, reaching a maximum value before decreasing again as the threshold value approaches 1.

Figure 8: Fraction stopped consignments correctly identified by logistic regression model as function of fraction selected



The results based on regression and neural network models are shown in Table 6. The coefficients of determination (or R^2) are somewhat lower that the figures reported by Laporte; this indicates that for our dataset the input–output relationships could be somewhat weaker. This coefficient is also lower for the test set compared to the training set, indicating that the models extracted from the training data do not fit the test data quite as well, also in line with expectations.

Apart from some marginal cases, the regression models could not significantly improve on the performance of the simplistic models. This indicates that the additional input variables, over and above the three variables used in the simplistic models, do not significantly contribute towards model performance. It can also be seen that the neural networks perform slightly better compared to the regression models, indicating that its nonlinear modelling capability is of some benefit in capturing the input–output relationships.

Table 6: Classification results for predicting infraction using different model types

	Linear regression		**Logistic regression**		**Neural network**	
Model Type	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
R^2	0.18	0.15	0.20	0.16	0.20	0.16
Fraction infractions found	**Fraction of observations inspected**					
50	0.32	0.48	0.32	0.48	0.32	0.46
80	0.64	0.74	0.63	0.73	0.65	0.66
90	0.81	0.85	0.81	0.83	0.83	0.75
95	0.90	0.90	0.89	0.90	0.91	0.81

In Table 7 we show similar results to predict the incidence of customs stops. It can be seen that all techniques perform much better compared to the case of predicting infractions. This confirms that the modelling techniques used do have the ability to capture the underlying behaviour present in the customs decision-making process with reasonable accuracy.

Table 7: Classification results for predicting stops using different model types

Model	**Linear regression**		**Logistic regression**		**Neural network**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
R^2	0.16	0.26	0.39	0.61	0.43	0.66
Fraction infractions found	**Fraction of observations inspected**					
50	0.27	0.22	0.30	0.21	0.26	0.20
80	0.49	0.34	0.65	0.30	0.49	0.29
90	0.68	0.42	0.81	0.83	0.68	0.39
95	0.74	0.65	0.99	0.53	0.77	0.57

5.3 Classification trees

A different approach had to be followed in the case of classification trees, as this technique does not allow a threshold to be adjusted in order to achieve the desired selection accuracy. Instead of changing a threshold, the model performance can be moderated by limiting the allowed tree level. As shown in Table 7, the maximum tree level when extracting the model from the training set was 1371. The case with tree level equal to 1 represents the trivial case where all observations are classified as true. As the tree level increases, fewer observations from the training set are classified as true. At the same time, the fraction of events that are correctly selected also goes up as more branches are added to the tree; this behaviour is displayed in Figure 9. As is evident from Table 7 and Figure 10, the fraction correctly selected infractions from the test set reaches a maximum value for an optimal tree level of about 20. For this level of tree complexity, 23 per cent of test set observations are selected to produce a selection accuracy of 58 per cent of infractions. This is slightly superior compared to the best results for the simplistic or econometric models.

We can also compare our results with those of Davaa and Namsrai (2015). Their risk engine increased the incidence of infractions from 0.05 per cent in the lowest risk categories to 0.22 per cent (i.e. by approximately a factor of 4) in the highest risk categories. Our classification tree increases the average incidence of infractions from 0.6 per cent (when all observations are selected) to 1.7 per cent in the test set when the optimal tree level is used (i.e. by approximately a factor of 3). We can, therefore, state that, for our dataset, the performance is comparable to the results achieved by Davaa and Namsrai.

Table 8: Classification results for infraction using classification trees

Tree Levels	1	2	3	15	20	101	536	1371
Training								
Fraction selected	1.00	0.20	0.26	0.27	0.28	0.27	0.25	0.22
Hit rate	0.008	0.026	0.023	0.022	0.022	0.024	0.028	0.032
Fraction correct	0.52	0.72	0.73	0.74	0.74	0.77	0.81	0.83
Fraction correct no event	0.00	0.80	0.75	0.73	0.72	0.74	0.76	0.78
Fraction correct event	1.00	0.65	0.72	0.75	0.77	0.79	0.86	0.88
Validation								
Fraction correct	0.52	0.72	0.73	0.74	0.74	0.76	0.80	0.83
Fraction correct no event	0.00	0.80	0.75	0.73	0.72	0.74	0.75	0.78
Fraction correct event	1.00	0.65	0.71	0.75	0.76	0.78	0.85	0.87
Testing								
Fraction selected	1.00	0.14	0.20	0.22	0.23	0.21	0.19	0.16
Hit rate	0.006	0.021	0.017	0.017	0.016	0.017	0.016	0.016
Fraction correct	0.47	0.67	0.67	0.68	0.68	0.68	0.66	0.63
Fraction correct no event	0.00	0.87	0.80	0.79	0.77	0.79	0.81	0.84
Fraction correct event	1.00	0.44	0.53	0.56	0.58	0.54	0.48	0.39

We display similar results for the case of customs stops prediction in Table 9. It can be seen that classification trees display a high level of accuracy to correctly model the decision-making processes used by Customs to decide which consignments to stop: by selecting only 13 per cent of cases 86 per cent of those that will be stopped by customs will be identified.

Table 9: Classification results for stops using classification trees

Tree Levels	1	2	3	15	20	104	456
Training							
Fraction selected	0.00	0.25	0.23	0.17	0.16	0.10	0.08
Fraction correct	0.51	0.86	0.87	0.89	0.90	0.94	0.96
Fraction correct no event	1.00	0.75	0.77	0.83	0.84	0.90	0.92
Fraction correct event	0.00	0.97	0.97	0.96	0.96	0.99	0.99
Validation							
Fraction correct	0.51	0.86	0.84	0.82	0.83	0.79	0.68
Fraction correct no event	1.00	0.75	0.77	0.83	0.84	0.90	0.92
Fraction correct event	0.00	0.98	0.91	0.81	0.82	0.68	0.43
Testing							
Fraction selected	0.00	0.17	0.16	0.14	0.13	0.09	0.06
Fraction correct	0.49	0.89	0.88	0.86	0.86	0.77	0.63
Fraction correct no event	1.00	0.83	0.84	0.87	0.87	0.91	0.94
Fraction correct event	0.00	0.94	0.93	0.86	0.86	0.63	0.34

Figure 9: Fraction training infractions correctly selected as function of classification tree level

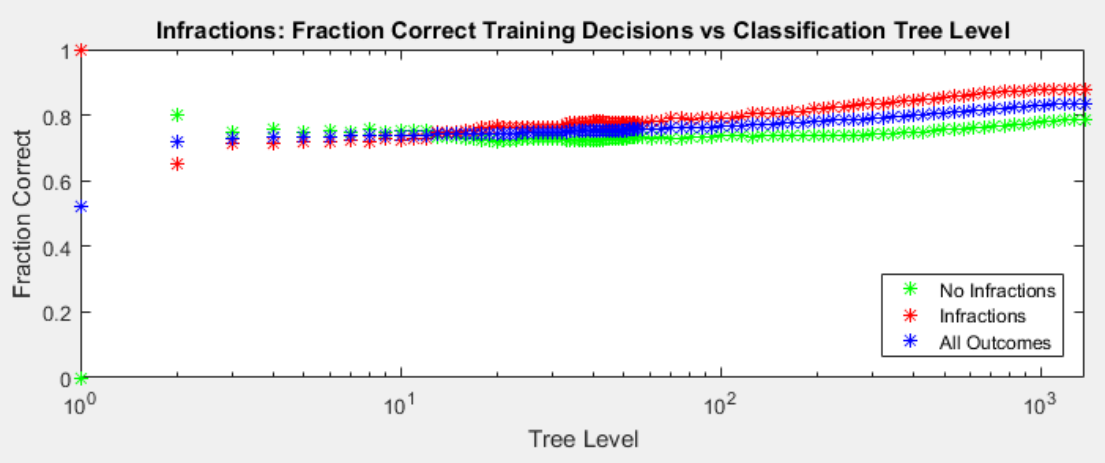
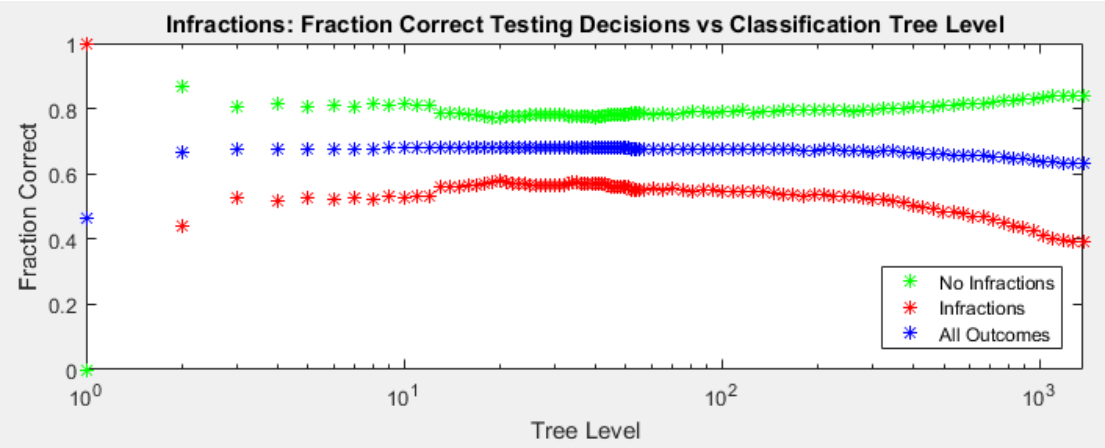


Figure 10: Fraction testing infractions correctly selected as function of classification tree level



6. Discussion

When comparing these results against the figures reported by Laporte (2011), significant differences are observed. This could have been expected as the models were extracted from very different datasets: Laporte used just over 100,000 observations generated by Senegalese Customs, whereas we used about 3.5 million observations generated by SARS Customs. It is quite possible that totally different behaviour would be present both in terms of the real incidence of infractions (resulting from the behaviour of consignors and their service providers) and in terms of the efficiency of the respective customs authority. We specifically note that, while the demonstrated approaches to risk assessment do improve the accuracy of selection compared to a random process, the accuracies of selection achieved on our dataset are not as high as those reported by Laporte. While Laporte claimed the ability to reach accuracies of 90–96 per cent by inspecting only 20–26 per cent of consignments, in our case the inspection rate had to be pushed up to 75–83 per cent to achieve the same. For lower accuracies the difference in performance is also prominent: while Laporte achieved accuracies of more than 50 per cent by selecting only 2–3 per cent of consignments, with our data the same techniques must select about 20 per cent of observations to exceed a 50 per cent accuracy.

It can be observed that other researchers applying similar techniques also did not achieve nearly the same level of selection accuracy as was reported by Laporte (2011). Selection accuracy in this case refers to the ability to correctly select only cargo consignments that include infractions for inspections. The difference in results may be partially explained by the fact that Laporte may have extracted his selection model from the entire data set, and that his selection accuracy results were then extracted from the same data used to produce the models. As can be seen from our results, empirical models will produce more accurate results for the datasets from which they were extracted than when the model is applied to previously unseen data.

Differences in effectiveness of the same techniques applied to two different datasets can also be explained by the nature of the input dataset as well as by the approach followed by the respective customs authority. To achieve the level of accuracies reported by Laporte the infractions present in the total population must be restricted to a relatively small number of the categories amongst which the observations are distributed. Our analyses of category behaviour as displayed in section 3.2 clearly shows that for our dataset, and for all category variables, the incidence of infractions is distributed across too many different

categories to allow 95 per cent of infractions to be found within as little as 20 per cent of the population. Even combining the different category variables as inputs into a neural network or classification tree cannot reach the same level of accuracy as was possible for the Laporte dataset.

Another factor that will influence the results is the fraction of consignments that were subjected to physical inspections before infractions were declared. In the case of SARS Customs, who already employs a relatively sophisticated risk management system, only a small fraction of consignments (about 2%) are physically inspected. In the case of Senegal, as with most other African countries, no systematic risk management approach was used at the time when Laporte performed his experiments. It is quite likely that they employed a much higher inspection rate, as it is common in many African countries to inspect virtually 100 per cent of all import consumer goods. If the current system used by SARS Customs was not effective in selecting those consignments for inspections that did include infractions, then the available data set will not include the majority of infractions actually present in the cargo processed. In such a case it will be impossible for an empirical model, extracted from data with such obvious limitations, to perform as well as in a case where almost all consignments were physically inspected and as a result more of the infractions that were present were found and thus represented in the data.

While the available data set may not necessarily accurately reflect the true presence of infractions, it does accurately reflect the decisions made by customs. A more realistic test for the techniques that we employed is therefore to determine how accurately they can predict customs decisions based on prior data. Table 9 shows that by selecting 16–17 per cent of consignments, 93–94 per cent of customs decisions to stop can be predicted out-of-training-sample. This provides evidence that the techniques that we applied work well on datasets that contain the required level of consistency in input–output relationships.

7. Conclusions and recommendations

Against the background of the previous sections, we can reach the following set of conclusions and recommendation, by addressing each of the stated research questions:

1. *Capability of identified input factors to predict Customs outcomes:* The results displayed in Table 5 provide a quantified answer to this question. Consignor identify was clearly the individual explanatory variable that contains the most prediction ability; in all likelihood it plays an important role in decisions taken by this customs authority.

2. *Combination of input factors that provide the best risk-prediction capability:* Apart from consignor, the other inputs that made a significant contribution include country of origin, transport mode (which in the case of our dataset is closely related to customs office as the two largest customs offices are served by primarily one transport mode each) and CPC code. Increasing the number of explanatory variables to a maximum of 9 did not appreciably improve prediction ability.

3. *Most effective empirical modelling technique:* Regression models performed slightly better than the simplistic techniques, with neural networks slightly outperforming regression models and classification trees providing the most satisfactory performance within the set of techniques that were investigated. As expected, the ability to correctly predict the outcome of the customs process was much superior to the ability to predict the incidence of infractions.

4. *Potential to improve the current South African Customs decision-making process:* From the results in sections 3 and 5, it is clear that the risk models can improve current processes as they all produced infraction incidence rates in the highest risk categories that are significantly higher than the incidence rates in the lowest risk categories. The level of improvement will depend on the minimum accuracy that is required for finding infractions. Based on our results, it is suspected that the primary limitation to extract accurate empirical models to be used in a risk engine is not the quality of the modelling techniques but the quality of the infraction incidence data that was generated by historical customs processes.

5. *Methodology to be followed:* The processes as described in section 4 are all applicable. If a technique is required that allows accurate control over the fraction consignments selected for inspections, then neural networks combined with setting an optimal risk score threshold value will be the most suitable; if a technique is required that allows the selection of the smallest fraction of consignments for a reasonably high selection accuracy then classification trees may be the best choice. In order to ensure that the models that are extracted remain current, the proposed methods of model extraction should be repeated on a regular basis, as the evidence shows that some of the underlying relationships change significantly over time.

Future work will focus on specific vertical markets that are severely impacted by customs operations. This work will involve extending the set of data fields extracted and measuring time delays in the process across the entire value chain, rather than focussing on the customs process only. This should indicate in which part of the value chain the most improvements can be achieved based on the ability to predict and detect eventualities at an early stage.

References

Baldwin, R. (2016). *The great convergence: information technology and the new globalisation*. Cambridge: Harvard University Press.

Bishop, C. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Creskoff, S. (2016). *What you need to know to go global: A guide to international trade transactions*. Archway Publishing.

Davaa, T., & Namsrai, B. (2015). Ways to modernise customs risk management in Mongolia. *World Customs Journal, 9*(2), 24–37.

Finger, M., Hintsa, J., Mannisto, T., Hameri, A. P., Thibedeau, C., Sahlstedt, J., & Tsikolenko, V. (2010). *Customs risk management: a survey of 24 customs administrations*. World Customs Organisation (WCO) Working Paper, EPFL-WORKING-173319.

Hoffman, A. J., Grater, S., Schaap, A., Maree, J., & Bhero, E. (2016). A simulation approach to reconciling customs and trade risk associated with cross-border freight movements. *South African Journal of Industrial Engineering, 27*(3), 251–264.

Khan, O. & Zsidisin, G. A. (2012). *Handbook for supply chain risk management: Case studies, effective practices, and emerging trends.* Fort Lauderdale, USA: J. Ross Publishing.

Komarov, O. V. (2016). Risk management systems in Customs: the Ukrainian context. *World Customs Journal, 10*(1), 35–44.

Laporte, B. (2011). Risk management systems: using data mining in developing countries' customs administration. *World Customs Journal, 5*(1),17–28.

Manners-Bell, J., Cullen, T., & Roberson, C. (2014). *Logistics and supply chains in emerging markets.* The Chartered Institute of Logistics and Transport. London, UK: Kogan Page Limited.

Manners-Bell, J. (2017). *Supply chain risk management*, 2nd ed. London, UK: Kogan Page Limited.

Prokop, D. (2017). *Global supply chain security and management: Appraising programs, preventing crimes*. Butterworth-Heinemann.

SARS. (2010). *Customs modernisation: Moving into the future. Trader pocket guide*. Retrieved from http://www.saaffkzn.co.za/web/webroot/files/articles/1294/60983_TRADERSPOCKETGUIDEv4.pdf

SARS. (2017). *Customs modernisation.* Retrieved from http://www.sars.gov.za/ClientSegments/Customs-Excise/AboutCustoms/Pages/Modernisation.aspx

Truel, C. (2010). *A short guide to customs risk*. Surrey, UK: Gower Publishing Limited.

United Nations Economic Commission for Africa (UNECA). (2013). *Trade facilitation from an African perspective.* Addis Ababa: UNECA.

United States Agency for International Development (USAID). (2015). *LEAP-II trade intervention agenda SA final report.* Washington, DC: USAID.

Viner, J. (1950). *The customs union*. New York: Carnegie Endowment for International Peace.

Widdowson, D. (2007). The changing role of customs: evolution or revolution? *World Customs Journal*, *1*(1), 31–37.

World Customs Organisation (WCO). (2015). *SAFE Framework for Standards to secure and facilitate global trade.* Retrieved from http://www.wcoomd.org/-/media/wco/public/global/pdf/topics/facilitation/instruments-and-tools/tools/safe-package/safe2015_e_final.pdf?la=en

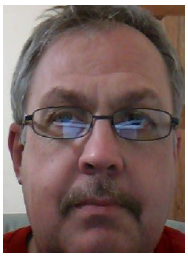
Alwyn Hoffmann

Alwyn Hoffman has Masters in Engineering, MBA and PhD from the University of Pretoria (UP). He has worked in the South African high technology electronics industry (1985–1994 and 2001–2008) and at Northwest University (19955–2001 and since 2009). Alwyn's recent research work focuses on the application of new technologies in the fields of transport and logistics, particularly transport corridors and has worked with many international organisations, including the World Bank and regional organisations, such as SADC (Southern Africa Development Community) and the EAC (Eastern African Community). This work is mainly aimed at establishing permanent corridor performance monitoring systems and promoting the concept of Green lanes for Authorised Economic Operators.

Sonja Grater

Sonja Grater is an associate professor in the School of Economics at the North – West University in South Africa. Before joining the university, she gained practical experience in the freight forwarding industry. She lectures undergraduate and post-graduate courses in the field of international trade. Her research specialisation is trade in services and trade facilitation and she has authored several publications in this field. She has also been invited to participate in several research projects for organisations such as the Department of Trade and Industry (dti) in South Africa, trade promotion organisations such as Wesgro and TIKZN, and private sector organisations such as the South African Association of Freight Forwarders (SAAFF). She has also undertaken projects for the International Centre for Trade and Sustainable Development (ICTSD), Department for International Development UK (DFID) and the World Trade Organisation (WTO).

Willem C Venter



Willem Christiaan Venter received a Masters of Engineering in electronic engineering in 1986 from the Potchefstroom University for CHE while working as a design engineer for Lektratek in Potchefstroom, South Africa. He received his PhD in digital signal processing from Iowa State University, Iowa, USA in 1989. He is a professor at the North-West University in South Africa. His areas of expertise include digital signal processing, digital image processing and software engineering.

Juanita Maree

Juanita Maree is Executive Director of Savino Del Bene South Africa (SDBSA), which is a multi-billion-rand international enterprise with 70 000m2 of warehousing space and some 421 employees in the South African operation. Juanita holds a Bachelor of Commerce and a Masters of Commerce. She is a passionate leader in the logistics field and is active in many Customs-related organisations, including the World Customs Organisation (WCO) Private Sector Consultative Group (PSCG), the International Chamber of Commerce (ICC), South Africa Association of Freight Forwarders (SAAFF), the Southern Africa Customs Union (SACU) and Business Unity South Africa (BUSA).

David Liebenberg

David Liebenberg is a Customs and Trade Consultant in South Africa. Before opening his own Consulting Company in 2014 he worked for South African Customs in Durban (1990–1996) and then the freight forwarding industry. David has also served on the South African Freight Forwarders Association Board as a Director of Customs. In his current capacity as a consultant he provides services to a number of global companies that include importers, exporters and freight forwarding companies. He also provides training at a leadership level and guest lectures at the North-West University in South Africa.