cebra
Centre of Excellence for
Biosecurity Risk Analysis

CEBRA 1401C/D: AIMS & SAC Text Mining, Stage One

CEBRA 1401C/D

Matthew Chisholm, CEBRA, University of Melbourne
Aidan Lyon, CEBRA, University of Melbourne
Lloyd Grant, Compliance Division, Australian Department of Agriculture
Tarik Zaman, Compliance Division, Australian Department of Agriculture

August 7, 2015

Contents

1

Executive Summary

We have conducted a preliminary investigation of the free text data held in the AIMS and SAC databases. Much of this data contains information that cannot be obtained from the structured fields in the databases, if at all. We have identified the following benefits to using text mining techniques.

1. Identifying direction failures. The result of each direction applied to a consignment is recorded in one drop-list field in AIMS, but the result values available don't always indicate unambiguously whether the result is a pass or fail. For example, the result *Admin Food Final* tells us nothing about pass/fail status, but an associated free text comment may say *cadmium detected*, indicating a failure. The free text can be analysed to categorise ambiguous directions results as pass or fail, which then allows non-compliance rates and performance indicators to be more accurately estimated.

2. Identifying reasons for direction failure. The direction result values don't always give any insight into the *reason* for a failure result, but this is often recorded by officers in several free text comments fields. For example, the free text alongside a failed fumigation direction could give reasons such as non-compliant certificates, incorrect treatment rates, unslashed plastic wrap etc. These free text fields can be analysed to inform tactical or strategic measures for targeting and reducing future non-compliance, be it by brokers, importers, suppliers, third-party treatment providers, or third-party certifying bodies.

3. Identifying incorrect tariff classifications. Preliminary analysis of the free text goods descriptions indicates that the tariff code selected by the broker is not always correct for the goods described — sometimes as high as 45%. Such inaccuracy may be accidental or deliberate, and may be indicative of other types of non-compliance. Since Agriculture's profiles operate largely on the tariff code, misclassifications can affect the effectiveness of targeting. Analysing the free text can identify these misclassifications, which can then inform corrective action such as issuing directions, providing targeted feedback or training to brokers, re-evaluating broker accreditation etc.

4. Identifying problematic Import Permits. The Import Permit numbers for a consignment are often recorded in various formats in a free text field, sometimes with comments about problems such as expired or invalid permits. These numbers and issues can be extracted from the free text and analysed in conjunction with direction results to identify permit holders who don't comply with permit conditions, or who attempt to use permits fraudulently. This can facilitate non-compliance action, and inform fit-and-proper character tests when assessing permit applications.

5. Identifying refrigerated containers ('reefers'). Officers often (but inconsistently) make notes

in one of several free text fields that indicate that the container is refrigerated. This information is only otherwise available by making a data request to Customs. The use of a refrigerated container for goods that don't normally require refrigeration could be indicative of non-compliance.

6. The potential to identify risk factors in SAC consignments. Targeting goods of concern and non-compliance in SAC consignments is limited by the lack of structured data in SAC declarations; most fields in SAC forms are free text. Text mining can identify terms associated with goods of concern, that can then be used to improve profiling. Improved profiling should reduce the SAC assessment workload of inspectorate staff.

7. Informing enhancements to the AIMS system. The shortcomings of the direction results field described above could be addressed by redesigning the field to unambiguously indicate pass/fail status and the failure reason. Analysis of the free text will inform such a redesign, which will make future analysis much simpler. In addition, many free text comments are variations on a theme, which are either auto-populated by AIMS, copied and pasted from regional cheat sheets, or manually typed by officers. Such comments could be standardised so that AIMS can auto-populate all of them based on the attributes of an entry, or by the officer simply clicking a radio button. Finally, some types of information are inconsistently recorded across multiple free text fields, and other fields are used very rarely and inconsistently. This makes data entry and analysis unnecessarily inefficient and prone to error. A more complete review of field usage can inform changes to the structure of AIMS which should improve the efficiency and accuracy of data entry and reporting.

2

Introduction

Within the Department of Agriculture (Agriculture), the Compliance Division is responsible for enforcing biosecurity import requirements at Australia's international borders. This includes the assessment, screening and inspection of passengers, mail articles and cargo consignments in order to intercept biosecurity risk material (BRM). The scale of this task, and the imperfect nature of border processes, together make it impossible to capture all arriving BRM.

For imported cargo, Agriculture relies on the information provided by importers and brokers when declaring the consignment to the Australian Customs and Border Protection Service (Customs). The information requirements for these declarations depend on the declared value of the goods. Full Import Declarations (FIDs) are required when the value of the goods is $1000 or more, and Self-Assessed Clearances (SACs) when less than $1000. FIDs require more information than SACs, and of better quality.

Declarations are generally provided in electronic form to the ICS database, where they are assessed against profiling rules that indicate whether biosecurity risk material (BRM) is likely to be present. FIDs of concern are referred directly to Agriculture's AIMS database, and SACs of concern are referred to its SAC database.

Since SACs are typically not subject to GST, SAC data quality is not a Customs priority. As a result, most SAC data is unvalidated free text. Agriculture staff *manually* assess all SACs referred from ICS to decide whether further action is required, such as presentation of documents or physical inspection. Consignments requiring further action may have those actions recorded directly in SAC, or they may be 'upgraded' into AIMS. Upgraded SACs have all subsequent processing recorded in AIMS, and not in the SAC system.

Agriculture staff manually process all upgraded SACs and directly-referred FIDs in AIMS, where they apply directions to goods based on the actions required, such as inspection or treatment. Much of the information critical to supporting risk-based intervention on both FIDs and SACs, such as the reason for direction failure, is recorded in various free text fields in AIMS by Agriculture staff. This is further complicated by the fact that information on any pests and diseases intercepted during inspections is (optionally) recorded in yet another database, Incidents. We note that the department is planning to implement guidelines on the use of AIMS which includes when and how to make free text comments, and is also planning to make the use of Incidents mandatory. These measures, if implemented successfully, should improve the utility and value of the data captured for future consignments.

Approximately 27 million SACs and 3.7 million FIDs were processed through the ICS in FY 2013–14, of which 620 000 SACs (2.3%) and 430 000 FIDs (11.6%) were referred to Agriculture for further assessment.

3

Background

This project is a synthesis of two approved scoping study projects, *Data profiling for border compliance*, which focused on free text data in SAC, and *Entry process outcomes in AIMS*, which focused on free text data in AIMS. Since the core analytical work of these projects is the same—text mining—the project leaders decided to combine them into one.

The combined project document originally proposed trialling text mining techniques that might be suitable to the task. Based on feedback from the then project sponsor, this was changed into a two-stage approach. In the first stage of the project, a small snapshot of data was to be examined to assess the quality and utility of free text information held in the AIMS and SAC databases, particularly information that is not readily available elsewhere. Based on the findings of this stage, the project sponsor may then decide whether the data are of sufficient value to warrant proceeding to the second stage: the formal trial of text mining techniques.

4

Purpose

This report constitutes the deliverable for Stage One of the combined project. It describes

- some characteristics of the free text data held in the AIMS and SAC databases
- the prospective value of the data to the business, such as the ability to detect goods of concern via improved profiles, and
- the retrospective value of the data, such as enhanced ability/accuracy in analysing historical data.

This report also identifies some factors in SAC data that may indicate different levels of biosecurity risk, and some common terms used by officers when populating the free text fields of AIMS. This satisfies the stated aims of the two original projects.

5

Previous work

A three-month data set of failed Cargo Compliance Verification (CCV) AIMS entries was manually categorised by Cairns staff by failure reason, as best as could be assessed from the information recorded in the entry. The data were from the period Jul–Sep 2013 inclusive. This methods and results of this analysis would have been instructive to our project, but were not considered because they were not provided to us.

Two recent ANAO reports for Customs dealt with data mining, namely *Processing and Risk Assessing Incoming International Air Passengers*, and *Risk Management in the Processing of Sea and Air Cargo Imports*.

Some time ago, the then Border Compliance Division investigated the free text data from AIMS entries containing a tailgate direction, looking to categorise them by failure reason. No data or results from this exercise have been provided.

Information on why a consignment failed a direction (document assessment, physical inspection, treatment etc.) may be encoded in the free text AIMS comments, and/or in a related Incidents record. *ACERA Project 1001B Study J* partly included a combined analysis of datasets from these databases, but it was found that Incidents records can't always be easily linked back to their associated AIMS entry.

The Interim Inspector-General of Biosecurity (IIGB) report on *The effectiveness of controls for imported uncooked, cooked and cured pig meat* notes the following (p. 8):

> 17. For reporting and analytical purposes, [Agriculture] is unable to differentiate accurately between uncooked, cooked and cured pig meat import data. The distinction can only be made if one has a copy of every individual entries from [AIMS], which would be a large administrative exercise. The IIGB noted there are inaccuracies as the data provided is derived from tariff codes used in entry documentation. These tariff codes may be incorrectly or inadvertently provided by the importer/broker. For example cured pig meat products (such as Parma-type or Serrano hams) may have a tariff code description as 'Meat of swine, fresh, chilled or frozen' and [Agriculture] records will capture this as uncooked pig meat. This was evident in the data received.
>
> ... AIMS determines whether imported pig meat is uncooked, cooked or cured through the profiling questions, which are answered by the importer/broker. These profiling questions are recorded in each AIMS entry. [Agriculture] is unable to extract this information into a consolidated report to detail and account for the volumes of uncooked, cooked and cured pig meat imports.

While the IIGB review did not analyse free text *per se*, we note that text mining techniques would have been useful had such an analysis been done; they are well-suited to differentiating between uncooked, cooked and cured pig meat across large datasets.

6

Methods

For this report we analysed a snapshot of AIMS data to characterise the free text fields. This dataset was the same that had previously been provided to ACERA for project 1101C-1 *Plant quarantine inspection and auditing across the biosecurity continuum*. The data were from the four years 2007–2011, for 34 variables (i.e. fields in AIMS) across 584 000 rows where the goods had been classified by the broker under Tariff Chapter 7 (Edible vegetables and certain roots and tubers).

Data snapshots were also provided by Agriculture for approximately 500 randomly selected entries in each of the following groups:

- FID-based AIMS records
- Screened-free SAC records
- Upgraded SAC records, along with their corresponding AIMS records
- CCV AIMS records.

Exploratory data analysis was conducted using R and Python. Analysis relied heavily on the use of regular expressions, which are character sequences used to define patterns in strings of text. They can account for spelling variations, non-printing characters, trailing spaces and other irregularities common in free text data. For example, we can find all instances of 'non-compliant', 'non-compliant' and 'non compliant' in a given text by using the regular expression `non-?compliant`. Fuzzy matching was also used as part of our analysis. This technique compares text strings to determine their similarity in terms of an edit distance, such as the Levenshtein distance.

When used in combinations, with or without fuzzy matching, regular expressions can be a powerful tool for identifying patterns that would otherwise be difficult to isolate, due to mis-spellings, typographical errors and user-to-user variations. They are supported in many programming languages, including those used by data analysts in Agriculture.

In interpreting the data, reference was made to the procedural material provided by Agriculture, listed at Appendix A.

7

Results

7.1 AIMS data for Tariff Chapter 7

All free text fields in the data sample are described below, listed in descending order of non-null usage. All fields except `Goodsdescription` are populated by staff.

7.1.1 Goods Description

`Goodsdescription` was non-null in 100% of rows (584k). There was a lot of variation, with 10k unique values observed.

7.1.2 Standard Comments

`Standard.Comments` was non-null in 52% of rows (303k). There was a lot of variation, with 85k unique values observed.

- 48% of rows null
- 0.36%: *Inspect for infestation/contamination.*
- 0.15%: *Quarantine direction only. Entry is subject to Imported Food Inspection Scheme and must be presented for processing.*
- 0.12%: *Inspect as per work procedure.*
- 0.12%: ¶[1]
- 0.09%: *ALL INFORMATION TO BE RECORDED IN FIELD RECORD COMMENTS AS PER EXAMPLE¶¶*

Some terms appear in multiple ways, partly due to typographical variations, and partly due to additional text. The following standardised comments were most common:

- 0.5% of rows, but in 68 different ways: *Outstanding Entry*
- 0.1%: in 988 different ways: *Inspect as per work procedure*

[1] The pilcrow character ¶ indicates a line feed or new line control character.

7.1.3 Direction Comments

`Direction.Comments` was non-null in 49% of rows (287k). There was a lot of variation, with 41k unique values. Many terms appear to be from AIMS shortcuts or copy/pastes from local cheat sheets:

- 51% of rows blank
- 1.6%: *Released on Documents¶¶Invoice and BOL/ AWB Sighted*
- 1.15%: *Phytosanitary Cert OK. Original documents to be presented at inspection point*
- 1.15%: *Seal prior to movement to Crewe Place for inspect. Vero orig docs & check packing secure. Goods arriving at Menzies, Patrick & AAE to be sealed at that depot only by phoning AQIS 0418 476464. Cut-off time for sealing is 3pm or overtime must be booked*
- 1.15%: *Goods to be unpacked and inspected at designated premise where goods are to remain consignment intact.*
- 1.10%: *(A shortcut was used to process this entry)¶ All documents presented were acceptable.*

Many terms turn up in multiple ways, partly due to typographical variations, and partly due to additional text. The following standardised values were most common:

- 2.6%, in 1323 different ways: *Original phyto/docs to be presented at inspect*
- 2.7%, in 116 different ways: *Released on docs*
- 0.7%, in 12 different ways: *Category not to be finalised until outstanding directions completed*
- 3.2%, in 11 different ways: *All docs presented acceptable*
- 1.8%, in 9 different ways: *Broker made a major amendment, re-assess risk before closing*

Other points of note for the terms in this field:

- Comments against fumigation directions sometimes include time in/out, actual rate, and officer initials.
- Comments against the `Direction` values *Check BSE Cert* and *Check FC Doco* often include a number; these are reportedly certificate numbers.
- Some `Direction.Results` values are ambiguous as to whether they constitute a Pass or a Fail, but the `Direction.Comments` may help in disambiguating. For example, in row 487 402, the `Direction.Results` value is *Admin Food Final*, which tells us nothing about Pass/Fail status, but the `Direction.Comments` value is *Cadmium detected*, which would suggest a Fail.
- Some terms have been truncated. It was later discovered that there is a character limit to this field, and that the overflow text gets saved into a second field, which was not provided in the dataset analysed.

7.1.4 Permit Comment

`Permitcomment` was non-null in 46% of rows (270k). Almost all terms are simply the Import Permit numbers in various formats, sometimes with multiple permit numbers, some truncated. They sometimes mention which line/goods the IP number applies to, and a few say *not needed*, *expired*, *application pending* etc.

7.1.5 Phytosanitary Comment

`Phytosanitarycomment` was non-null in 38% of rows (220k). Points of note:

- Many contain 15-digit numbers, which we believe to be CIQ[2] certificate numbers.
- Many explicitly mention *CIQ*, but the associated `Country` values for these entries are rarely *China*.
- Many are variations of *nil khapra*, *khapra free*, or *nil khapra dec ok*.

7.1.6 Manufacturer's Comment

`Manufacturerscomment` was non-null in 33% of rows (192k). Of the non-null values:

- 16% mention *-18℃*, but in many different ways—it appears in 1900 (20%) of 9400 unique values
- 6.1% say *as per permit*, in 3.3% of unique values
- 5.4% mention *blanched & frozen*, in 4.4% of unique values
- 4.4% say *see comment*, in 1.6% of unique values

7.1.7 Field Comments

`Field.Comments` was non-null in 32% of rows (186k). There was a lot of variation, with 47k unique values. This field appears to be notes by or for officers performing inspections in the field, or by/for officers making appointments. Several values include email addresses, reportedly for the depot contact.

7.1.8 Packing Comment

`Packingcomment` was non-null in 28% of rows (164k). Almost all values were some variant of *unlined*, *reefer*, *frozen*, or *airfreight*.

7.1.9 Fumigation Comment

`Fumigationcomment` was non-null in 13% of rows (76k). Most give detail of the treatment applied, sometimes with dates and certificate numbers. Some state the treatment provider or AFAS[3] number. Some also mention problems, e.g. *wrong rate*, *no temp or duration* etc.

7.1.10 Health Comment

`Healthcomment` was non-null in 5.8% of rows (34k). Most values appear to be certificate numbers, but some say *cert not ok* or similar.

7.1.11 NZMAF Comment

`Nzmafcomment` was non-null in 5.5% of rows (32k). Most values appear to be certificate numbers.

[2]China Inspection and Quarantine Organization
[3]Australian Fumigation Accreditation Scheme

7.1.12 Cleanliness Comment

`Cleanlinesscomment` was non-null in 5% of rows (29k). Usage appears almost the same as `Packingcomment`, plus some other terms, such as the type of container, LCL/FCL, *illegible*, and *fringe rural*.

7.1.13 Timber Packing Comment

`Timberpackingcomment` was non-null in 4.3% of rows (25k). The vast majority say *ISPM 15 ok* or similar. Others indicate freezing, reefer, fumo, nil timber, or use of ply/inka[4]/chipboard pallets in various formats.

7.1.14 Treatment Comment

`Treatmentcomment` was non-null in 4.2% of rows (25k). About half mention refrigeration, and the remainder mostly say *not sighted*. Some give detail of treatments applied, others mention problems with certs.

7.1.15 Rarely used fields

- `Plasticwrapcomment` was non-null in 0.6% of rows (3526).
- `Strawcomment` was non-null in 0.3% of rows (1821). Most values indicate that a certificate was sighted, ok or not, or give details of certificate. Some mention treatment rates, use of reefers etc.
- `Barkcomment` was non-null in 0.1% of rows (681). Values indicate whether bark present/declared or not, timber used or not, ISPM, or reefer.
- `Fieldtestedcomment` only used in 0.03% of rows (183). Usage appears to be by mistake, noting that the data analysed were only for Tariff Chapter 7; this field is reportedly used for new machinery that has been field-tested, which falls outside Tariff Chapter 7.
- `Mosquitocomment` was only used in 5 rows. Usage appears to be by mistake, noting that the data analysed were only for Tariff Chapter 7; this field is reportedly used for tyres, which fall outside Tariff Chapter 7.

7.2 SAC data

The main field in SAC entries used for identifying the type of goods is the `Goodsdescription`. Most other fields are timedatestamps, airwaybill numbers and address fields. We focus our attention on patterns detected in the `Goodsdescription` field.

7.2.1 Goods description — screened-free records

The dataset for screened-free SAC records consisted of 1685 rows of data across 576 SAC records.

The most common types of commodity were body-building supplements and e-cigarettes and their associated liquids. Using preliminary regular expressions, we found that body-building supplements account for 649 (38.5%) of the 1684 rows, and that e-cigarettes and their liquids

[4] `http://www.inka-palet.com/`

account for 192 (11.4%) of the rows. These two regular expressions, plus another for `flavou?r`, jointly identify 901 (53.5%) of the rows.

7.2.2 Goods description — records upgraded to AIMS

The dataset consisted of 12 885 rows of data across 500 SAC records. The large number of rows was mainly due to multiple directions being applied to each line of goods.

The most common types of commodities were

- biological substances, including DNA and tissue samples – 179 rows (35.8 %)
- laboratory reagents – 42 rows (8.4%)
- coffee – 31 rows (6.2%)
- khat leaves – 24 rows (4.8%).

Collectively these account for 276 rows (55.2%) in the data.

8

Discussion

8.1 AIMS data fields

8.1.1 Goods Description *vs* Tariff

We developed a draft script to determine whether the broker used the correct tariff code for the goods declared. Preliminary analysis of goods declared as *potatoes*, *garlic* and *asparagus* indicate that the broker assigned the correct tariff code in 76%, 68% and 55% of cases respectively. It is possible that accurate classification may be correlated with the value of the goods, the practice of 'tariff shopping', or the expectation of further Agriculture interventions. Note that the tariff code is used by Agriculture's profiles in ICS to determine whether additional CP and CRA questions will be asked of the broker on lodgment, so the accuracy of the data provided by the broker is critical to these profiles working correctly.

8.1.2 Import Permits

We developed a draft script to extract standardised Import Permit numbers from all fields that appear to contain them, namely `Goodsdescription`, `Standard.Comments`, and `Permitcomment`. This may be of use when analysing compliance of permit-holders. The script could be easily adapted to extract other certificate numbers.

In the `Permitcomment` field, terms like *expired*, *application pending*, *wrong goods* sometimes appear. These would be worth further investigation to identify non-compliant permit holders, or importers presenting fraudulent permits.

8.1.3 Certification failures

Using the `Plasticwrapcomment` field, we developed scripts to isolate the following pass/fail reasons for fumigation certificates that were assessed:

- Fumigation done prior to plastic wrap (i.e. pass)
- No plastic wrap used at all (i.e. pass)
- Wrap not slashed before fumigation (i.e. fail).

We may find more reasons among the text, and may be able to correlate with `Direction.Result` values.

`FumigationComment` often identifies the Australian Fumigation Accreditation Scheme (AFAS) provider and/or the type of problem with the fumigation certificate. This should be of use in identifying failures by AFAS providers.

`TimberPackingComment` usually indicates ISPM 15 compliance or the use of non-timber materials. This field should be useful in determining the rate of ISPM 15 failures for specific countries, suppliers etc.

Some `TreatmentComment`, `StrawComment` and `BarkComment` values indicate problems with or lack of certificates. These will probably be of use in identifying failures by certifying bodies and/or fraudulent documents.

`Phytosanitarycomment` may also be used to identify failed certificates, but the variants of *nil khapra*, *khapra free*, or *nil khapra dec ok* are ambiguous in meaning. For example, *nil khapra* could be interpreted either as 'certificate certifying nil khapra', or as 'nil certificate regarding khapra'.

8.1.4 Reefers

We developed a draft script to identify whether the container used for the goods was refrigerated, or a 'reefer'. This is reportedly of interest, since there is no other way to identify reefers, except by sending a data request to Customs. The script currently uses text in the `Packingcomment`, but could also draw on text in other fields, particularly `Goodsdescription`, `Manufacturerscomment`, `Treatmentcomment` and `Timberpackingcomment`.

8.1.5 Refactoring AIMS

Based on the uses of the various comments fields, many of them could be replaced with a few properly structured fields. For example, all certificate types could be listed in AIMS, with check boxes against them to indicate *ok*, *not ok*, or *n/a*. A *not ok* value may trigger a separate set of check boxes to indicate the reason, particular to the certificate type.

We also found that direction results values aren't always being used for the purpose intended. For example, a failure due to an unacceptable fumigation certificate might be recorded as *fumo not ok* instead of *doco not ok* as it should be; the true reason for failure can only be gleaned from the comments, if at all.

8.2 SAC data fields

The SAC dataset analysed was only a very small snapshot of the SAC records that are processed by Agriculture staff. We have found that regular expressions are a fast and powerful way to automatically identify patterns of text in the `Goodsdescription`. The regular expressions we developed are only preliminary, and further refinement will undoubtedly improve their accuracy and coverage.

8.3 Potential benefits of text mining

By applying text mining algorithms to AIMS, the following benefits may be realised:

- Validate tariff codes against goods descriptions, so that entries can be more accurately categorised for subsequent analysis. Mis-codings by brokers may also prove to be an indicator of non-compliance, which, when confirmed, could be used to assist in auditing broker performance.
- Automatically join Incidents records to their AIMS entries using the entry number, and validate this join with other fields such as date, importer etc. Where validation is imperfect, provide a list of candidate matches.
- If coupled with OCR, could extract information from scanned phytosanitary certificates, packing certificates etc., and populate relevant AIMS fields automatically.
- Find new patterns in goods descriptions etc. that can improve ICS profiles.

By applying text mining algorithms to SAC, the following benefits may be realised:

- Automatically interpret the free text goods descriptions to determine if an AIMS upgrade is warranted or if the goods can be released. At present this task is done manually by inspectorate staff. In a hypothetical implementation, an algorithm should be able to categorise the majority of SACs as definitely of interest or definitely not of interest. The small number of residual 'not sure' entries would be referred to an officer for decision, as would a random sample of 'sure' entries for confirmation. Modern algorithms self-adapt based on this kind of human feedback, so should continually improve both specificity and sensitivity over time.
- Where an upgrade to AIMS is needed, automatically determine the most appropriate tariff code for the goods description, and automatically standardise the consignee name, consignor name and destination address against known names and addresses. This task is also performed manually at present, and is prone to error.
- Find new patterns in goods descriptions etc. that can improve ICS profiles.

8.4 Notes for Stage Two

Stage two of the project, if approved by the project sponsor, will determine the most appropriate text mining methods for extracting and interpreting unstructured data in the AIMS and SAC fields of interest, as identified in Stage One. This work will seek to satisfy the following elements of the two original projects:

- enable the analysis of SAC data to differentiate levels of biosecurity risk
- guide targeting in the SAC system to improve efficiency in resource allocation
- improve the recording of assessment and inspection results in AIMS, such as by replacing commonly-used free text with drop-boxes, radio buttons etc.

Once the first set of key terms are developed, different types of mining algorithms using various parameter values would be run over the first snapshot of data. The best models produced will then be refined iteratively over subsequent datasets. The end product of Stage Two will include a prototype model/algorithm for identifying records of interest based on the patterns identified.

The details of the algorithms and their implementation will depend on the context of the analysis. If analysing large amounts of aggregated historical data, the algorithms may be tailored to identifying patterns that may be of interest to the analyst for developing new profiles or for informing system enhancements. However, if analysing live ICS data for profile matches, the algorithms will be tailored to finding pre-identified patterns, and computational efficiency will be critical. Note that we have not determined whether ICS can support regex-based classification of consignments.

Since the data being analysed are subject to continuous change, so too should the algorithms be adaptable. The specificity and sensitivity of the algorithms with regard to most recent data will need to be validated by subject matter experts. Machine learning techniques can use feedback from these experts so that algorithm performance is maintained.

Some more specific points that may be considered during Stage Two include the following.

- The patterns we've found in the AIMS data are based only on the sample dataset, which was restricted to lines for Tariff Chapter 7. An analysis of an unfiltered dataset is recommended, as it is likely to produce different results, and provide additional insight.
- In addition to the free text discussed in Section 8.1.4, refrigerated containers may also be identified from the Tariff code. This assumes that the code is correct, which can in turn be validated from the Goods Description as discussed in Section 8.1.1.
- While free text analysis may indicate an incorrect tariff code, the degree to which these errors adversely affect the profiling (e.g. CP and CRA questions) should be investigated.
- Further analysis of tariff codes may identify instances of misuse due to ambiguity, rather than any attempt at deception.
- Our draft scripts and regular expressions will need to be refined and tested for specificity and sensitivity. This was not possible due to the restricted data available to us.

Before implementing any text-based algorithms, Agriculture should determine whether the algorithms would perform better than human assessments of free text. Performance comparisons should include (but may not be limited to) sensitivity, specificity and cost. Agriculture should also consider the experiences of other regulatory organisations in implementing text-mining techniques.

Appendices

Appendix A

Procedural documents provided by Agriculture

In addition to the data, the following documents were provided to assist in interpreting the data.

- `Entry processing - understanding an AIMS entry.pdf`
- `Entry processing - document assessment.pdf`
- `Entry processing - recording outcomes of a document assessment and issuing directions in AIMS.pdf`
  - Explains free text that indicate why documentation failed assessment, options for brokers/importers, coded details of offshore treatments, extra information for subsequent officers, timeframes for directions, permit numbers and post-entry requirements. Mentions some structured fields whose info may be duplicated in free text. Also mentions that comments may be copy-pasted from the `Preview Comments` field and ICON entries.
- `Cargo Compliance Verification.pdf`
- `Entry processing - assessing Cargo Risk Analysis profiles in AIMS.pdf`
  - Describes the types of profiles in ICS. Also mentions free text comments used when reviewing profile-related directions, such as checking for fraudulent documents, cancelling directions etc.
- `Self Assessed Clearance (SAC) Declaration Processing - Training.pdf`
- `Self Assessed Clearance (SAC) Declaration Processing.pdf`
  - Full overview of the SAC clearance process done by the SAC NCC; looks the same as the Training document above, only without the screen shots.
  - Mentions use of some comments fields.
  - Mentions document assessments, inspection results, upgrades to AIMS, and seizures made directly on SACs without upgrade to AIMS.
- `Guideline - Introduction to Phytosanitary Certificates.pdf`
- `How to process plant product entries part-processed by the Q-ruler.pdf`
- `self_assessed_clearance_declarations.pdf`
- `ICS Customs Data Dictionary.mht`
- `AIMS User Guide.pdf`
  - Very comprehensive document, explaining the import process, fundamental terminology & acronyms, and the various fields in AIMS and their permitted values and

485 usage. Gives some info on which fields are auto-populated from ICS, SAC, QPR etc. Written primarily for frontline inspectors.

- `AIMS Quarantine Shortcuts and Directions Business Policy.pdf`
- `Creating an Incident.pdf`
  - Not only describes how to raise an Incidents entry, but also describes some of the 490 changes to the system made in the last redesign project. Describes the values available in various structured fields.
- `Integrated Cargo System (ICS) User Guide.pdf`
  - Description of the ICS and detailed instructions on how to use it to perform biosecurity functions.
  - 495 Table 4 (p. 33) has a neat list of groups within the Department and the functions they perform, e.g. SAC NCC in Sydney.
  - Chapter 2 gives an overview of the import process, systems used and how profiles work. Also mentions that the SAC NCC staff not only assess SACs automatically referred from ICS, but also SACs where the profile can't make a definitive decision.
- 500 `ICON Query user guide.pdf`
- `ICON Permits user guide.pdf`
- `ePermits user guide.pdf`
- `Quarantine Management of Imported Goods SOP.pdf`
  - Describes the legal and administrative framework for the quarantine management of 505 goods.