Artificial Intelligence in the Customs Selection System through Machine Learning (Sisam)[1]

Abstract

In this work we describe the artificial intelligence module of the Customs Selection System through Machine Learning (Sisam), today in use in all customs units of Brazil's Federal Revenue (RFB). We show that the deployment of this system was only possible due to the tackling of more than a dozen technical challenges and scientific innovations of direct interest of RFB. Doing that we characterize this process not as an example of the already acknowledged application of data mining techniques in RFB, but as an innovative case of technological research and development within the institution. We also show that Sisam brings real performance gains in the detection of several infractions during the course of customs clearance with special emphasis on the important and complex error in the classification of goods. We present Sisam's interface with customs officers and highlight the way by which it allows human and machine knowledge to be united fast enough for timely decision-making. We point the fact that Sisam is the first online artificial intelligence to be developed within RFB and the first with fully generalized use in its field of application. We also show that the interest in Sisam's technology is not restricted to its original goal having given origin to a reasonable future work list and being effectively applied in two other contexts both in the field of internal taxation.

1 Introduction

The Customs Selection System through Machine Learning (Sisam) is an artificial intelligence (RUSSEL; NORVING, 2004) which learns with the history if import declarations (DIs) and has the goal of helping Brazil's Federal Revenue to reduce the percentage of goods verified during customs clearance and thus reduce cost for Brazil's economy. At the same time it helps to reduce tax evasion and non-compliance with administrative requirements like failing to obtain consents from the ministries of health, agriculture or defense.

A reduction in the percentage of verification without precision gains would obviously lead to an increase in tax evasion and breach of administrative requirements, the opposite of the goal. Thus, it is desired to select better which import declarations will

1 This text is a translation of (JAMBEIRO, 2015), which was written to take part in RFB's creativity and innovation award in 2015.

be automatically cleared (green channel) and which will be distributed to an inspector for customs verification (yellow, red and gray channels).

Since RFB's Normative Instruction number 680/2006 the customs inspector is not obligated to verify the whole DI. He is only required to verify the part of the DI that motivated the selection of the DI for the yellow, red or gray channel and has the freedom to verify or not any other aspect of the DI. So, it is also important to help the customs inspectors to choose which goods they will effectively verify.

Sisam increases the precision of the selection of DIs for conference channels and also the choice of individual goods for verification. It analyzes every item of every addition of every DI and, for each of them, calculates the probability of the presence of several types of error. It also indicates possible correct values for the fields that have the errors and for each possible value calculates its probability and its consequences, both for taxes and for administrative requirements. With this, Sisam is able to estimate the relevance of every possible verification from the point of view of RFB and act both deciding automatically which verifications should by carried on and helping a customs officer[2] who happen to be responsible for making those decisions.

The history of import declarations of Siscomex (Brazilian Integrated Foreign Trade System) is accumulated since 1997 when it was activated and contains more than ten million DIs and a hundred million goods descriptions. In this database, the verified DIs (about 15% of the total) appear both in their original version and the version that was cleared by the customs inspector. Thus, it is possible to identify what changed from one version to the other and know what was wrong in the first version what generates great potential for the application of machine learning (MITCHELL,

2 Brazil's Federal Revenue is in charge both of Brazilian customs service and Brazilian internal revenue service. Brazil's Federal Revenue employees responsible for customs clearance and tax audit are the same and are in general called Fiscal-Auditors. In this text we call them customs officers or customs inspectors when they appear performing customs selection, verification or clearance.

1997).

However, the data in this history are complex and developing a precise mechanism to select goods for verification during customs clearance is not a simple task. Fiscal-Auditor Marcos Ferreira (2003a) analyzed this problem and identified the main obstacle to the statistical treatment of this base: the presence of high cardinality nominal attributes.

A high cardinality attribute is an attribute that may assume many different values. The Mercosur Common Nomenclature (NCM) code, which is Mercosurs' version of the Harmonic System (HS) code, is an attribute that may assume about ten thousand different values. The identifier of the importer takes tens of thousands of different values in Siscomex database. The countries involved in transactions with Brazil are around of 200. These attributes when combined, generate an exponential explosion that induces a problem that permeates the entire artificial intelligence, overfitting. In the presence of overfitting the artificial intelligence does well over the training cases, but very badly when tested over new cases, as if it were a person who memorized rather than understood a subject.

Marcos Ferreira made a wide bibliographical survey and concluded that the best available strategy to deal with the problem would be the use of linear methods, since they avoid the combinatorial explosion of the attributes. He selected the strategy recommended by Pearl (1988), the Noisy-Or. That research showed significant gains regarding the selection of DIs for conference and Ferreira was awarded with the second place in the Prize for creativity and innovation of the RFB (then called Schontag Award) in the year 2003 (FERREIRA, 2003b).

However, linear methods like the Noisy-OR have the deficiency of discarding non-linear interactions between the attributes. For a linear model, for example, if a certain importer has a high risk and a certain NCM is often erroneously declared then it is is inevitable to conclude that, when this importer appears together with this NCM, an operation whose risk is also high (usually higher) is taking place. In practice, there is a non-linear interaction between the attributes “importer” and “NCM” which, from time to time, makes this conclusion false. It is caused by the fact that when an importer is caught committing an error it tends to stop committing it, but may easily go on committing different errors. Doesn’t matter how strongly the result of verifications show that we now have a safe exception, a linear model cannot adapt to it. Insisting in mistakes that are obvious to customs officers jeopardizes the credibility of the system and therefore discourages its adoption. The lack of that adaptation also prevent improvements in the system's success rate.

The relevant non linear interactions among important import declarations attributes are many. Thus, since Marcos Ferreira work, developing a method capable of handling nonlinear interactions among nominal high cardinality attributes without incurring in overfitting became a technological advance of particular interest for RFB. Some years latter, that advance was achieved by Fiscal-Auditor Jorge Jambeiro Filho (2007a). That work was presented in the main conference of artificial intelligence, the International Joint Conference of Artificial Intelligence (IJCAI), that took place in India in 2007 (JAMBEIRO FILHO, 2007b) and published in the main artificial intelligence journal, the Journal of Machine Learning in 2008 (JAMBEIRO FILHO, 2008).

In spite of that, several important technical aspects for the effective delivery of a system like Sisam were not solved, among them:

- handling attributes of different types at the same time and non-linear interactions among them all, including nominal attributes like the importer identifier, hierarchical attributes like the NCM code, which involves concepts like chapters, headings and subheadings, continuous attributes like tax rates, weight and price and natural language attributes like the names of the foreign suppliers and manufacturers and, most importantly, the free description of goods;

- handling multiple unknown variables at once to deal, for example, with the fact that the declared tax regime, the declared legal base of the tax regime and the declared import duties may all diverge for their real values at once and they all have important consequences over each other;
- having resources to cut the hypothesis space to compensate the exponential explosion that emerges from the presence of multiple unknown variables and to minimize the losses for not considering every hypothesis;
- undertaking supervised and unsupervised learning at the same time to learn from verification results, but also from unverified import declarations and, with this, notice deviations from typical patterns which raise suspicions even if no infraction has yet been detected in a similar context;
- being able to learn with new import declarations without having to redo the training with old ones and therefore make daily updates possible, even after having been trained with millions of import declarations;
- being able to learn fast, in particular, be able to learn to stop infractions that repeat in very similar ways even from a single example;
- having an inference engine that admits interventions ad hoc to deal with peculiarities of the environment of RFB and impose structural constraints that avoid giving excessive degrees of freedom to statistical models and

consequently avoid increasing sensitivity to noise;

- being fast enough to take into account a history of millions of import declarations and produce answers in real time handling the constant flow of customs clearance;
- avoiding to learn with wrong behaviors, even if repetitive, not being induced to error by importers;
- not getting lost when dealing with goods descriptions that include more words associated with the business context that the merchandise itself;
- being able to handle mutant classes, since the legislation may determine, for example, that what was previously classified in a certain NCM code, now belongs to another, creating a problem that does not exist in traditional machine learning;
- not allowing importers to identify the system's behavior and learn how to stay bellow the system's radar;
- generating explanations that allow customs officers to add, to the decision process, knowledge that extrapolates the scope of Sisam's artificial intelligence and combine that knowledge with the system's conclusions.

Each of these requirements imposes restrictions on the architecture of the probabilistic inference engine. This causes the biggest challenge in building the artificial intelligence of Sisam to be dealing with multiple challenges at the same time.

It is important to notice that many machine learning tools, like, for example, Weka (WITTEN; FRANK, 1999), offer long lists of features, but that doesn't mean to have the ability of using all resources at once to solve a single problem.

Although being imperfect in the way it handles most of its requirements individually, Sisam meets them all at the same time. There was no solution for that, neither in the market nor in academia.

The importance of data mining technologies is growing in the world, and much is heard of “big data”, the expression that denotes the exploitation of giant databases that surfaced on the internet and within major institutions. RFB has already recognized the importance of the application of these technologies and has promoted events like the Seminary of Data mining and Artificial Intelligence that took place in Bauru in March 2015. In addition, it has been planning to acquire a robust data mining platform (currently in the process of preparation of bidding terms). Moreover, at least two works which have data mining at their core have already received RFB's award for creativity and innovation.

FERREIRA (2003b) investigated the application of various techniques of artificial intelligence to the selection of import declarations and, for that, indicated the use of Bayesian Networks with Noisy-Or gates and CARVALHO (2014) investigated methods, techniques and tools that provide fuzzy logic application to RFB's databases. Nonetheless, we are not aware of any work that claims to have created new technology in the field of data mining and artificial intelligence in the direct interest Brazil's Federal Revenue. In relation to that, Sisam is unique.

Furthermore, Sisam is the first artificial intelligence of generalized use within RFB. It is available online to all customs units of RFB, it is integrated to Siscomex and it process 100% of the import declarations registered in Brazil.

Besides that, the technology developed in this work does not have applicability only in the selection of goods for customs verification. Many future works are planed by

RFB customs coordination (COANA): supervision of export goods, supervision of postal and express consignments, supervision of the granting of licenses to operate in foreign trade, monitoring of customs transit and surveillance of accompanied luggage. In fact, during the development of Sisam's technology, two parallel applications were developed and made available through the Contágil system (FIGUEIREDO, 2008), both in the field of internal taxation. The two functions are the function of inexact names pairing used for comparisons of payrolls and the mechanism of detection of encoding errors in NCMs and CFOPs in invoices (MDECNF). This shows the generality and the importance of the research carried out.

Sisam puts the Brazilian Customs Office on the edge of technological development and arouses attention in other countries. Canada, for example has expressed interest in knowing the details of the system.

2 General view of the system

The best starting point to understand Sisam is its interface with the users. This interface was build within the Intelligent and Integrated Analyzer of Customs Transactions (Aniita) (COUTINHO, 2012), which was in use to select import declarations for verification and support to customs clearance before Sisam and which presents in a single place information from different systems of RFB.

The backbone of Sisam's interface is an interactive spreadsheet with colorful highlights. This spreadsheet may be configured by the user to include any field of the import declaration, Aniita alerts and Sisam results that happen not to be included by default. In Figure 1, we show Sisam's spreadsheet in its default configuration.

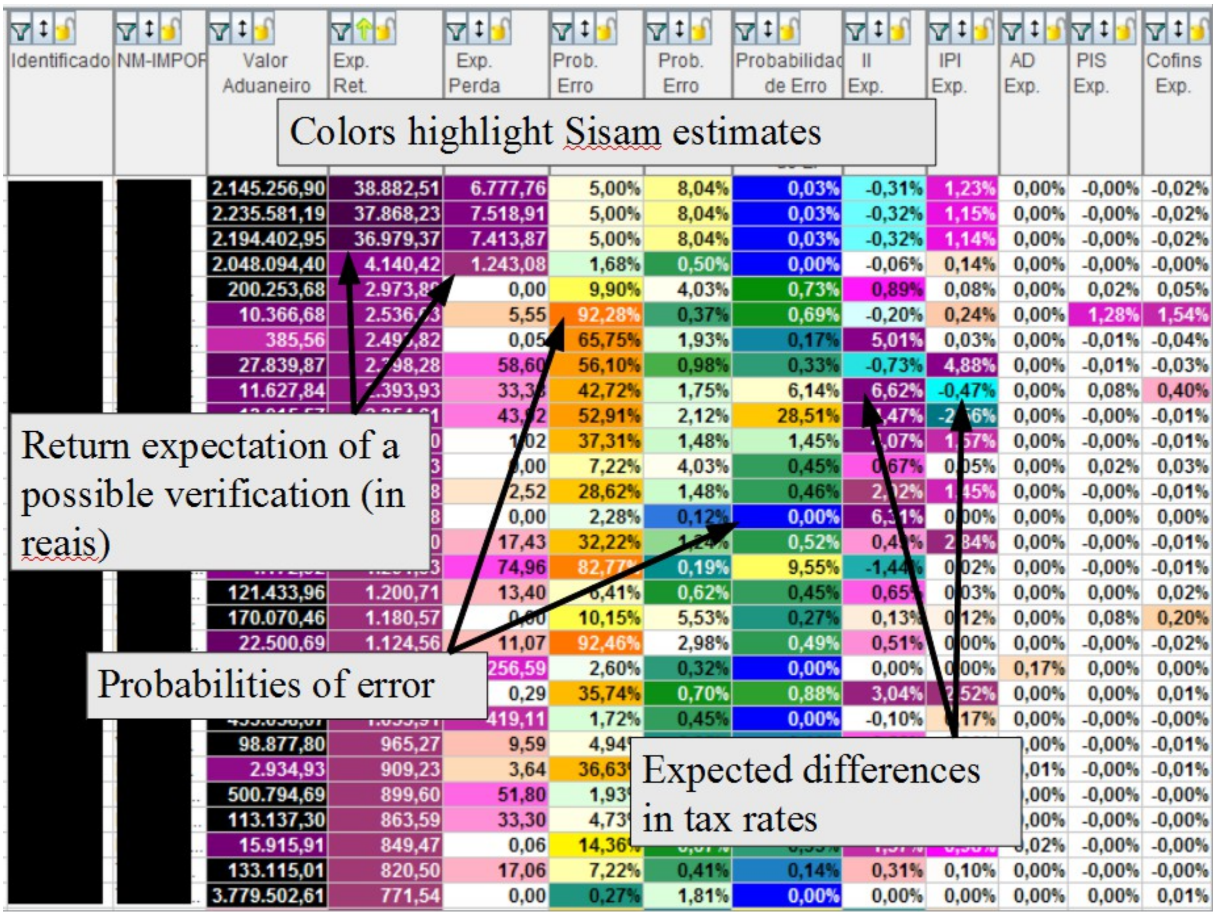

Figure 1: Sisam's default spreadsheet

To understand the spreadsheet it is necessary to know the basic structure of an import declaration. Each DI provides data about several products at the same time. These products are grouped according to some common attributes such as the declared NCM code, the declared manufacturer and the declared supplier. Groups of products are called *additions* of the DI. The products possess some individual data within an addition. The data related to an individual product is what we call an *item* of the DI.

Sisam users can choose to view the worksheet data segmented by DI, by addition or even by item. In Figure 1, the data was segmented by addition what means that each row of the worksheet corresponds to an addition.

The first two columns identify the transaction and the importer and for confidentially issues have been covered with black stripes. The third one shows the customs value of the goods. Since the spreadsheet is segmented by addition this corresponds to the total value of the goods in the additions identified in the corresponding rows. That column could show the total value of the DI or the value of a particular item if the segmentation had been different.

The forth column shows the first value estimated by Sisam. It is the return expectation of the verification. For every item on the DI, Sisam calculated the probability of each possible error and the probability of each correct value in case of error (correct NCM code, for example). It considered the tax consequences (the duty rate of the correct NCM code, for example) and administrative consequences (a possible import license requirement associated with the correct NCM code, for example) for each of these possible values and combined everything to reach this return expectation, that matches, in reais, the interest of RFB in the verification of the item, the addition or the DI represented in that row of the spreadsheet. In the last two cases, the values of the return expectation are the sums of the return expectations of the items contained in the addition or in the DI.

Administrative consequences have been mapped to reais by brazilian customs coordination (Coana) and may be changed. For example, an error in the NCM code without changes in tax rates and without implications in terms of administrative requirements can be mapped to R$1000,00. So, if Sisam considers that there is 10% likelihood that an error like this is present, the return expectation will be increased in R$ 100,00. Informing exactly what are the administrative aspects mapped and what values they have received is beyond the scope of this work.

If a customs officer performs 1000 verifications, all of them with a return expectation of R$500.00 (either for tax or administrative issues), he or she is expected to recover to RFB a value of about R$500000.00. Of course, Sisam is not always going to get the estimate right, but, at least, the user can understand exactly what Sisam is trying to predict and what its developers are trying to have it predicting.

It is not expected that, when doing a single verification with a return expectation of R$500.00, the custom officer is really going to recover that value. In general, if the suspicions raised over the goods are not confirmed, the return is zero and if they are confirmed, it is greater than R$500.00, in such way that the average is R$500.00.

The return expectation guides the customs officers in order for them to make the best use of their time and energy. The verification of a product of intermediary value may have a high return expectation if the error probabilities are high. The verification of a product of low value may have a high return expectation if there are high probabilities of error with relevant administrative consequences, like dodging the requirement of an import license. On the other hand, a product of very high value may deserve the attention of a customs officer even if the probabilities of error are relatively low.

The fifth column shows what we call the expectation of loss. It corresponds to the possibility that the verification ends up finding differences in tax rates that favor the taxpayer and therefore cause financial loss to RFB. It is up to the customs officer to decide whether or not to take action to correct these cases.

A single item may have both high return expectation and high expectation of loss. That happens, for example, when it is suspected that there is an error in the declared NCM code with two possible correct values, one of them with a higher tax rate and the other with a lower tax rate than the declared one.

The sixth column shows the probability that the declared NCM code of an item is wrong. If the spreadsheet is segmented by addition or by DI, the presented probability is the probability that there is at least one item with an incorrect NCM code within them.

The seventh column shows the probability that the declared country of origin is wrong.

The eighth column shows the probability that the goods need an import license that was not presented. That may or may not be a consequence of an error in the NCM code.

The last columns show expectations of the differences between the declared and actual tax rates for Import Duty (II), Tax on manufactured products (IPI), Antidumping duty and Social contributions (PIS and COFINS). The effective tax rates are the ratio of the tax that must be collected to the value of its ad valorem calculation basis. Note that if there is suspicion of specific tax, the values to be collected will be divided by the value of the ad valorem basis so that we can show all tax differences in a single scale.

The differences in tax rates may be consequence of errors in the NCM code or in the county of origin code, whose probabilities are explicitly presented in the spreadsheet. At the same time, they may be consequences of errors in EX tariffs, tax regimes and tariff agreements which, by default, do not have their probabilities displayed in the spreadsheet.

To make the identification of priority verifications faster, the spreadsheet can be sorted by any column and the values are accompanied by colors.

In Figure 1, we sorted the spreadsheet by return expectation. That column follows a gradient that goes from white, for return expectations equal to zero, till black, used for very high values, passing by pink, lilac and purple. In Figure 1, we can't see the white color in this column, because we sorted it and made the smallest values go outside the screen.

The customs value of the goods follows the same gradient. We can notice that high values tend to be associated to high return expectations, but the relationship is not straightforward.

The probabilities of error follow a gradient where white corresponds to an error probability that is considered to be mild, which was fixed in 3%. As the probabilities of error go up, the color becomes yellow, orange and then red. Sisam is not limited to identifying signs of the presence of errors. Learning with the history of import declarations, also allows it to identify particularly low error probabilities. This usually follows from a history of verifications that have always confirmed what had been declared. As the probability of error goes down the color becomes green and then blue.

The expectations of difference in the tax rates, for positive values, follow the same gradient of purple shades as the return expectations and, for negative ones, they follow a gradient of blue shades.

Leaving the mouse over any cell of the spreadsheet, the user receives some kind of explanation or useful information. In Figure 2, we left the mouse over a return expectation of R$ 38882.51. Sisam decomposed such expectation in respect to its nature, which can be tax related or administrative[3] and also in respect to its primary cause[4].

3 In the figure, tax related consequences are labeled with the Portuguese word “tributárias” and administrative consequences are labeled with the word “subjetivas”.

4 In the figure, the possible causes of the return expectations appear in Portuguese and are translated here:
red = error in the NCM code and any consequences;
green = error in the country of origin code and any consequences;
blue = missing import license that is not a consequence of an error in the NCM code neither of an error in the country of origin code;
yellow = error in the import duty rate that is not a consequence of an error in the NCM code neither of an error in the country of origin code;

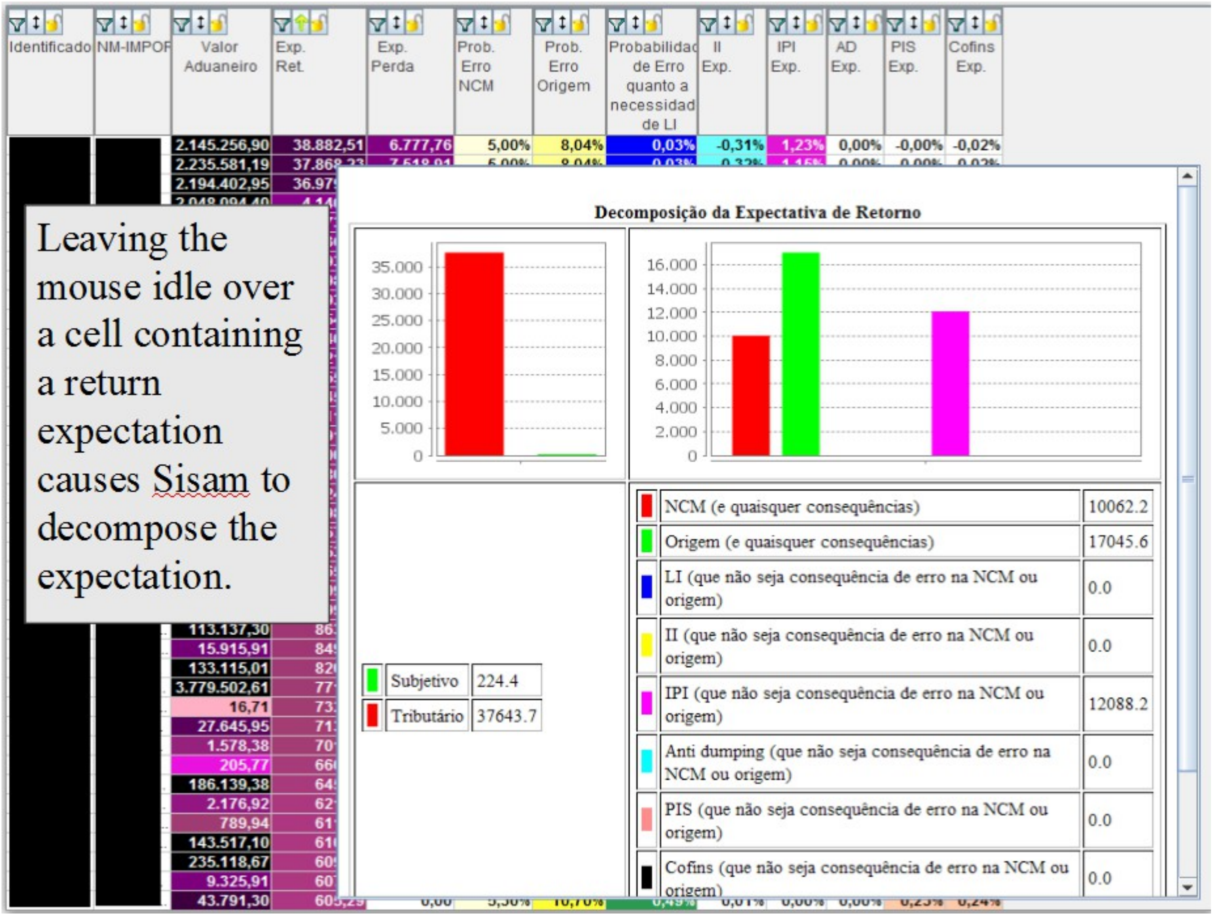

Figure 2: decomposition of the return expectation

The first decomposition conveys to the customs officers whether they should expect to find differences in the collected taxes or non-compliance with administrative requirements. The second conveys to the officers where they should look for the errors. If, for example, there is an expectation of difference in the import duty rate (II) due to the suspicion of an error in the NCM code, it will have been computed under the label "NCM code error and any consequences". If the expectation is due to an error in the tax regime of the import duty then it will have been computed under the

lilac = error in the tax on manufactured products that is not a consequence of an error in the NCM code neither of an error in the country of origin code;
light blue = error in the anti dumping duty rate that is not a consequence of an error in the NCM code neither of an error in the country of origin code;
salmon pink = error in the PIS social contribution that is not a consequence of an error in the NCM code neither of an error in the country of origin code;
black = error in the Cofins social contribution that is not a consequence of an error in the NCM code neither of an error in the country of origin code.

label “Error in the import duty rate that is not a consequence of an error in the NCM code neither of an error in the country of origin code”. The same thing may happen with any other tax. The error in the country of origin code also may impact the return expectation, for affecting tariff agreements.

In Figure 3, we left the mouse over a probability of error in the NCM code that had been estimated in 92.46%.

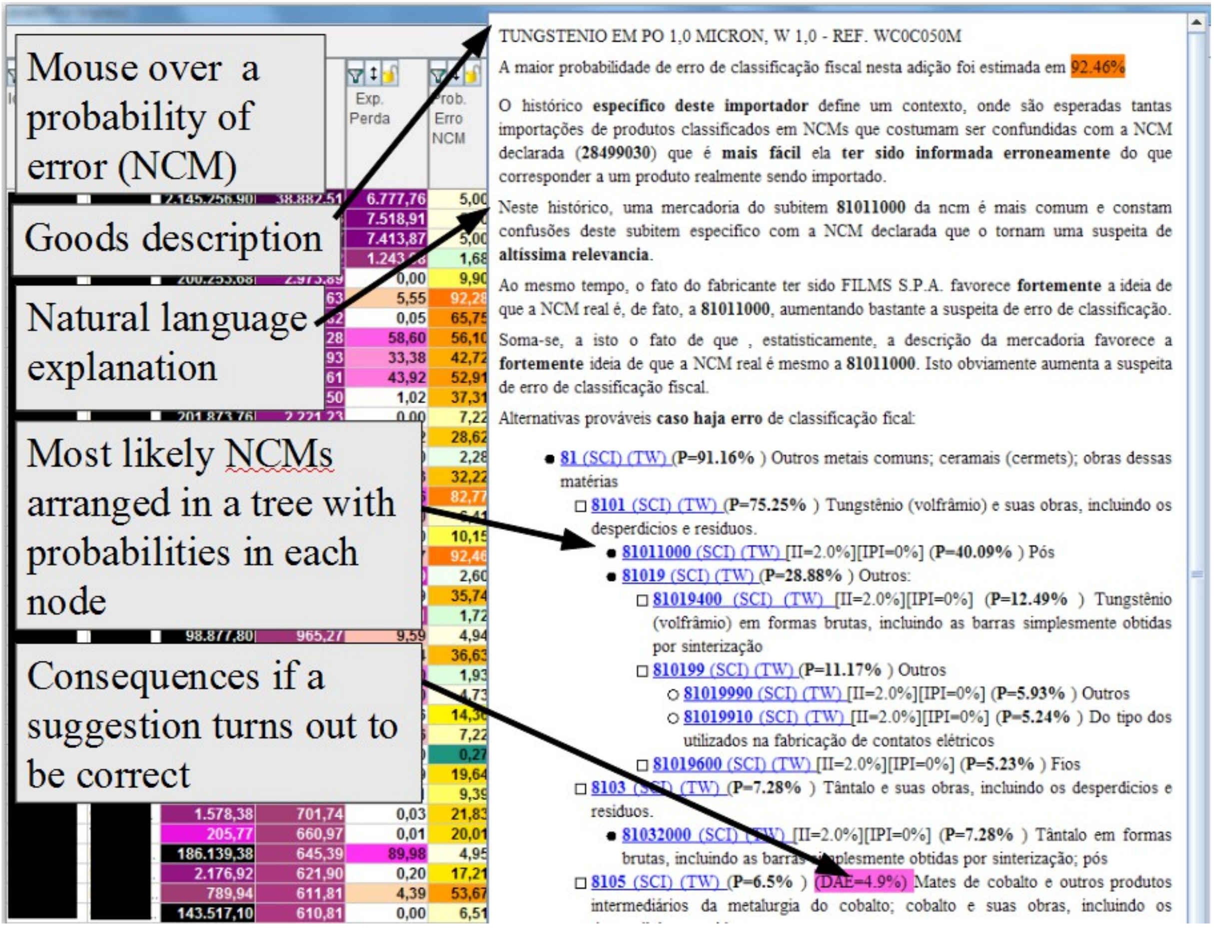

Figure 3: explanation of the suspicion of error in the NCM code

Sisam shows a text[5] containing the goods description as informed by the importer, a natural language justification for the alleged suspicion and a tree of suggestions for the NCM code. The tree is a summary of TEC[6] containing the most likely headings considering everything that Sisam has analyzed. Each NCM code is accompanied by

5 The translated text appears ahead.
6 TEC is the table containing NCM codes and duty rates that is used in Mercosur, it's based on the Harmonic system and thus has a tree like structure.

the estimated difference in the tax rate in case that code happens to be confirmed. The requirement of import licenses is also indicated in the tree.

Clicking on any code the user can see the whole TEC, the official solutions for classification of goods consults and, when Aniita is correctly configured, even open support tools like TECWin. In all cases the users are directed to the information related to the NCM code in which they clicked.

The description of the goods presented in Figure 3 is

> *"TUNGSTEN POWDER 1,0 MICRON, W 1,0 - REF. WC0C050M."*

Sisam reports the probability of error saying

> *"The probability of error in the declared NCM was estimated in 92.46%."*

the natural language explanation is the following:

> *"The history of this **specific importer** establishes a context in which we expect so many import operations of products that are used to being confused with the declared NCM code (**28499030**) that i**t is easier to find it as the consequence of an error** than to find it correctly informed.*
>
> *In this history, articles of the subheading **81011000** are more common and there are confusions between this specific subheading and the declared one, that make **81011000** a **highly relevant** candidate to be the correct subheading.*
>
> *At the same time, the fact that the manufacturer of the product was XXXX S.P.A. **strongly favors** the idea that the correct NCM subheading is, indeed, **81011000**, reinforcing the suspicion that there is an error in the tax classification.*
>
> *We add to this the fact that, statistically, the description of the goods **strongly favors** the idea that the correct NCM subheading is 81011000. This obviously makes the suspicions higher."*

In the example, the goods had been classified as belonging to the subheading 2849.90.30 (Tungsten carbide). Sisam scanned its knowledge base and realized that other NCMs are often confused with it. Reviewing the history of the importer and considering its national classification of economic activity (CNAE), Sisam noted that many of these suspected NCMs are expected for it. In fact, this importer is so strongly expected to import goods of NCMs which are usually declared wrongfully as if they belonged to 2849.90.30, that it is more expected that this code appears as the consequence of an error than correctly applied. This led the system to generate the first paragraph of explanation.

If Sisam had found relevant confusions, but not so much, it wouldn't have said that " *it is easier to find it (the NCM code) as the consequence of an error than to find it correctly informed".* Sisam could have said that given the errors found, the transaction required *"some attention"* or even *"a lot of attention"* without making a statement as strong as done. The system seeks to regulate the tone of text according to the strength of the evidence.

In the second paragraph, Sisam informs to have found a particular NCM code which is more common in the context of the importer and that is used to being confused with the declared NCM code. This NCM code is 8101.10.00 (Tungsten powder) and Sisam tells the user it is a suspect of extremely high relevance. Again, if evidence was weaker Sisam would have been more moderate.

In the third paragraph, Sisam reports to have found that the suspected NCM code is in the roll of goods the informed manufacturer is used to selling to Brazil. For this, Sisam looked at the history of other importers that have made purchases from the same manufacturer. That information increased the suspicion. Note that Sisam introduces the paragraph with the expression "*At the same time"*, an expression that conveys the idea of going on in the same direction. Had Sisam found that the manufacturer does not sell goods of the suspected NCM code to Brazil it would have opened the paragraph in a compatible manner, using "However", "In contrast" or some other expression of opposition.

The explanation ends with Sisam saying that the description of the goods also points to subheading 8101.10.00, what looks correct reading the description of the goods and the NCM code text.

It is still possible that the NCM code that was declared by the importer was actually correct and the description of the goods had failed to include the information that it was a carbide. However, considering every thing that Sisam has found in the knowledge base, that is unlikely. If the goods are verified, the result of such conference, negative or positive, will teach Sisam, reinforcing or mitigating similar suspicions.

Lets see another example of explanation produced by Sisam:

> *"FC-100/F-ULTRAVIOLET LAMP 100W, 230V, FAN COOLED8 PRI, 8 SEC.*
>
> *The probability of error in the declared NCM was estimated in 44.85%.*
>
> *In the history of this **specific importer**, articles of the subheading **85437099** are more common and there are confusions between this specific subheading and the declared one, that make **85437099** a **highly relevant** candidate to be the correct subheading.*
>
> *It is worth pointing out that this importer has had goods belonging to NCM code **85437099** checked by customs officers thrice and **in all cases they had been wrongly declared** as belonging to code **90275090**.*
>
> *The influence of this manufacturer (XXXXXXXXX) was small, but slightly reinforced the idea that the correct NCM subheading would be **85437099**.*
>
> *Moreover, statistically, the description of the goods favors the idea that the correct NCM subheading is **85437099**. This increases the suspicion that the NCM code has been wrongly declared."*

The example corresponds to an ultraviolet lamp misclassified as a machine for physical or chemical analysis. Sisam didn't find a wide range of errors that put the declared NCM code under suspicion, but it has found one particular NCM code that is usually confused with the declared subheading, NCM 85437099. More than that, it also pointed the fact that this importer has never declared any goods to belong to 85437099 spontaneously. All three times that this code appears in the history of the importer, it was due to a reclassification made by customs officers from NCM 90275090. Naturally, it is possible to conclude that if he imports a lamp belonging to 85437099 again he may perfectly well inform that he is importing a measurement instrument of the subheading 90275090. The manufacturer of the product and its description reinforced suspicion, but not as much as in the first example.

When Sisam is able to find concrete cases of error that are very similar to the one that may be going on, that usually contributes a lot to the conviction of the customs officers.

It is worth noting that neither Sisam needs many cases to make such notes (references to single events are not rare), neither the users need many cases to consider the information to be very relevant. This goes against the popular belief that a statistical method needs lots of cases to work well. The ability to learn with few cases is a peculiarity of the highly non-linear models that are employed by Sisam and therefore a benefit of the technology that was specifically developed to meet RFB needs.

The most sophisticated analysis conducted by Sisam is undoubtedly that of the NCM code error, an error that is important enough to have a specialized division within RFB (DINOM) and that has been the focus of a job that received a creativity and innovation award of the RFB (SIFUENTES, 2014).

Nonetheless, Sisam also produces textual explanations for several others errors. To save space we will only show two other examples, this time related to errors in the declared country of origin.

Example 1:

> *“The probability of error in the declared country of origin of this item was estimated in 20.42%.*

*In the history of this importer and of the routes of cargo involving this country of purchase and provenance (**USA**) there are errors in the declared countries of origin which make the possibility that this item was produced in another country (**China**) a relevant suspicion.*

Besides that, the fact that the declared NCM code was 84433111 favors the idea that the true country of origin is, indeed, China and contributes to the suspicion of error in the declared origin."

Example 2:

"The probability of error in the declared country of origin of this item was estimated in 18.57%.

*The influence of the declared NCM code (29291021) was small but favored the idea that the country of origin is actually **Japan**.*

*Moreover, the fact that the supplier of the goods was XXXXX AG greatly favors the idea the country of origin is, indeed, **Japan** instead of India, increasing the suspicion of error in the declared country of origin."*

In both examples, we notice how Sisam analyzes the route of the goods comparing the country of origin to the countries of purchase and of provenance in the light of the errors that have already been detected by customs officers. We also see that the NCM code influences the analysis, as well as the historical trends of the supplier for the type of product being imported.

Sisam does not quite have a poetic fluency and with attention, imperfections in its comments can be perceived, but it is not uncommon that users who see them for the first time really think that there is a person writing the texts. No person could do that for all goods being imported in Brazil.

When they are working in the selection of import declarations for conference during customs clearance, customs officers may select a DI directly from within Sisam`s spreadsheet. When they are working in customs clearance itself, they can navigate in the DIs following Sisam`s spreadsheet in the order that they find to be most convenient to speed up their work.

3 Basic structure of the system

Sisam is a system written in Java language and currently has a platform of seven servers collocated in SERPRO (Brazil`s Federal Data Processing Service). Two of these machines are used for development, tests and homologation. The other five are used for production, three of them to execute the artificial intelligence engine and two for load balancing, database, integration with Siscomex and communication with users. All machines have 12 physical processing cores (24 virtual processors). The machines running the AI in production and the test machines have 64 GB of RAM. The two other machines have 24 GB of RAM.

When an import declaration is registered in Siscomex, it sends the DI to Sisam through a messaging system. After being received, the DI is stored in a MySQL database and a module named "broker" sends it to one of the 3 IA machines. The chosen machine then performs the requested analysis and the result is placed back in the database.

When a user opens a set of DIs in the Aniita system, which runs on the users own machines, it asks Sisam for the analysis. In general the analysis is ready and is sent immediately. If the DI is old or for any other reason has not been processed yet, Aniita sends the DI to Sisam, that makes the analysis at the moment. The results are displayed in Aniita as described in section 2.

When an DI is cleared or rectified, Siscomex also sends this information to Sisam in such a way that it can learn from the history of rectifications of DIs.

All this flow is complex, follows security standards defined by RFB`s technology coordination (COTEC) and is optimized to meet the daily volume of import declarations. The description of the data flow is outside the scope of this work which is focused only the artificial intelligence aspects.

The AI module keeps one knowledge base in each machine where it runs. For speed issues, the knowledge base is kept in files with proprietary form without the use of any database system.

The knowledge base is initially generated offline from a big amount of already cleared DIs. Today in its compressed format it takes about 80GB of space.

Later, in low demand moments (at night or weekends) one of the production machines stops analyzing registered DIs and starts to learn with cleared and rectified DIs. This machine generates a differential knowledge base that is later sent to the other AI machines and is added to the main knowledge base in each machine. The process does not require Sisam to go down, since it can be performed in one machine at a time while the broker redirects the analysis to the other machines.

Only recently the integration with Siscomex went into work. Till then Sisam performed all analysis when Aniita sent to it the DIs that had been downloaded to the users machines. knowledge base updates occurred only when a set of cleared DIs were delivered to Sisam through a manual data extraction performed by Serpro.

Today, analysis are performed in an integrated way and occur as Siscomex sends the DIs to Sisam. The daily updates planned for Sisam are not yet active and the knowledge base is updated by the above described procedure only when initiated by human operators. The complete automation of the process should happen in the following months.

4 The structure of Sisam's artificial intelligence

The core learning structure of Sisam is a set of Bayesian networks (PEARL, 1988). The inference engine of the probability distribution of each variable given their parents in these Bayesian Networks (BNs) have been changed from traditional conditional probability tables for variants of the smoothing hierarchies described in (JAMBEIRO FILHO, 2007a), with the name of Hierarchical Pattern Bayes (HPB), an advance of particular interest for RFB.

HPB hierarchies have been modified to gain speed and precision when the target variable and not only the explanatory variables has high cardinality. Special nodes were introduced in the networks to handle exact conditional implications among variable values (like, for example, the relations between a tax regime of exemption and the tax rate) and to provide the exploitation of structural aspects like the number of subdivisions of a NCM heading.

Another special need that was met was to use the history of non verified DIs of one importer to analyze the others, but not himself, thus preventing the system from considering that a wrong behavior is right just because of its repetition. This also required modifications to the operation mode of a traditional BN.

We also introduced resources to exploit hierarchical relations among variable values like the NCM code that has the structure of a big tree and the countries which were grouped according to economic zones. Such resources allow the system to analyze one variable and its correlations to other variables in all its levels, extracting the maximum amount of information from the database. HPB's smoothing hierarchies, which have in their core, the progressive combination of attribute values, started to deal also with the hierarchical aspects of a single attribute.

Such hierarchies were also extended to handle free texts. Machine learning is used a lot for text classification, but in general, that involves correlating the texts to only one external variable that is called “class”. Using smoothing hierarchies, Sisam correlates the text to several external variables at once and latter extracts the influence of the variable of interest which is the NCM code. That maneuver allows Sisam to perform the so called “explain away” (PEARL, 1988) of the irrelevant parts of the text, understanding them as consequences of the importer, the language spoken in the country of origin, the supplier or the manufacturer and not as a consequence of the NCM code.

We also extended the hierarchies to handle continuous variables and recently sophisticated such extension to gain speed, adaptation to sharp borders, multimodality and asymptotic convenience in face of outliers.

All extensions to hierarchical smoothing mechanisms were done in a way that preserved the possibility of incremental learning, a feature of traditional Bayesian Networks that doesn't exist in several other machine learning methods.

Sisam handles changes in the NCM system incorporating the “from / to” relations established by legislation in its calculations. If one NCM code is divided in two, data collected before the split are still capable of separating the two new NCM codes from the other 10000 and a much smaller amount of data may be enough to separate the two from each other, speeding up learning and avoiding a phase of gross errors that would otherwise follow from legislation changes.

While Sisam makes its calculations, it registers its steps in a structure that can latter be scanned by the natural language generation mechanism. This scanning allows the identification of the patterns that have most affected the calculations, making it possible to choose which comments will be made and to calibrate the tone of the employed language. The generation of texts is based in this scan and in context variables which allow proper inflexion of words in respect to gender and number as well as the choice of conjunctive adverbs of addition or opposition that brings fluency to the text.

Given the amount of unknown variables, it was not possible to treat them all in a single Bayesian network. We had to use several separated networks and connect them “off the books” in a more restricted way then a global unification. The BNs that handle errors in the NCM code and errors in the country of origin are applied first and their results are used to cut the search space of the networks that handle other errors (tax regime, tariff agreement, tax rates, etc.). Besides that, we created a mechanism that groups together attribute values that are estimated as less likely by faster marginal analyses into a single value called “others”, with that reducing the amount of hypotheses evaluated by the full probabilistic model. The presence of the “others” value also requires modifications on the way BNs usually work.

Sisam's knowledge base includes about five billion patterns (combinations of attribute values). The information related to these patterns goes from simple frequencies till non parametric distributions of continuous variables (HASTIE et al., 2001). This knowledge base does not fit in memory and is manged by an aggressive system of swap to disk. It is able to anticipate which data will have to be loaded in memory and activate the disks while calculations proceed in a massively parallel way without having to stop to wait for them.

The data that need to be stored during training phase are automatically determined when Sisam runs the “compilation” of probabilistic models. The existence of this compilation allows each model to be separately specified while their physical application benefits from every common part of them. The compilation also determines the positions where the models will be put in memory and, for maximum efficiency, generates big blocks of data that are defined according to the order in which the inference engine scans the hypothesis space. That reduces swap, boosting the efficiency of a cache system that is structured in three levels: central associative cache, per thread associative cache (to minimize synchronization actions) and positional (to reduce calculation of hash functions and comparisons).

Data models are not represented in memory by traditional Java objects, but by huge vectors of bytes that are interpreted by system as statistical models. That reduces the use of space in 14 times and reduces the number of live objects in the JVM in orders of magnitude and thus avoid known performance problems of the garbage collector when heaps of several gigabytes are employed.

As already described, Sisam's learning involves the identification of the differences between registered and cleared versions of import declarations. That identification must be done before we try to train the Bayesian networks and is not trivial. DI items do not have unique identifications and when they are rectified they are frequently moved to DI additions that are different from the one where they were in the original version of the DI. That is frequently necessary because that item stops having the same attribute values as the ones of the other items in the addition, what is not admitted by Siscomex. It is not rare that they are moved to other import declarations (that fortunately share the same bill of lading number), since Siscomex, for a long period of time, did not admit the creation of more additions in an already registered DI. Sisam needs to find the item wherever it goes to compare it with its original version and discover what changed. That search is done by a mechanism that was adopted by the Contagil system for name comparison in payrolls. The mechanism aligns the original DI with all final versions of DIs with the same bill of lading number pairing the items.

Besides working as a decision support tool for customs officers, Sisam has the ability of selecting DIs for conference automatically using a share of Siscomex automatic selection. This resource is not yet active, but is implemented and works using decision theory, which recommends the selection of the DIs with highest return expectation, and game theory, that recommends being unpredictable to the opponent, making uneven random selections where the probability of the selection of a DI is proportional to its return expectations. That prevents anyone from feeling comfortable and staying bellow Sisam's radar.

5 Performance measures

In this section we show two types of performance measures. In the first, predictions made by Sisam's artificial intelligence module acting standalone over goods that had already been verified by customs officers were compared to the actual results of the verifications. In the second, we present results related to the use of Sisam in production, working as a decision support tool for customs officers. We also included measures that show that Sisam is able to meet the daily volume of DIs registered in Brazil.

Accuracy measures of the artificial intelligence acting standalone

To make measures that represent the performance of Sisam artificial intelligence as a standalone entity, we selected a test sample containing 624517 items, each one corresponding to a particular product. The items came from 187417 additions of 36291 DIs, all cleared after verification by customs officers (if we included non verified DIs in the test sample, we would not know whether predictions were right).

Sisam was trained with 98798545 items from 27909669 additions of 5509000 DIs, where 88% of them had been cleared without verification and 12% had been verified before clearance and thus were more informative.

Sisam AI module was isolated form the rest of the system and applied offline to analyze all items in the test sample. No DIs were both in the training sample and the test sample.

Performance in the prediction of errors in the declared NCM codes

To evaluate the system's performance we measured the recall for each selection rate, S, with S going from 0% to 100%. For each value of S, we considered that the $S \cdot N$ items with greatest probability of error would be selected, where N is the total number of items in the test sample and that all other items would be cleared without verification.

Recall is given by

$$Recall = \frac{Total\ number\ selected\ items\ with\ error}{Total\ number\ of\ items\ with\ error}$$

The total number of goods with classification errors in our test base is of 6007 what corresponds to a little less than 1% of the total number of items.

In Figure 4 and 1 we show recall for errors in NCM codes. To understand the chart it's just necessary to notice that when selection rate is zero, obviously, no error is captured. When selection rate is 100%, also obviously, all errors are captured. What is important is that the curve rises fast showing that with small selection rates we can capture a high percentage of the errors, that is, it is important to have high recall with low selection rates.

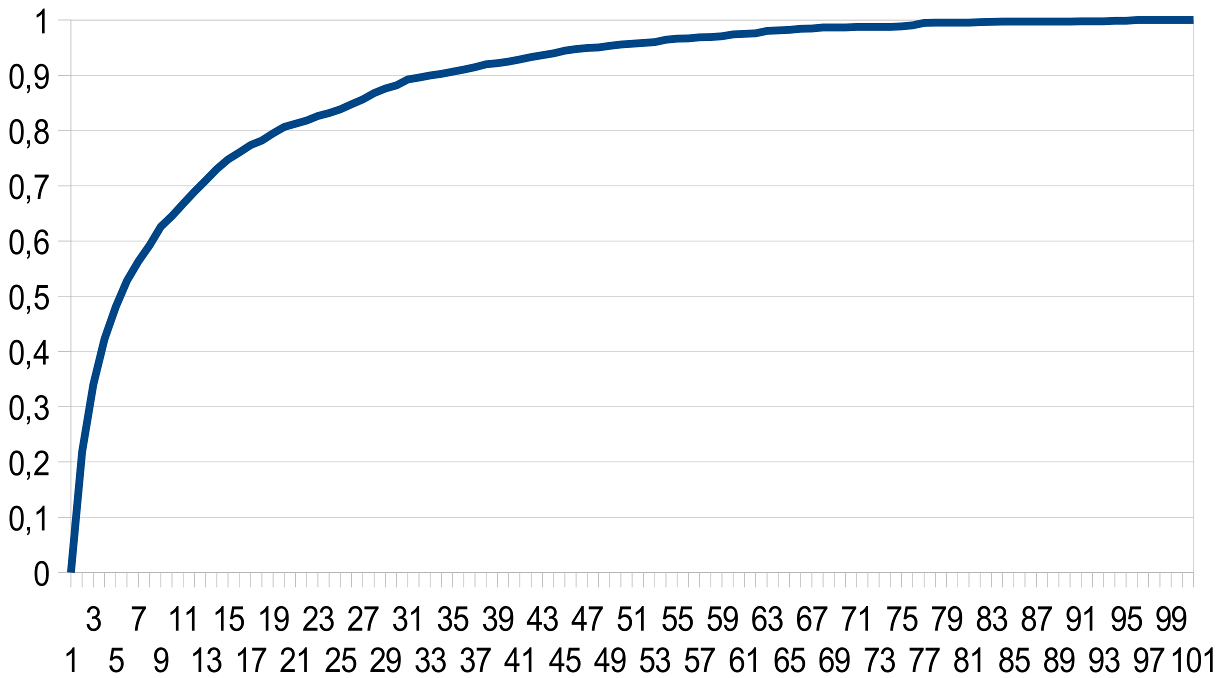

Figure 4: recall curve for errors in NCM codes

Selection rate	1%	2%	5%	10%	20%	50%	75%
Recall	22%	34%	52%	66%	81%	96%	99%

Table 1: recall rates for errors in NCM codes

Observing 1, we note that if, instead of having verified the more than 600000 items in out test base, the customs officers had verified only 2% of it, as long as the 2% were the 2% with greatest probability of error in the NCM code according to Sisam, 34% of all errors that have been captured still would, what corresponds to a leverage of 17 times in relation to the selection rate. At the same time, if they had checked 10% of what they did, they would have captured 66% of all errors in the sample. If, on the other hand, they had verified 75% of all they did, 25% of the effort would have been saved, but only 1% of errors would have escaped because of that.

Observation of Figure 4 and 1 let us safely assert that the system brings advantage for every chosen selection rate.

Naturally, if the customs officers had verified 10% of the number of items that they actually verified, they could have verified many other items that were released in the green channel (cleared without conference). If those errors had probability estimates that were similar to those in the test sample, the efficiency in the capture of the errors would be multiplied by 6.6.

The most adequate selection rate varies according to RFB's policies and with the context. For customs officers that work selecting DIs for conference during customs clearance and handle all DIs registered in their customs unit (in some cases, there are more than 10000 items each day), selection rates of about 1% are relevant, since they usually don't have time to select more than that. For that selection rate, leverage is of 22 times because in this range there is a concentration of promising items.

For officers that work directly in customs clearance and work only with DIs that were distributed to them and who, in the past, were forced to check 100% of the items, much higher selection rates are reasonable (note that now we are talking about the percentage of items of a DI that will be verified). Today, officers that work in customs clearance only need to check the parts of the DI that motivated its selection and, at the same time, have the freedom to check whatever they consider pertinent. It is perfectly reasonable to select 20% of all items in one DI and check their NCM codes even physically verifying the goods. Doing that they would still be capturing more than 80% of errors and releasing themselves for other tasks. It is also reasonable to try to check the whole DI but in the order indicated by Sisam. If, in this case, they end up only managing to check 75% of items, only 1% of the errors would escape.

In 3, we show that Sisam has a good capacity of indicating the correct NCM code too. We see that 40% of times the code with highest probability is, indeed, the correct one and that 65% of times the correct code is within the first 5 suggestions. That is a very favorable result considering that there are 10000 possible codes, that several quite different products belong to the same code and that the classification process follows many non trivial rules.

Position in suggestions list	1	2	3	4	5
Percentage for all items	40%	52%	57%	61%	65%
Percentage for the two percent of items with the greatest probability of error	56%	67%	70%	75%	77%

Table 2: position of the correct NCM code in the suggestion list

We only counted as correct the cases where Sisam hit the exact subheading of the NCM. Even when it cannot hit the subheading it can still hit the heading or at least the chapter speeding up the work of the customs officer.

In the last row of 3, we see that when Sisam produces a high estimate for the probability of error it can hit the correct NCM code more frequently. That is expected because the same evidence that made Sisam to be suspicious of the error also lead it to identify the correct code. This tendency is advantageous because customs officers tend to verify the cases that are most likely to have errors and in this cases it is desired that they quickly find the correct NCM codes.

Performance predicting errors in the declared country of origin

Our measures for errors in the declared country of origin are completely analogous to the measures for NCM code errors. The total number of errors in the test sample was 1890.

In Figure 5 and 3, we show recall for errors in the declared country of origin.

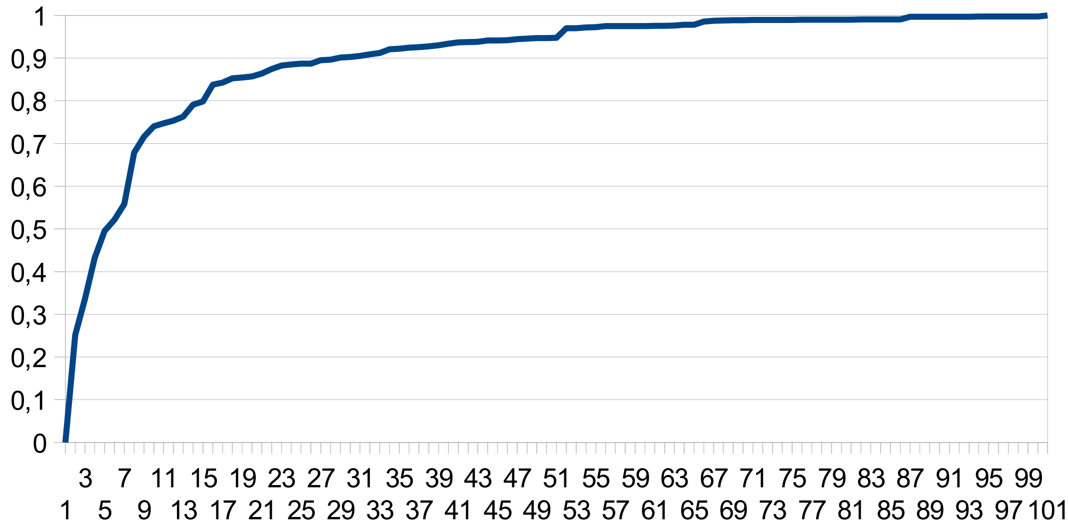

Figure 5: recall curve for errors in the declared country of origin

Selection rate	1%	2%	5%	10%	20%	50%	75%
Recall	25%	34%	52%	75%	86%	95%	99%

Table 3: recall rates for errors in the declared country of origin

In 4, we show the position in which the correct country of origin appears in the list of suggestions.

Position in the list of suggestions	1	2	3	4	5
Percentage for all items	53%	64%	72%	82%	83%
Percentage for the two percent of items with the greatest probability of error	74%	86%	90%	90%	90%

Table 4: Position of the correct country of origin in the list of suggestions

In general, hit rates for countries of origin are superior to the ones for errors in the NCM code. Sisam is more sophisticated in the way it treats the NCM code than the country of origin. In spite of that, Sisam makes it right more easily in the second case because the problem is inherently easier. There are only 200 countries instead of 10000 and the complexity of the confusions that are used to happening is smaller.

Performance in the prediction of errors in license requirements

In Figure 6 and 5, we show recall for errors in license requirements. That includes both the case where a required license is missing and the cases where a license was presented when it was unnecessary. In general, errors are of the first type, but the second also exist. In or test sample there were 986 errors in license requirements.

With a selection rate of just 1%, 51% of the errors in license requirements are captured, an enormous leverage. Great part of them follow from NCM code errors, where a product has its NCM code changed from a subheading that does not require a license to one that does. However, within some subheadings, a license is required for some products and not to others, making errors in license requirements possible even if the NCM code is correct.

Results in Figure 6 and 5 include all errors in license requirements.

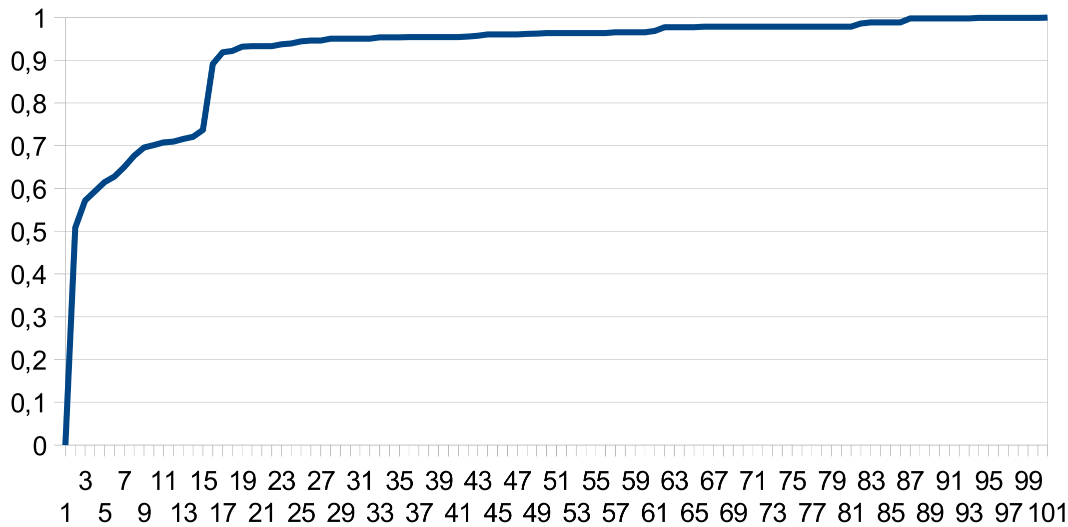

Figure 6: recall curve for errors in license requirements

Selection rate	1%	2%	5%	10%	20%	50%	75%
Recall	51%	57%	63%	71%	93%	96%	98%

Table 5: recall rates for errors in license requirements

Although the success rate are in general very good, the fact that we observe an almost stabilization of recall between 10% and 15% with a sudden rise just after, shows a deficiency in our mechanism. One of two things must be happening: either we are underestimating the probability of error for a relevant group of goods, moving them towards the end of the ranking or we are overestimating the probability for a group that is being moved towards the beginning of the ranking pushing other items toward the end.

We know that changes in legislation that establish and remove license requirements are quite common. The speed of adaptation of Sisam to changes in law is still low and that may explain many mistakes in the predictions related licensing. We intend to speed up that adaptation and approach the issue in section 7.

Performance predicting other types of error

In Figure 7 and 6, we show recall for errors in tax regimes, in the existence of a reduction in the calculation base of social contributions and in the informed tariff agreements. 6 also contains the number of errors found in our test sample for each type. The level of effort to improve Sisam performance in this errors was, till now, smaller than the effort made to handle errors in NCM code. In spite of that, several recall rates are very high.

The explanation for that is that error in tax regimes almost only happen when a special regime is requested (though Sisam also point cases where a special regime should have been requested and was not). The same thing happens with tariff agreements and to the reduction of the calculation basis for social contributions. This makes these errors more easily predicted.

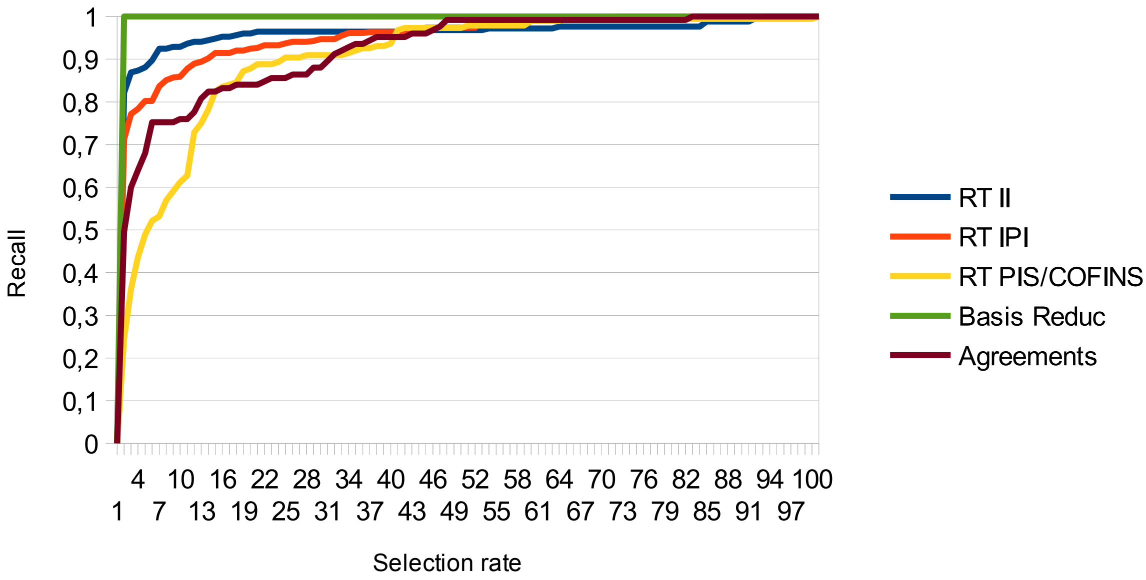

Figure 7: recall curves for several types of error

	Total	1%	2%	5%	10%	20%	50%	75%
TR II	252	82%	87%	90%	94%	96%	97%	98%
TR IPI	490	71%	77%	80%	88%	93%	97%	100%
TR PIS/COFINS	188	24%	36%	52%	62%	89%	98%	99%
Reduction in the calculation basis for PIS/COFINS	13	100%	100%	100%	100%	100%	100%	100%
Tariff agreements	125	50%	60%	75%	76%	84%	99%	99%

Table 6: recall rates for several types of error

We plan to invest more in the way we handle these errors. When we do that we will also make more precise measures that distinguish success rates for each requested regime. We will also select a test sample that includes more cases of special requests. In the tests we made, there was, for example, only 13 cases of error in the reduction of the calculation basis of social contributions deforming the results.

What is more important in Figure 7 and 6 is that Sisam is already able to help in the detection all those types of error.

Performance in the prediction of differences in the effective tax rate

We define the effective tax rate by

$$Effective\,tax\,rate = \frac{Total\,Collected\,Tax\,Amount}{Customs\,Value\,of\;the\,Goods} \quad .$$

Thus, it already includes the import duty (II), the tax on manufactured products (IPI), the social contributions (PIS and Cofins) and Anti dumping rights. The effect of tax regimes, tariff agreements and reductions in the calculation basis of social contributions are also embedded in the effective tax rate.

As we see in the equation that defines it, the effective tax rate is an ad valorem concept, but values collected as consequences of the existence of specific taxes are also included.

In Figure 8 and 7, we show recall for differences in effective tax rates. The concept of recall in this case is somewhat different to reflect not only the existence of of differences, but also their values.

$$Recall = \frac{\mathrm{Sum}\,of\;the\;differences\,\mathrm{in}\,the\;selected\;items}{\mathrm{Sum}\,of\;the\;differences\,\mathrm{in}\,all\;items} \quad .$$

For small selection rates we have a great leverage. Selecting only 1% of the goods for conference we could recall 27% of the sum of all observed differences in effective tax rates. So, we could multiply efficiency by a factor of 27.

The goods selected when selection rate is low are those with high probability of error and big tax consequences. Sisam can frequently point such cases successfully as consequences of errors in NCM codes, countries of origin, tax regimes, tariff agreements and ex tariffs.

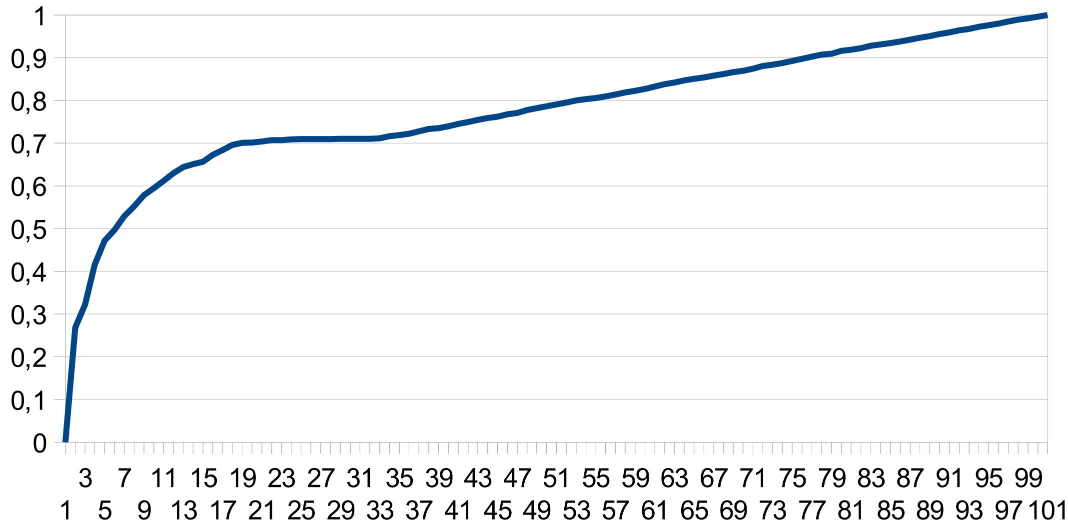

Figure 8: recall curve for differences in effective tax rates

Selection rate	1%	2%	5%	10%	20%	50%	75%
Recall	27%	32%	50%	61%	70%	79%	90%

Table 7: recall rates for differences in effective tax rates

For higher selection rates we observe a performance drop in the predictions. We assign that drop to slow adaptation to new legislation that changes tax rates. As we will see in section 7, we already have plans to improve this.

Measures of suitability to the daily volume of import declarations

In 8, we show the amount of DIs, additions and items and processing time observed in our tests, using only one machine with 12 processing cores and 64 GB of RAM.

Total number of DIs	Total number of additions	Total number of items	Total time	Time per DI	Time per addition	Time per item
36291	183417	624517	47665 s	1,3 s	0,26 s	0.08 s

Table 8: volume and processing time

In a business day, about 10000 DIs are registered and, in rare peak days, that number can go near 20000. That corresponds to 100000 additions and 340000 items.

DIs are sent by Siscomex to Sisam that has to analyze them as fast as they arrive. DIs concentrate in business hours and between 4 pm and 5 pm they peak. In this time interval 13% of DIs are registered. So in one hour 2600 DIs need to be analyzed. Since Sisam analyzes DIs item by item it is important to consider that this corresponds to 44200 items.

The number of items in a DI varies a lot. Many DIs have only one item and others (rare cases) have more than 30000. If all biggest DIs were registered at once there would be no way to handle them. The best we can do is to leave a safety margin in the required processing power to handle DIs of average size.

As shown in 8, Sisam is able to process 45.000 items in one hour using a single machine. Since it has three machines in production, there is a good safety margin.

Time consumed by communication with Siscomex and with user machines are also satisfactory, but their details are beyond the scope of the present work. In practice, in production environment, Sisam can handle about 2 DIs per second what is near to what happens when AI is applied standalone.

Measures with the system being operated by customs officers

In this section we show measures related to the use of Sisam by customs officers both in the task of selecting DIs for verification during customs clearance and in customs clearance itself.

Selection of DIs for verification during customs clearance

To evaluate the performance of the redirection of DIs from green channel (clearance without verification) to yellow, red and gray channels (clearance upon verification) held by customs units, Brazil's customs coordination (Coana) made a survey using a sample of 7201 DIs redirected by customs officers from April to May 2015.

Any redirect is accompanied by a textual justification which is registered in Siscomex. The customs officers had been instructed to include the word “SISAM” when the motivation for the redirection had been originated by this system. Other reasons for redirects are also identified by text expressions.

In 9, we have the list of the most common motivations for redirects. Sisam is the most frequent motivation, followed by suspicions associated to NCM codes whose justifications did no include the word “SISAM”. We believe that a relevant part of those redirects were influenced by Sisam that was not mentioned by accident, but we will not count that.

Sisam	NCM	Weight	Incomplete description	Undeclared corporate links	Import License after shipping	Foreign Exchange coverage	Others
31,62%	16,78%	12,80%	5,93%	5,46%	2,49%	0,97%	23,96%

Table 9: redirect motivations

The percentage of Sisam redirects with positive results (indicated by the presence of rectifications in the DIs are in 10.

Yellow channel	Red channel	Gray channel
61%	75%	85%

Table 10: percentage of redirects with positive results per channel

Most Sisam based redirects have errors in the NCM code as their main suspicion, what cannot always be confirmed in the yellow channel, which does not include physical verification of the goods. In red and gray channels we have more than 75% of positive results. The rule based selection of Siscomex have an average of 30%.

Today, the data available in Brazil's customs data warehouse does not allow us to know exactly what was rectified in the DIs, but it is known that some type of rectifications are much easier to be obtained. Some of them are spontaneous, like the frequent changes in the legal proxy. Changes in goods descriptions that may or may not be relevant are also very common. Together with other small rectifications that don't represent detections of infringement, those changes establish a floor to the percentage of rectification making it much easier to achieve 30% than 75%. It should also be considered that Siscomex chooses first and may select all obvious cases. Redirects are held from the DI sets that remain in the green channel, where errors may be much more subtle.

Thus we can say that redirects held with Sisam help are really much more precise than Siscomex rule based selection. On the other hand, we cannot claim that that is a result of the system itself. That is a result obtained by the customs officers using the system. What really shows the usefulness of Sisam is the fact that customs officers choose to perform more than 30% of all their work based on the information produced by it. That only happens because Sisam identifies relevant suspicions, makes plausible suggestions for corrections, presents those suggestions timely and explains the reasons for the suspicions and suggestions, so that they are quickly understood by the customs officers who have to make selections among thousands of goods every day.

Customs clearance

To survey the usefulness of Sisam during customs clearance itself we made the Aniita system ask the customs officers if Sisam had helped them. The question is made when the customs officer finishes the work with a set of DIs and removes them from Aniita. To avoid disturbing the customs officers the question is made only once for every DI set that was about to be removed.

The officer may answer that Sisam did not help him, that Sisam helped in some cases, that Sisam helped in many cases and that Sisam helped in all cases.

When the customs officer answers that Sisam helped him, that is stored in the database. Unfortunately we failed to register how many times the question was presented and we don't know how many times the officers answered that Sisam did not help them. We collected, however, 3884 claims that Sisam has been useful, and they are segmented according to 11.

The fact that we had 56% percent of claims that Sisam was useful in many cases or in all cases is very favorable. In particular the fact that 32% of the DIs were included in groups where, according to the officer, Sisam was useful in all cases is really surprising, specially considering that 94% of the DIs had not been selected by Sisam, but by Siscomex (80%) or other criterion of redirection.

Useful in some cases	1680	44%
Useful in several cases	947	24%
Useful in all cases	1217	32%

Table 11: usefulness of Sisam during customs clearance

That confirms our previous claim that Sisam does no need to have been the reason for selection to be useful during clearance.

6 Derivative work

Sisam's technology is already employed in two functions of the Contagil system, both in the field of internal taxation. The first is the inexact match mechanism that in Sisam is used to align the registered and the cleared versions of import declarations (see section 4). Within Contagil, the mechanism is used to align lists of people names in the supervision of payrolls, where the lack of the CPF, Brazil's Federal Revenue identifier for natural people, is common.

The second application of Sisam's technology is the error-detection mechanism for NCM codes and CFOP codes in invoices (MDECNF) and is described in Contagil's manual. The goal is to detect undue credits of social contributions (PIS and Cofins).

These credits are usually due when a company acquires feedstock, but not when it acquires materials for internal use or resale. The Fiscal Operation Code (CFOP), indicates the type of use that will be made of any purchased goods and thus determines if the company has the right to credit. The detection of wrong CFOPs is, therefore, essential.

Based on a set of invoices, MDECNF learns correlations between the economic activity of the company, indicated by its CNAE, the type of goods, indicated by the NCM code, and the type of use of the goods, indicate by the CFOP. Since, to disguise a CFOP error, it is common for the companies to inform wrong NCM codes, it is important to detect errors in such codes too.

The specific history of each taxpayer, the descriptions of the goods and the nature of third parties with whom the company does business are taken into account to estimate error probabilities.

MDECNF does not count on the platform for probabilistic model compilation and knowledge base generation that was latter developed to Sisam, what makes its updates more laborious. Is also does not count on central servers and needs to download the full knowledge base to the user machine, forcing us to keep this base small. It has about 160MB and is, so, 500 times smaller than Sisam's base. Moreover, there is no automatic system to update the knowledge base, what impairs its performance.

We hope that in the future MDECNF receives all Sisam's resources and achieves a comparable performance. For the purposes of this work, the important thing is that MDECNF proves that the interest in Sisam's technology is not restricted in any way to the customs and could be useful in all fronts of RFB's work.

7 Future work

Over the next few years, Coana, not only intends to continues to improve import supervision, but also to optimize supervision of all other customs procedures, among them: supervision of export goods, supervision of postal and express consignments, supervision of the granting of licenses to operate in foreign trade, monitoring of customs transit and surveillance of accompanied luggage.

Sisam is inserted in Coana's plans for all those areas, in fact, the experience with import operations shows that it may make important contributions in all of them.

It is worth remarking that Sisam is not the only resource that is planned to be used in any of the intended optimizations. In the same way as on import, it will be integrated with existing tools and other tools that are still to be built. In particular Sisam must be integrated with new Aniita modules that are being constructed for several new areas.

The Aniita system optimized the application of the intelligence of customs officers directly. Sisam learns with customs officers and with the importers and adds machine intelligence to the process. In this way, artificial intelligence is added to human intelligence and may be decisive or merely supportive, but nevertheless positive.

Improvement in import supervision

In the import area we intend to improve Sisam's sensitivity to time, speeding up the way it adapts to legislation changes. Time is usually handled with the use of time series (HOLT, 2004), but it is hard to conciliate such series with all requirements that are already met by Sisam.

Our approach to deal with that will be to treat time as one more continuous variable within Sisam's Bayesian networks. Time is, anyway, a continuous variable and thus the approach makes sense. However, simplifications that are usually reasonable for other variables like a normality assumption or even the weaker monomodality assumption are completely unacceptable for time. Other variables vary with time and mathematically we can transform that into an oscillation of time as a function of other variables. This inversion allows the fit of time in the structure that we already have and that solves so many other problems. So, if we are able to model continuous variables in a way that contemplates oscillations we can reach our goal. The mechanism that we have just developed to handle continuous variables and mentioned in section 4, will be essential in this task.

Besides the improvement in time treatment, we intend to improve our handling of variables like price, quantities and weight, which are also continuous. They have explanatory interest in the detection of the errors that Sisam already handles and are themselves target variables, specially in what regards to over and under invoicing.

Another planned improvement is the prediction of errors in goods descriptions. Today, Sisam already considers the possibility of error in goods descriptions when handling errors in NCM codes. However, it does not suggest explicitly that a description is wrong, something that it should start to do. Sisam won't be restricted to saying that there is suspicion of an error in the description. It will also inform which words it believes that should not be part of the text. This is possible, because, when a customs officer requires a change in a description, the change is registered. Today, Sisam already aligns the initial and final texts finding out what parts of the original text were replaced by what part of the final text (this is part of the alignment of DI items). Using this information more intensively, Sisam may start to say things like:

> *“in the description of the goods, where it is written '**aluminum screw'** consider the possibility that it should be written '**stainless screw**'”.*

The point till which such improvement will be implemented will depend on the costs in processing power and memory that are perceived when we start to test it.

We also intend to improve the treatment of codes that appear in the middle of text descriptions using a modified version of n-grams (GUSFIELD, 1997). With this Sisam will be able to to deduce that if a HP810 is a printer and a HP820 is another printer, then a HP830 should also be a printer even without ever having seen such code.

Another important point is to activate the automatic selection of Sisam, since, today, it is only working as a support tool for human decision-making. This automatic decision-making is less precise than the decision-making of a customs officer, but allows a higher percentage of selection. This percentage would replace part of Siscomex rule based selection that is much less precise than Sisam. Besides that, game theory aspects that customs officers are unlikely to take into account would be, what would lead to selections that even if not maximizing the immediate results would induce spontaneous behavior more efficiently (see section 4).

It is also possible to make Sisam start to interact directly with importers, giving them the opportunity to rectify certain declarations before they are submitted to an officer. That procedure would be similar to what RFB does for income tax supervision and would allow more corrections without costs with human workforce.

Using data originated in other databases of RFB besides Siscomex will also bring import gains. Indexes related the health and structure of companies should be very useful, as well as the relations among individuals which are available in Contagil's social network.

Another important point is to lead Sisam to help in the detection of rarer and more serious errors like counterfeiting, fraudulent third-party intervention and drug trafficking. Because companies involved in these activities, after being caught, generally stop acting, individual histories for machine leaning are small if not null. However, the analysis of the general structure of the companies and of the relationships among individuals are the hooks that will make the statistical approach of Sisam possible also in this areas.

Sisam in other customs procedures

Although they are already among our goals, details about Sisam extensions for monitoring of customs transit, surveillance of accompanied luggage and supervision of the granting of licenses to operate in foreign trade are not defined yet.

Application of Sisam to export area is natural, given its similarities with the import area. The NCM code error-detection mechanism should be completely reused. However, there is a lot of worry with errors that Sisam does not handle yet, like over and under invoicing, counterfeiting and drugs. So, the new resources that were mentioned as improvements in import supervision will be even more important here.

Supervision of postal and express consignments also demand a good system to assess prices, but with an extra obstacle. In these operations, the NCM code does not need to be informed, what impairs statistics about the price of goods. The statistics have to be done only by the text descriptions. The texts will be considered directly, but an approach to obtain a more structured starting point is to use the database of normal import declarations to deduce the NCM code for the postal consignment and analyze the price considering the code. For this, the import declarations module will need to gain the ability to act as an online support to a different module, starting an integration of different machine learning processes that, in an ideal world, would contaminate the whole RFB.

8 Conclusion

We presented the Customs Selection System through Machine Learning (SISAM), an artificial intelligence that learns with the history of import declarations and boosts the efficiency of customs clearance.

We showed, by its list of requirements and by the description of its architecture, that one could not expect to find an equivalent system in the market, neither build it only applying preexisting academic theories.

We also showed, both with tests where the artificial intelligence acts standalone and with tests where it interacts with customs officers that the gains are palpable.

At last, a list of relevant future works associated to Sisam technology and the existence of two applications derived from this technology already in production allow us to claim that Sisam goes beyond the application of data mining techniques and opens the way for RFB in their research and development. This puts the institution in a position that is compatible with the potential and the complexity that are inherent of its gigantic and amazingly rich information environment.

9 References

CARVALHO, S. C.. Um Método de Inferência Difusa para Classificação de Sonegadores Fiscais. Prêmio de Criatividade e Inovação da RFB, 2014.

COUTINHO, G. L.. Aniita – uma abordagem pragmática para o gerenciamento de risco aduaneiro baseada em software. Prêmio de Criatividade e Inovação da RFB, 2012.

FERREIRA, M.A.C, Uso de redes de crença para seleção de declarações de importação. Tese de Mestrado, Instituto Tecnológico de Aeronáutica, 2003.

FERREIRA, M.A.C, Seleção Probabilística, - Melhorando a eficiência da conferência aduaneira. Prêmio de Criatividade e Inovação Auditor-Fiscal José Antônio Schöntag, 2003.

FIGUEIREDO, G. H. B.. Um Novo Paradigma na Auditoria em Meio Digital. Prêmio de Criatividade e Inovação Auditor-Fiscal José Antônio Schöntag, 2008.

GUSFIELD, D.. Algorithms on Strings, Tress and Sequences: Computer Science and Computational Biology. Cambridge U. Press, New York, 1997.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H.. The Elements of Statistical Learning. Springer, 2001.

HOLT, C. C.. "Forecasting Trends and Seasonal by Exponentially Weighted Averages". International Journal of Forecasting 20 (1): 5–10, 2004.

JAMBEIRO FILHO, J. Tratamento Bayesiano de Interações entre Atributos de Alta Cardinalidade. Tese de Doutorado, Instituto de Computação da Universidade Estadual de Campinas (IC/UNICAMP).

JAMBEIRO FILHO, J.; WAINER, J.. Using a hierarchical bayesian model to handle high cardinality attributes with relevant interactions in a classification problem. In Proceedings of the International Joint Conference of Artificial Intelligence (IJCAI). AAAI Press, 2007.

JAMBEIRO FILHO, J.; WAINER, J. HPB: A model for handling BN nodes with high cardinality parents. Journal of Machine Learning Research (JMLR), 9:2141–2170, 2008.

JAMBEIRO FILHO, J.. Inteligência Artificial no Sistema de Seleção Aduaneira por Aprendizado de Máquina. Prêmio de Criatividade e Inovação da RFB, 2015.

MITCHELL, T. M.. Machine learning. McGraw-Hill, New York, 1997.

PEARL, J.. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., 1988.

RUSSEL, S.; NORVING, P. Inteligência Artificial. . 3 ed. Rio de Janeiro: Campus, 2013. 1021 p. ISBN 9788535211771.

SIFUENTES, M. C.. A Nova Sistemática de Solução de Consultas em Classificação de Mercadorias: a busca pela segurança jurídica. Prêmio de Criatividade e Inovação da RFB, 2014.

WITTEN, I. H.; FRANK, E.. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers Inc., 1999

ZADROZNY, B.; ELKAN, CHARLES. Transforming Classifier Scores into Accurate

Multiclass Probability Estimates. In proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 694-699, 2002.