Artificial Intelligence Initiatives in the Special Secretariat of Federal Revenue of Brazil

Jorge Jambeiro Filho – jorge.jambeiro@rfb.gov.br

In this document we describe the Artificial Intelligence Initiatives in the Special Secretariat of Federal Revenue of Brazil (RFB) going on in 2019. RFB encompasses both the customs department and the internal revenue service in Brazil and there are AI initiatives in both areas. One of them is a mature artificial intelligence working in national scale whose core technology was created and developed by RFB: the Customs Selection System Through Machine Learning (Sisam). Still in customs field we have other AIs in production like the expert system in the Aniita system and the Iris face recognition system. Some other initiatives are partially implanted, in tests or in advanced stages of development. Such initiatives exist both for customs and the internal service. There is also an increasing number of initiatives in their first steps.

We also have generic AI resources that have been internally developed and which are shared by customs and the internal service. Among these resources is the ContÁgil system, which includes a full featured machine learning environment. ContÁgil is available to all RFB's employees, is distributed to Brazilian state tax administrations and is used in the famous Car Wash Operation.

Near the end of the document we describe a simulator of selection strategies for audits with which we can establish the importance of high quality predictive methods and also their limits. The simulator also show some unintuitive results like the conclusion that the strategies that win the game against tax evasion do not maximize the immediate results of the audits. At last we describe the approach of RFB's General Coordination of Information Technology to handle all parallel initiatives in AI within the institution.

AI in Brazilian Customs

Customs Selection System Through Machine Learning

The Customs Selection System Through Machine Learning (Sisam) [1, 2, 3] is an artificial intelligence that, since 2014, analyzes every import declaration registered in Brazil. Sisam learns both from inspected and non inspected import declarations and for every item in each import declaration calculates the probability of about 30 types of errors, including false descriptions of goods, errors in harmonized system (HS) codes, errors in the declared countries of origin, missing import licenses, non-applicable tax regimes, wrong preferential tariff and “ex-tariff” claims and simply the use of wrong rates for the calculation of import duty, the tax on manufactured products, social contributions and anti-dumping duties. Sisam's upcoming version also handles miss-invoicing.

Besides the error probabilities, Sisam calculates the probabilities of alternative values for each field that can be wrong, analyzes the consequences of such values both for applicable taxes and administrative requirements and uses them to estimate, in Reals, the return expectation of each possible inspection. Sisam also produces natural language explanations for the presented probabilities and

expectations. These explanations are important to allow the customs officers to join the information that comes from the system to the information that already exists in their minds. The explanations are also very convenient when ethical concerns are raised [4].

We have assessed that inspecting the items that are most likely to have errors according to Sisam is 20 times more effective than making random selections and more than 30% of selections made by customs officers in Brazil are due to suggestions of this system.

User feedback is very good including assertions that: many errors that are captured by Sisam would certainly escape within the thousands of daily imports; the natural languages explanations look to have been humanly written; novice officers become productive faster and the system has induced importers to committing less errors.

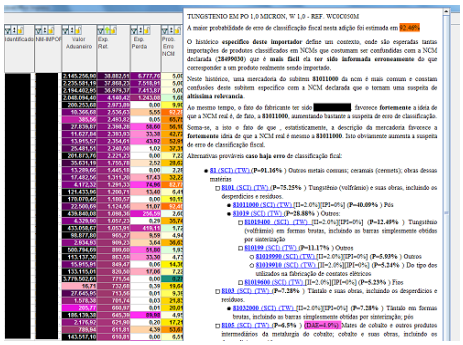

Figure 1: Sisam Main Screen

Figure 2: Presentation at WCO

Sisam's core technology has been developed in the interest of Special Secretariat of Federal Revenue of Brazil. It is essentially a set of Bayesian Networks whose conditional probability tables have been replaced by smoothing hierarchies [5]. Sisam's knowledge base is updated daily and currently contains more than 8.5 billion patterns originated from 150 million imported items. The system is implemented in Java and employs no machine learning tools or libraries.

Sisam is mainly focused on the risk of individual import operations and is typically used before and during customs clearance. It has recently been extended to generate aggregated data about importers with goal of helping post clearance revisions.

Expert systems for selection of goods for inspection

The hype over machine learning should not prevent us from using AI systems that are based on humanly created rules, the traditional and very useful expert systems. Besides Sisam, Brazilian customs officers count on Aniita [3, 6], which is a tool that gathers in a single place all information that is relevant for customs clearance and include an embedded expert system. This system points risk factors in import declarations, export declarations, express couriers, postal consignments, and export declarations. Expert systems are based on rules created by humans. They are simpler to implement, light weighted to execute and immediately scale up the application of human knowledge. They are an indispensable resource in every fraud detection domain. In our experience, the key points for successful expert systems are the flexibility of the rules, the provision of the ability of creating rules both for regional and national experts under an adequate privilege control scheme and a sharing mechanism that allows good rules to be propagated from region to region and possibly become national.

Document Mismatch Detector

The Document Missmatch Detector (BatDoc) [7], looks for mismatches between import declarations and auxiliary documents like invoices and bills of ladings, which become available, as digital images, after an import declaration is selected for inspection. It detects divergences in company names, addresses, prices, quantities, HS codes, incoterm codes and others. To do this, it applies optical character recognition to the auxiliary documents, identifies relevant fields and performs transformations to the data to handle differences in how theses fields are presented in each document. BatDoc is available to all customs officers in Brazil and is widely used.

BatDoc is implemented in Java and was tested with two optical character recognition tools: Tesseract, which is free and Abby, which is a commercial tool whose license had already been acquired by RFB. The best results were obtained with Abby, which is currently in use with BatDoc. We consider the possibility of enhancing BatDoc's text post processing abilities to reduce the pressure on the OCR tool and reduce our dependency on Abby. We also consider the possibility of testing new versions of Tesseract or other free tools.

One relevant point where BatDoc is not yet very successful is the comparison between the description of the goods in the invoices and their description in the import declarations. The former tend to be more concise and are generally presented in English. The latter are longer to satisfy some legal requirements and are always in Portuguese.

We plan to write a second version of BatDoc which will have the ability to use our past data to learn the patterns involved in this loose translations and employ them to gain accuracy. However, till now, we have not assembled a team handle the job.

Container X-Ray Image Analyzer

The AJNA system [8, 9] is focused in the analysis of scanned images of containers. AJNA is under development in the Port of Santos, the largest in Brazil, with the support of RFB's Innovation Laboratory of the State of São Paulo (Labin08). In Santos, all containers leaving or entering the country are scanned. AJNA colects the resulting images, associates them with the corresponding declarations and makes the images available to customs officers in their desks whenever convenient.

AJNA employes convolutional autoencoders to find similar images within a database that already includes more than 1.5 million past import or export operations. Suspicions are raised when the declarations associated to the images found are not compatible with the declaration being examined. In other suspicious cases, the autoencoders indicate high divergence between the image of a container and the images of other containers supposed to contain the same type of goods.

Figure 3: Arms x-Ray

The autoenconders are generated by one of AJNA's deep neural network models. A second deep learning model is used to detect containers which have been falsely declared to be empty. This simple model led to immediate results that are very easy to be confirmed visually.

Currently 3 deep learning models are being developed: one to detect the presence of drugs, one to detect the presence of weapons and one for goods classification in the Harmonized System. The first two models are expected to produce results during 2020. The third is expected to be part of a planned integration between AJNA and Sisam.

AJNA is implemented in Python and uses the TensorFlow library for deep neural networks. It also uses SciKit-Learn and random forest regressors to estimate the quantities of the goods observed in the x-ray images.

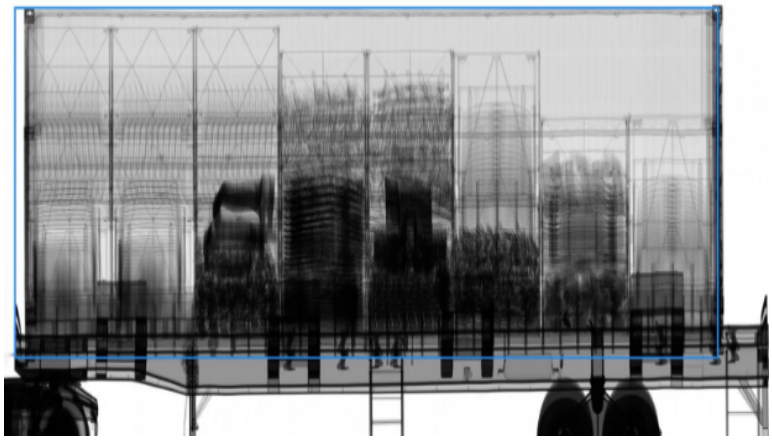

Figure 4: AJNA Image

Travelers Control System

RFB's Travelers Control Sytems (e-DBV) [10, 11] includes a facial recognition system (Iris) [10] and a GeoProcessing system (Vivii). Besides that, we are starting a partnership with the Institute of Aeronautical Technology (ITA), one of the most important universities in Brazil to develop a reinforcement learning module for (e-DBV).

In RFB, we have a preference for developing our own AI systems. This allows the exploitation of the peculiarities of our target problems and better integration to our environment. However, a face recognition system is complex and the type of faces that we need to recognize are the same human faces that are important to everybody else. Thus we decided to acquire a commercial face recognition software, which happened to be the NEC software. However, we were careful to guarantee that the supplier of the commercial solution could be replaced at any time in the future without disrupting the IRIS system as a whole.

A predefined list of passengers of interest associated to smuggling, drug trafficking, etc is kept by Iris. If a passenger is found in the list he is more likely to be selected for inspection. Besides that, once the identity of a passenger is confirmed, a lot of information about him or her can be retrieved from all RFB's systems. That includes data about prior travels, annual income and related people. An expert system embedded in e-DBV then estimates the risk level for the traveler.

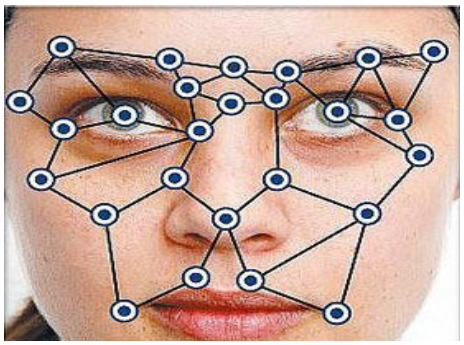

Figure 5: Face in analysis by Iris

The Vivii module is under development in the Viracopos airport, one of the most important in Brazil, with the support of RFB's Labin08. Vivii analyzes travel routes and displays information in concentration maps. So, customs officers can see, for example, outliers among the addresses of the passengers in a certain flight and see the most common origins and destinies of travelers caught carrying drugs. Vivii is still under development, but it is expected to use machine learning over the mapped data to calculate risk levels automatically.

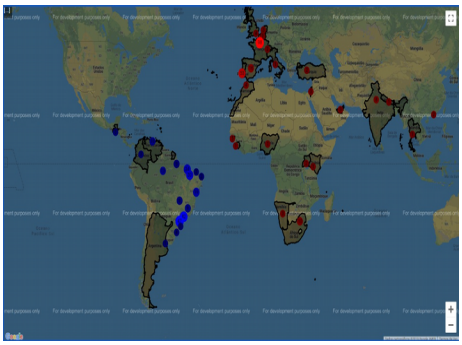

Figure 6: Vivii drugs destinies concentration map

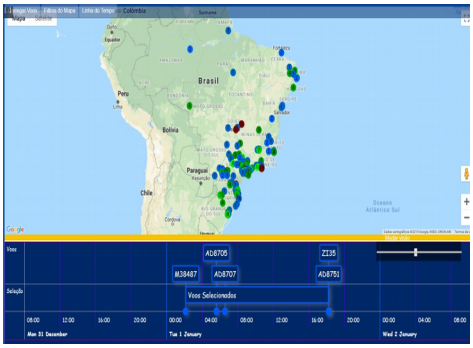

Figure 7: Vivii passenger address map

In several selection problems there is an issue related to the balance between exploration and exploitation. Selecting those whose risks are the greatest according to our current data produces the best immediate results, but prevents us from learning new niches of crimes or infractions. Recently we started a partnership with the ITA to develop a reinforced leaning module based on the Contextual Multi-Armed Bandit (MAB) algorithm, whose goal is to handle the exploitation versus exploration problem in the context of travelers control. If this research is successful we intend to apply the solution to several other fields.

Time series to Monitor Importers and Exporters Behavior

Legislative changes like the increase or reduction of aliquots, the establishment of anti-dumping duties, the imposition of countervailing duties and incentive policies may affect the behavior of importers and exporters. Sometimes, the behavioral change is an attempt to scape or exploit the new legislation in a fraudulent manner.

A project to monitor importers and exporters behavior using time series has just been started. Our expectation is to compare the predictions of the system to actual behavior and select the taxpayers which deviate more from expectations for auditing procedures.

This project is, for now, a term paper of one of RFB's employees graduation in statistics. If it is successful it may become part of the Brazilian customs risk management ecosystem.

AI in the Brazilian Internal Revenue Service

Neighborhood Black Sheep System

The Neighborhood Black Sheep System is under development in RFB's Innovation Laboratory of the States of Ceará, Maranhão and Piauí (Labin03) and is being tested in the same states before becoming national. The goal of the system is to employ geodata mining techniques to identify individuals with economic-fiscal characteristics (income, equity or financial movement) that differ from those living in their vicinity (spatial outlier detection).

The analysis of economic data together with geographic data is intended to identify, for example, taxpayers residing in regions of high economic standard, as identified by the pattern of taxpayers residing in the same area, who do not declare compatible income or equity, what is a strong indication of heritage concealment. In the opposite direction, the system detects those who have a great declared heritage but reside in less affluent areas, a strong indication of third-party fraudulent intervention (usage of stooges).

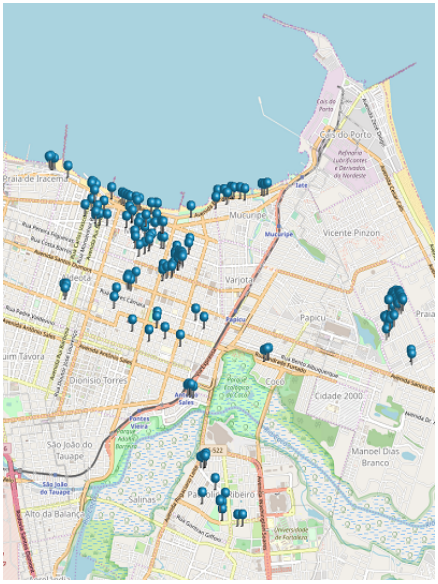

Figure 8: Black Sheep indications

CPF Issue Fraud Detection

The taxpayer identification number for natural people (CPF), administered by the RFB, is adopted by several public and private entities as the basic identification number of the citizen in Brazil. This is a necessary document, for example, to participate in social programs such as Bolsa Família, to request public services such as telephone and electricity, to open accounts, to obtain loans and to other services from the banking system. Registrations in the CPF database are made by a large network of service channels, formed, since the 1990s, by affiliated entities such as post offices, official banks and, in recent years, also by state agencies and notaries.

As the visibility and importance of the CPF grew, so did the risks of attacks on the integrity of the database. People interested in circumventing the controls of social programs and the banking system try to get registration numbers by presenting false information or documents in one or some of the many service units. Simple problems are detected by the old register system when someone attempts to insert data, but more complex cases are detected manually at a later time by RFB employees. The increasing

volume of application requests, however, has exceeded our operational ability to detect fraud attempts in a timely manner.

A machine leaning system is under development in Labin03 for automatic detection of CPF issuance fraud. It aims to statistically identify patterns in already encountered fraud cases and quickly apply these patterns to identify new cases. In this way, the knowledge acquired by the RFB teams in the treatment of fraud cases, regional or national, from the simplest to the most complex, will be able to be applied uniformly and efficiently to the register as a whole.

In September 2019, a pilot version of the system was applied to 1.7 million registers in the CPF database, all created by a single affiliated entity. Initially more than 70.000 high risk registers were found and they are currently under examination.

ChatBots

ChatBots are becoming common in the world and they can be very useful to RFB both to assist internal and external public. To build a chatbot we need a natural language understanding module and dialog manager module. These modules can be encapsulated in web-services and isolated from our main business. Considering this possibility, we started to probe commercial solutions from IBM, Microsoft and SAS for these modules. We ran into the facts that current legislation prevents the use of the clouds of private companies to host services that handle data related to individual taxpayers and the fact that on premise solutions turned out to be too expensive.

Since many companies, even small ones, are experience success developing chatbots using free tools we decided to investigate them. Currently we plan to use Rasa (RasaNLU + RasaCore), which is a very popular tool in the development of chatbots. Our first experiments showed Rasa to be a viable alternative, but other tools, like, BotPress may be investigated latter. Since these experiments took place, two state companies, Dataprev and Serpro offered chatbot solutions to RFB and we are considering their offers too.

Most chatbots that can be interesting for RFB depend a lot on the integration with other systems. For example, if a taxpayer asks about his own debits, this must be assessed from several RFB databases. Since RFB systems are hundreds and many of them do not offer APIs designed to be used by other systems, we believe that these integrations will be the most difficult part in the development of chatbots. The RFB's current systems development guideline considers the creation of mechanisms that facilitate machine-to-machine communication (API) as explained in the end of this document.

There is, however, one communication channel, FaleConosco, that was designed to help taxpayers without requiring them to be authenticated. This means that FaleConosco can only admit generic questions that don't depend on private information. Therefore, we chose FaleConosco to be our first chatbot endeavor. We expect results within 2020.

Acceleration of administrative processes

Many infringement notices result in appeals that need to be analyzed by RFB employees. These analyzes take time and currently there are about 150000 administrative processes going on while the taxpayers wait for a decision. Most of these process were originated from the tax on income of natural persons.

Artificial intelligences are being developed to help RFB to accelerate the analysis of old and new administrative processes. The two groups are very different and require different measures.

Today, most taxpayers already appeal against infringement notices using a software called e-Defesa. This software lists hundreds of common arguments and most taxpayers make their appeals just picking some of them. The number of taxpayers that pick the “other” option and write their own defense in natural language is still big. Our approach to handle that is to use clustering algorithms to reveal common arguments that had not been listed yet and list them as soon as possible. We also consider using a chatbot to make it easier for taxpayers to pick the best arguments.

We also intend to make the analyzes of taxpayers defenses more structured, allowing superior administrative instances to understand what was done in the first instances without depending of sophisticated natural language abilities. In this context, the role of AI will be to make predictions based on past results and grant some taxpayer requests without consuming human resources. Besides that, the new processes will soon also count on electronic invoices which will eliminate the need to exam medical receipts, which today are frequently handwritten and are the key elements in the old processes.

The old case files are completely written in natural language and their associated decisions, almost always, depend on analyzing elements of proof that are presented as digital images like handwritten receipts. We decided to handle the natural language written files using AI, but not to handle the images of elements of proof. Our first tests using optical character recognition, both using free or commercial tools, were poor for handwritten receipts. Since these receipts are doomed to disappear, we considered that developing an AI that could understand them was not worth it.

The first goal of the AI module that will handle the old case files is to transform the arguments presented in natural language into a list of predefined arguments that resemble the ones of e-Defesa. To do that, we first separated the files according to the types of infringements that gave origin to them. The infringements are known in structured form even for the old processes. Then we used a team of specialists to label the processes according to the arguments presented, with each process possibly receiving more than one label. After that we trained several supervised learning algorithms to predict the labels and tested them in a separate set. Results were mixed. For the most common infringements results were good (though not excellent), but for the least common infringements, the number of labeled files was clearly insufficient.

We are currently testing more algorithms and enhancing our data preparation, what is leading to significant improvements in the results. We are also building an interface to show the specialists the most ambiguous files and let them label these specific files. We expect this method, which is a type of active learning, to lead to improvements without requiring to much labeled data.

Information Extraction from Judicial Case Files

Judicial Case Files can contain important information for RFB whether RFB is directly involved in the judicial process or not. When RFB is involved in a process, our main interest is to identify which arguments typically lead to a win or to a loss. When RFB is not involved directly, it is still interested because taxes may apply to court determined compensations and because the judicial processes reveal relevant relationships among taxpayers. One of RFB's employees has just started an MSc program in applied data science in the Data ScienceTech Institute in France with the goal of extracting information from judicial case files.

Selection of personal income tax returns for examination

After the income tax declarations are received, a simple preselection for audit is applied to taxpayers. The preselection is mostly based on the detection of incompatibilities between the declaration presented by the taxpayer whose risk is under assessment and declarations provided by other taxpayers that happened to be involved in financial transactions with the taxpayer in focus and in the observation of violations of some expectations defined by our specialists. The preselected taxpayers are grouped according to their similarities and the history of audits of members of each group is them considered in statistical tests that lead to clearance of the lower risk taxpayers.

The current process does involve some manual steps that slows it down. The present effort is to change the system in a way that would make it fully automatic and set up the path to the use of more sophisticated AI techniques.

Rejection of Refund Claims

The machine learning project for rejection of refund claims is also in its first steps. The project was started by Labin08 in an innovative way. Instead of a lot of discussion and planning, a hackathon was organized. Labin08 prepared an obfuscated dataset related to refund claims and made it available for all RFB's employees. The labels, which are the percentages of rejection of the refund claims, were removed from some lines in the dataset. The goal of the participants is to predict the removed labels. Whoever gets closer to the correct answers will get an award. Latter, when deeper discussions about the project take place, these individual attempts to make the predictions will be considered as a start.

Taxpayer Profiling

Generating a taxpayer profile is about to become the first project of RFB's Innovation Laboratory of the State of Rio de Janeiro (Labin07). Such a profile is planned to summarize the data about a taxpayer and carry information for one area of RFB's activity to all others.

The profile has initially been specified to be a set of point of view profiles. Each point of view profile is a vector where each entry is the degree of pertinence of the taxpayer to the cluster of taxpayers that corresponds to the entry position.

So, to create a point of view we can just cluster taxpayers using partial data about them. For example, we can create clusters based on the HS codes of the goods that taxpayers sell in the internal market, create clusters based on the services that they offer, create clusters based on exports, on imports, on refund claims, etc. Once we have the set of clusters that corresponds a certain point of view, we will also have the vector where each entry informs how well a taxpayer fits in each cluster.

To gain precision, we don't need to restrain ourselves to generating one point of view for each dataset. We intent to generate more than one point of view per dataset using non redundant clustering techniques.

One important factor in the profile of any taxpayers is how it relates to other taxpayers. We could have, for example, the cluster of small farmers who sell groceries to big retailers. However, this requires other taxpayers to be classified as big retailers. So, at least partially, we need to generate the profiles of all taxpayers at once in a big convergence process.

Once we have the full profile of a taxpayer, we can use word clouds to make them understandable. The word clouds can be generated mining company names, web sites, contracts, tax declarations,

infringement notices and other texts related, not only to that specific taxpayer, but also to other taxpayers with similar profiles. The key point here is that some words that don't appear in any documents directly related to a taxpayer may be found to be important in their profiles.

Though we already had an experience in the direction of a taxpayer profile [14], we consider this work to be in a planning stage and the data science team is not yet closed.

Shared resources between customs and the internal revenue service

RFB's data lake (ReceitaData)

RFB has created a unified repository for all its data. This data lake is called ReceitaData and is based on the Hadoop architecture. ReceitaData minimizes the major problem in data science, the access to the data itself. Data from various sources are loaded in ReceitaData with periodicities that vary from 1 second to 3 days. It is our plan to have all data to which RFB has access available in this data lake.

Data scientists can access the data lake using SQL through HUE, using Python code through Jupyter Notebooks and even send Java packages to be executed by Hadoop MapReduce engine or Spark.

ReceitaData is our general solution to access offline data. That does not prevent our artificial intelligences from needing access to online data and to get it directly from our transactional systems. However, operations involving large amount of data can always be executed in the data lake.

Machine Learning Framework in the ContÁgil System

The ContÁgil system [13] is a data retrieving and data analysis tool developed by the Special Secretariat of Federal Revenue of Brazil that offers hundreds of features and is widely used within its boundaries. ContÁgil is also distributed to state tax administrations in Brazil and is used by the Car Wash Operation task force.

One of ContÁgil's resources is a full featured machine learning environment that includes most common supervised learning algorithms like decision trees, naïve Bayes, support vector machines and deep neural networks. It also includes algorithms for clustering, outlier detection, topic discovery and co-reference resolution.

This ML framework is available to all RFB's employes and can be used interactively through a graphic interface. All ContÀgil functions are also available through scripts that can be created visually or written in Javascript or Python. A few thousand ContÀgil scripts have already been created by RFB's community. ContÀgil interfaces with other ML tools like Weka, R, DL4J and Neo4J allowing its users to benefit from the extra features of theses tools without leaving the familiar ContÁgil environment. ContÁgil itself is written in Java.

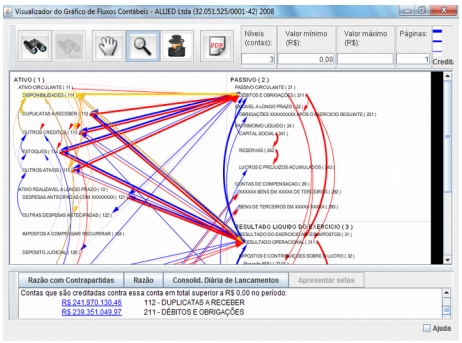

Figure 9: ContÀgil analyzing an accounting book

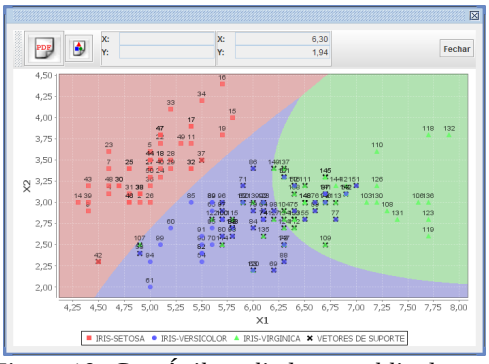

Figure 10: ContÁgil applied to a public dataset for demonstration

The ability to retrieve data and also to analyze it, makes ContÁgil a very powerful tool. It also act as a platform for other well known systems in RFB like Farol, which automates repetitive tasks involving accesses to RFB systems through the simulation of an extremely fast human user, frequently performing in one hour what would otherwise take a week.

ContÁgil's ML framework is integrated to RFB's Hadoop data lake and can run algorithms within the lake environment to gain efficiency. Such integration involves tools like Spark, Hbase, HDFS, Impala and Hive. Is can also retrieve information from most RFB's transactional systems directly and read dozens of different types of files containing relevant information like account books and invoice sets.

One very useful feature of ContÀgil is its social network analysis tool. ContÁgil scans the various datasources to which it has access and builds a giant graph representing entities like people, companies and their relationships. The graph can be filtered and visualized in different ways. The user can search for specific patterns in the graph and run different algorithms to find minimum paths, spanning trees and key players. This tool the most frequently used to dismount big fraud schemes like it was done int the Car Wash operation.

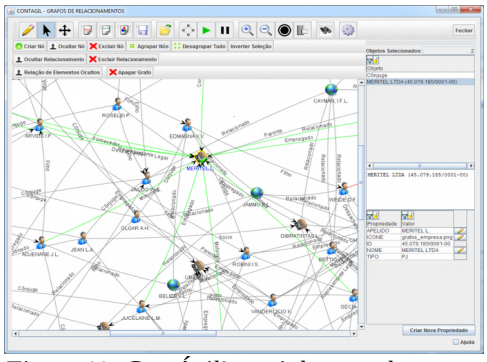

Figure 11: ContÁgil's social network analysis tool

GeoRFB

GeoRFB is a project whose goal is to provide geoprocessing resources to other projects like Vivii and Black Sheep. It was started as a combined effort between Labin03 and Labin08, when Google decided to eliminate the free quotas for GoogleMaps usages. Google maps were being used both by Vivvi and Black Sheep and they were both affected. We considered paying for the use o Google Maps, but, with

restricted resources and worries about privacy, we decided to host our own Open Street Maps Tile Server. Our plan is to concentrate all geoprocessing requirements that are common to several projects under GeoRFB.

Simulator of selection strategies for audits

RFB is investing a lot in the development of predictive methods using AI, but can a predictive method really win the game against tax evasion without association with an adequate selection strategy? No, it cannot. Should the goal of a selection strategy for audits, powered by a predictive method, be to minimize the percentage of audits without findings? Also not.

We built a simulator of selection strategies [14] for audits that creates a virtual environment where taxpayers decide to evade their taxes or not according to the probability of being audited that they estimate for themselves based on information that they gather from their neighbors in the simulation. The probability estimate is combined with the financial advantage that they will experience in case they evade their taxes and are not selected for an audit and with the financial loss that they will suffer in case they are audited and punished. Such combination is their return expectation for evading taxes. The more positive the expectation is, the more likely a tax payer is to incur in tax evasion.

In this simulated universe a predictive method is applied to all taxpayers and it indicates some of them as evaders and others as compliant. Since we are in a simulation where everything is known, the quality of the predictive method can be arbitrated through the specification of its statistical sensibility and its statistical specificity. The results of the predictive method become available to a selection policy that can, for example, spread audits over the whole environment or concentrate them in some groups. The policy can also only select positive examples (the ones indicated as evaders by the predictive method) or to make random selections. It can audit taxpayers without warning and punish the evaders with maximum force or give them the opportunity correct themselves (auto regularization) with discounts on fines or even with complete pardon with the hope that this will spare RFB's limited workforce.

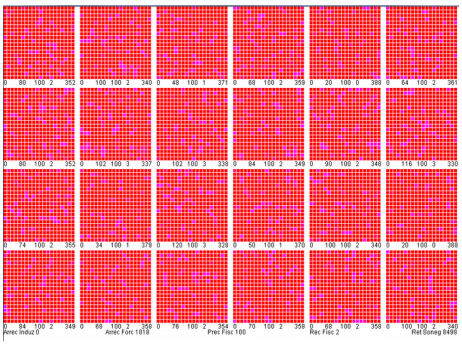

Figure 12: Simulation where a "scattered audits strategy" is going on.

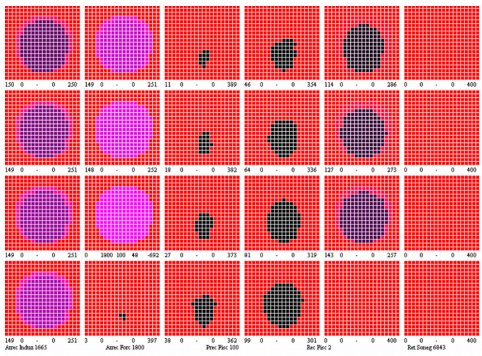

Figure 13: Simulation where "focus on worse groups strategy" is going on.

As an effect of the chosen policy, tax evasion can shrink or grow. A successful policy is one that leads to an environment that is almost free of tax evasion and a failed strategy is one that leads to an environment where it is ubiquitous.

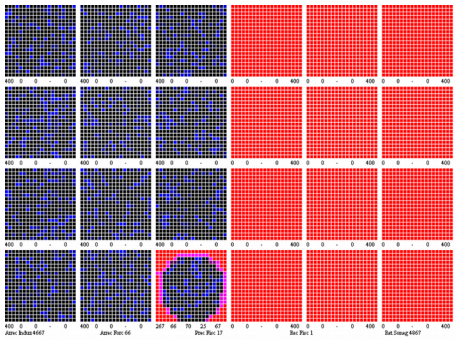

Figure 14: Simulation where a "Clean some groups strategy" is going on.

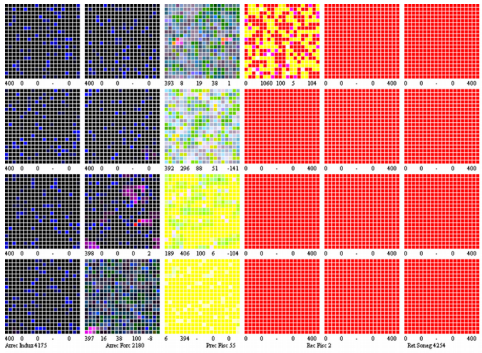

Figure 15: Simulation where auto regularization is exploited

We found, for example, that even if the predictive method is absolutely perfect (free of both false positives and negatives) naïve selection strategies are badly beaten by tax evasion while the smartest ones can succeed even if an imperfect predictive method is employed. On the other hand, when we make the predictive method bad enough any selection strategy fails. We also found that policies that try to minimize the percentage of audits without findings (audits of compliant taxpayers) fail to progressively clean the environment and eventually loose the game. On the other hand, successful strategies worry about the return expectation of tax evaders and keep it negative at least within some controlled groups. The number of controlled groups is increased whenever possible without letting evasion contaminate the groups that are already clean.

Our plan is to collect more data to make simulations more realistic before they become more widely used in RFB. However, some RFB operations are already considering the concepts involved in the simulations, one of them, the Cartórios Operation, resulted in more the 250 million reals in unpaid taxes.

Data science diffusion in RFB

In 2019, through Labin08, in association with the Foundation Institute of Administration of the University of São Paulo (FIA/USP), RFB promoted a course of data science where 40 new data scientists were formed. Python courses, SQL courses and Java courses also took place this year. Courses of ContÁgil, our most popular data analysis tools were also ministered. RFB is also hosting data science events and seminars, like the PyData that took place in Brasilia, in October 2019 and the World Customs Organization Data Science Seminar for Latin America, that took place in São Paulo, also in October 2019. Internal seminars are also increasing in number. In addition, two hundred RFB employees began, in November 2019, the postgraduate course in Data Science and Big Data offered by the Pontifical Catholic University of Minas Gerais (PUC Minas).

Management of AI and Data Science in RFB

Artificial intelligence and data science are not new to RFB. There is very relevant research in the field within RFB since the General Coordination of Customs Administration (Coana) started the Harpia project, an agreement between RFB and two of the most important Brazilian Universities the University of Campinas (Unicamp) and the Technological Institute of Aeronautics (ITA) in 2005. After the end of the agreement in 2008, RFB proceeded researching AI using its own staff.

The development of Sisam gave us the experience of leading one artificial intelligence project from the first discussions to national scale production and maintenance. Since then, several artificial intelligence initiatives have been started in the institution. However, RFB is moving from an environment of

individual initiatives in artificial intelligence to a coordinated and generalized application and development of technologies in the field. In this coordination, we are still newbies.

RFB's General Coordination of Information Technology (Cotec) has recently created the Center of Data Analysis and the Center of Excellence in Artificial Intelligence. The former uses existing resources to handle data analysis requests while the latter is focused on creating new resources.

The two centers are references that handle part of the work directly, but also harmonize the various initiatives of our vibrant data science community. Innovation laboratories are being created in all ten fiscal regions in Brazil and two of them are already well structured (Labin03 and Labin08). Several general coordinations a getting involved with data science and more and more data scientists are being formed.

Cotec has decided to support local initiatives, supervising them from the very beginning with the goal of preventing projects from overlapping, factorizing their common factors in separate projects, transferring knowledge from one project to the others, ratifying good projects to sponsors within the institution and guaranteeing that all products will eventually be able to be used nationally. Cotec is also involved in assembling multidisciplinary teams for each project including data scientists and business specialists.

Another important decision was to determine that all new transaction systems must be build with data analysis in mind. This means that they should offer APIs that allow other systems to access them and should register not only the data that is necessary to achieve immediate results like the application of fines, but also the data that will allow AI systems to learn. That imposes a cost on the fiscal-auditors, but allows their work to be replicated in large scale.

One, maybe initially unintended, boosting factor for our data science development is RFB's Creativity and Innovation Award that is conceded annually since 2002. Any employee can submit a monograph to the competition describing or proposing an innovation in any field of RFB's interest. Works in data science are frequently among the five winners, highlighting the importance of the field for RFB.

One key point for the decision of having many data science projects going on in parallel is the perception that they don't need to be all successful. When projects start, they don't demand much material investment. At the same time, RFB has many engineers, computer scientists, statistics and mathematics formed in the best Brazilian universities among its staff. We can guarantee that each of them will be productive in a small scale forcing them into usual work or risk the possibility that they will build data science projects with national impact. In the past, the first option would be chosen, but RFB's new policy is to break that tradition.

References

[1] Jambeiro Filho, Jorge. Artificial Intelligence in the Customs Selection System through Machine Learning (SISAM). Prêmio de Criatividade e Inovação da RFB, 2015.

[2] Jambeiro Filho, Jorge. Artificial intelligence in Brazil’s Customs in Study Report on Disruptive Technologies, pages 75-78. World Customs Organization. Rue du Marché 30. B-1210 Brussels.

[3] Coutinho, Gustavo; Jambeiro Filho, Jorge. Brazil’s New Integrated Risk Management Solutions. World Customs Organization News, 86. Junho, 2018.

[4] Köche, Rafael. L'intelligenza artificiale a servizio della fiscalità: il Sistema di selezione doganale attraverso l’apprendimento automatico (SISAM). Il ragionamento giuridico nell’era dell’intelligenza artificiale. 15 de novembro de 2018, Florença, 2018.

[5] Jambeiro Filho, Jorge; Jacques Wainer. HPB: A model for handling BN nodes with high cardinality parents. Journal of Machine Learning Research (JMLR), 9:2141–2170, 2008.

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

[7] Barbosa, Diego de Borba. Batimento Automatizado de Documentos na Importação – BatDoc. Prêmio de Criatividade e Inovação da RFB, 2016.

[8] Brasílico, Ivan. AJNA – Plataforma de Visão Computacional e Aprendizado de Máquina. Prêmio de Criatividade e Inovação da RFB, 2017.

[9] Brazílico, Ivan. AJNA - X-Ray Images for Customs. 2019. Available at https://ivanbrasilico.github.io/ajna_docs/

[10] Moraes, Felipe Mendes; Jezini Netto, Felipe. A aplicação de análise de dados em informações avançadas de passageiros internacionais e sua relevância para a fiscalização aduaneira brasileira. Prêmio de Criatividade e Inovação da RFB, 2017

[11] Thompson, Ronald Cesar et al. Projeto IRIS - Reconhecimento Facial de Viajantes. Prêmio de Criatividade e Inovação da RFB, 2016.

[12] Carvalho, Leonardo. Análise de Setores Econômicos, Relatório Técnico – COPES Brasília, 2015.

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

[14] Jambeiro Filho, Jorge. Gerência da Expectativa de Retorno do Sonegador e Simulação de Estratégias Fiscais. Prêmio de Criatividade e Inovação da RFB, 2019.