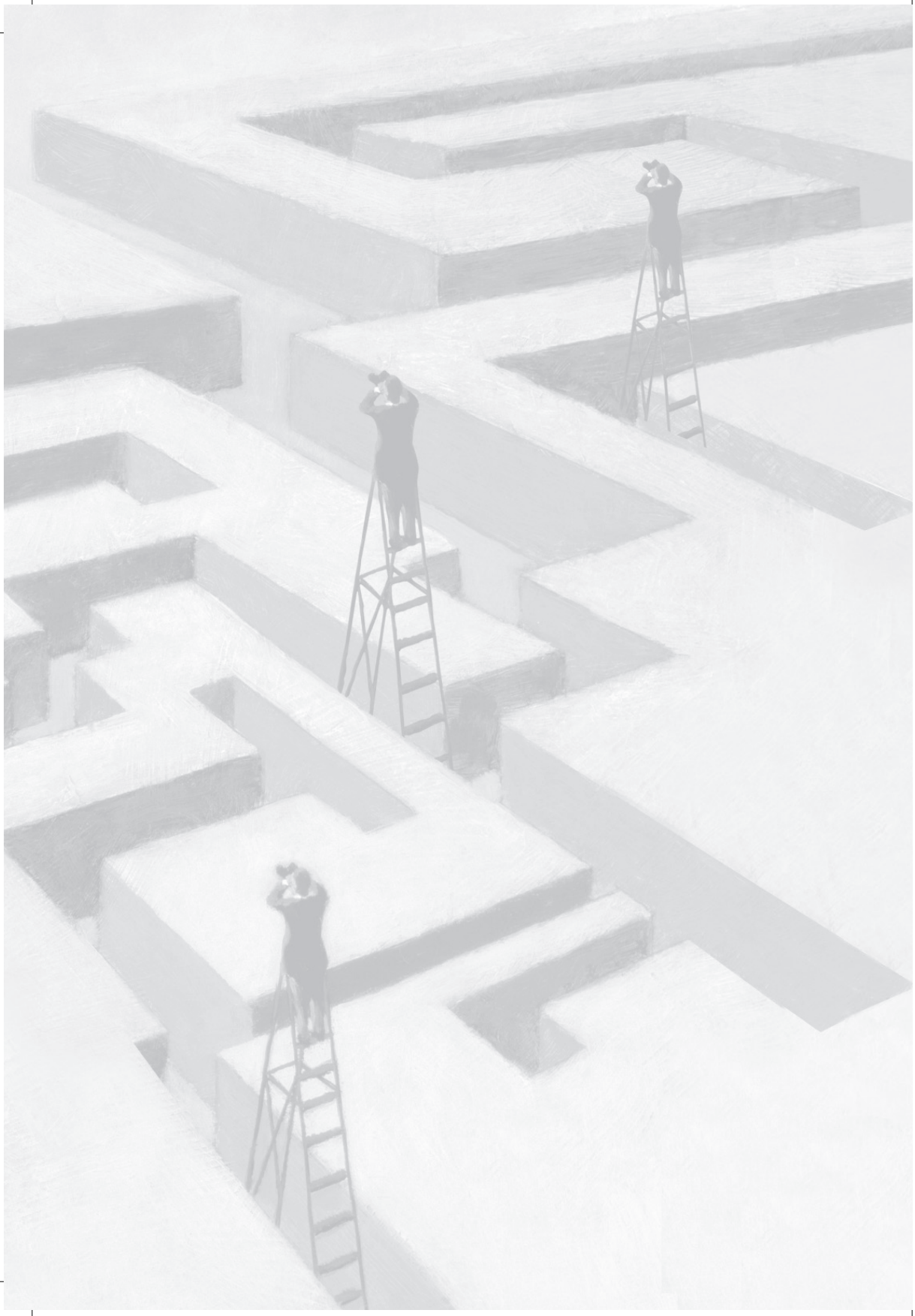
Inteligência Artificial no Sistema de Seleção Aduaneira por Aprendizado de Máquina

1º Lugar

JORGE EDUARDO DE SCHOUCAIR JAMBEIRO FILHO*

* Engenheiro de Computação, doutor em Inteligência Artificial (IA) pela Universidade Estadual de Campinas (Unicamp). Auditor-Fiscal da Receita Federal do Brasil, Delegacia de Barueri.



Inteligência Artificial no Sistema de Seleção Aduaneira por Aprendizado de Máquina

Resumo

Neste trabalho, descrevemos o módulo de inteligência artificial (IA) do Sistema de Seleção Aduaneira por Aprendizado de Máquina (Sisam), hoje em uso em todas as unidades aduaneiras da Receita Federal do Brasil (RFB). Mostramos que a implantação deste sistema só foi possível graças à solução de mais de uma dezena de desafios técnicos e inovações científicas de interesse direto da RFB. Fazendo isto, caracterizamos esse processo não como um exemplo da já valorizada aplicação de técnicas de mineração de dados na RFB, mas como um caso inovador de pesquisa e desenvolvimento tecnológico dentro da instituição. Também mostramos que o Sisam traz ganhos de desempenho reais na detecção de várias infrações características do despacho aduaneiro, com especial destaque para o importante e complexo erro de classificação fiscal. Apresentamos a interface da Sisam com os Auditores-Fiscais da Receita Federal do Brasil (AFRFBs) e destacamos a forma com que ela permite que o conhecimento de máquina e o humano sejam unidos com agilidade suficiente para tomada de decisões tempestivas. Apontamos o fato de o Sisam ser a primeira inteligência artificial *on-line* desenvolvida na RFB e a primeira de uso totalmente generalizado em sua área de atuação. Também mostramos que a tecnologia do Sisam não tem interesse restrito a seu objetivo original, já tendo dado origem a uma razoável lista de trabalhos futuros e sido efetivamente aplicada em dois outros contextos, ambos na área de tributos internos.

A) Objetivos básicos

Por meio de uma inteligência artificial (IA) que aprende com o histórico de Declarações de Importação (DIs) ajudar a Receita Federal do Brasil (RFB) a reduzir o percentual de mercadorias verificadas no despacho aduaneiro de importação e a, concomitantemente, reduzir a evasão fiscal na importação e o descumprimento de exigências administrativas.

B) Metodologia utilizada

Idealizar uma inteligência artificial que aprenda com o histórico de declarações de importação do Sistema Integrado de Comércio Exterior (Siscomex) e que atenda aos seguintes requisitos:

- calcular a probabilidade da presença de erros de classificação fiscal, erros de origem, erros em regimes tributários, erros em acordos tarifários, erros de falta de licenciamento e erros nas alíquotas de II, IPI, PIS, Cofins e *Antidumping*;
- realizar os cálculos das probabilidades destes erros para cada mercadoria (item), de cada adição, de cada declaração de importação registrada no Brasil;
- calcular a probabilidade de cada valor correto possível para cada campo suspeito de erro;
- calcular o impacto tributário e não tributário de cada valor correto possível e assim obter a expectativa de retorno de cada verificação possível para RFB;
- realizar estes cálculos, de forma rápida, o suficiente para que os resultados estejam disponíveis enquanto ainda é possível redirecionar a DI em questão;
- poder ser atualizada diariamente com as novas declarações de importação;
- aprender rapidamente de modo a barrar tentativas de fraude semelhantes em diversos pontos do país em intervalos curtos de tempo;

- aprender com as infrações detectadas pelos fiscais nas DIs que forem verificadas;
- aprender os comportamentos típicos e atípicos dos importadores mesmo com DIs liberadas em canal verde;
- adaptar automaticamente a seleção à carga e à mão de obra disponíveis;
- não se deixar enganar por um comportamento errado de um importador, mesmo que este comportamento se repita;
- não permitir que os importadores consigam prever o comportamento do sistema e assim descubram como enganá-lo;
- apresentar taxas de acerto elevadas;
- interagir eficientemente com os fiscais, inclusive, gerando explicações em linguagem natural de modo a permitir que eles se beneficiem do sistema quando ele acerta sem perder muito tempo quando ele erra;
- projetar, implementar e testar a inteligência artificial idealizada;
- implantar o sistema e colocá-lo em efetivo funcionamento em todas as unidades aduaneiras da RFB;
- fazer correções e melhorias a partir dos resultados experimentais; e
- analisar o desempenho do Sisam aplicando o sistema isolado a uma amostra de 665 mil itens já verificados por fiscais e comparar suas previsões com os resultados reais. Também reportar a efetividade do uso do Sisam, em produção, pelos fiscais que atuam na seleção para verificação no despacho aduaneiro e próprio despacho aduaneiro.

C) Adequação do trabalho aos critérios de julgamento

I – Criatividade e inovação

A aplicação das melhores tecnologias de mineração de dados disponíveis já foi reconhecida como altamente relevante para RFB, sendo

inclusive a parte central de dois trabalhos já agraciados com prêmios de criatividade e inovação da RFB.

O presente trabalho vai além e inova ao aplicar uma tecnologia concebida especificamente para atender a um interesse da RFB. Assim, a instituição é colocada na ponta, não apenas do uso das tecnologias deste campo, mas também de seu desenvolvimento.

Neste trabalho, resolvemos mais de uma dezena de desafios técnicos e elaboramos uma arquitetura única, capaz de satisfazer a todos os requisitos oriundos do complexo ambiente de dados da RFB e da meta de ter uma inteligência artificial *on-line* processando tempestivamente todas as declarações de importação registradas no Brasil e interagindo eficientemente com os Auditores-Fiscais da Receita Federal do Brasil (AFRFBs).

II – Relação custos versus *benefícios*

O Sisam emprega para desenvolvimento, homologação e produção sete servidores, cada um com 12 núcleos de processamento e 64 GB de RAM. Ao todo cerca de 7 TB de espaço em disco são empregados. O preço deste equipamento flutua, mas sua aquisição fica em torno de R$ 300.000,00.

O mecanismo de aprendizado de máquina do Sisam é automático e nenhuma regra de seleção ou regra de estimativa de risco precisa ser criada por pessoas. O Sisam não requer que os fiscais realizem nenhum procedimento que já não realizassem normalmente para aprender, visto que se baseia nas retificações das DIs.

Ele também aprende com as DIs liberadas sem verificação, capturando os comportamentos típicos e atípicos dos importadores sem que eles sequer precisem saber que o Sisam existe. Nenhuma adaptação da parte deles é ou foi necessária, dispensando custos com atendimento ao contribuinte e campanhas de informação e conscientização.

As análises produzidas pelo Sisam são apresentadas dentro do sistema Aniita, já usado pelos fiscais da RFB. Assim, os fiscais não são forçados a consultar um sistema a mais para realizar suas atividades e recebem informações do Sisam juntamente com dados que já teriam que examinar de qualquer modo. Isto novamente reduz custos.

O primeiro benefício do Sisam é a melhoria da qualidade da seleção para verificação aduaneira, principalmente com objetivo de capturar erros difíceis, como o erro de classificação fiscal.

O Sisam também melhora a efetividade da própria verificação no despacho, pois fornece relatórios interativos para todos os itens de todas as adições de todas as DIs, mesmo aquelas selecionadas por critérios que nada tem a ver com o próprio Sisam.

Outro benefício é aumentar a uniformidade de tratamento no despacho, tornando inútil para importadores mudar sua carga para Unidades da Receita Federal (URFs) que ainda não conheçam suas infrações.

O sistema não apenas captura sinais de presença de infração, mas também os sinais de que ela está ausente. Isto reduz o desperdício de mão de obra com verificações repetitivas e libera os fiscais para usar sua capacidade em análises mais produtivas.

Como consequência de tudo isto, o Sisam aumenta a sensação de presença fiscal e, portanto, induz o comportamento espontâneo.

Por fim, há o benefício do potencial uso da tecnologia do Sisam em outras áreas. Ilustramos este potencial com uma lista de trabalhos futuros já encampados pela Coana e com dois trabalhos derivados já em produção, dentro do sistema Contágil, ambos na área de tributos internos.

III – Aumento de produtividade

Medimos o desempenho do Sisam e mostramos que ele efetivamente gera ganhos de produtividade na seleção para verificação no despacho aduaneiro e no próprio despacho aduaneiro. Mostramos que a tecnologia do Sisam pode ser aplicada em outras áreas, gerando mais ganhos de produtividade.

IV – Viabilidade de implementação

O Sisam já está em uso em todo Brasil.

V – Melhoria da qualidade dos serviços prestados e dos resultados estratégicos

O impacto do Sisam está relacionado aos seguintes objetivos estratégicos:

Perspectiva de resultados:

1) Aproximar a arrecadação efetiva da potencial.
2) Elevar o cumprimento espontâneo das obrigações tributárias e aduaneiras.
3) Contribuir para o fortalecimento do comércio exterior e para a proteção da sociedade.
4) Aumentar a percepção de equidade na atuação da instituição.
5) Fortalecer a imagem da instituição perante a sociedade.

Perspectiva de processos internos:

6) Reduzir o tempo entre o vencimento do tributo e o seu recolhimento.
7) Elevar a percepção de risco e a presença fiscal.
8) Aumentar a efetividade e segurança dos processos aduaneiros.
14) Conhecer o perfil integral do contribuinte.

Perspectiva de pessoas e recursos:

17) Desenvolver competências, integrar e valorizar pessoas.
18) Adequar o quadro de pessoal às necessidades institucionais.
19) Assegurar soluções de tecnologia da informação (TI) integradas e tempestivas.
20) Adequar a infraestrutura física e tecnológica às necessidades institucionais.

Na cadeia de valor da RFB, os processos de trabalho a que se refere a monografia são:

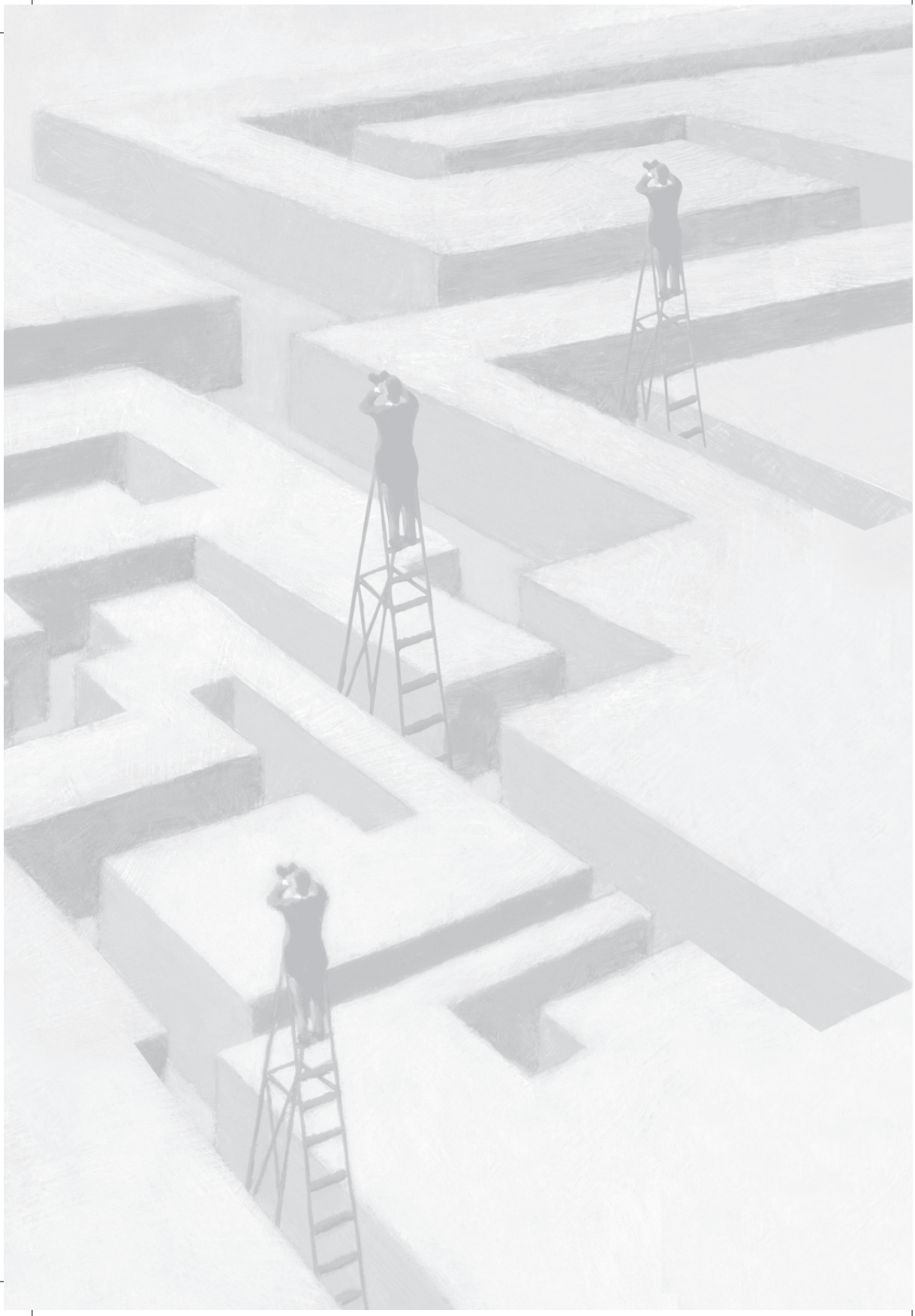
- ✓ Fiscalização tributária e combate a ilícitos.
- ✓ Controle aduaneiro.
- ✓ Governança de TI.

D) Principais lições aprendidas

Comprovamos ser possível desenvolver tecnologia na área de mineração de dados dentro do âmbito da RFB e aferir benefícios efetivos com isto. A instituição já havia apostado nesta possibilidade ao criar o projeto Harpia, envolvendo um convênio com a Universidade Estadual de Campinas (Unicamp) e com o Instituto Tecnológico de Aeronáutica (ITA). O fim deste convênio deixou fortes dúvidas quanto a se tal possibilidade era real, mas agora usando apenas pessoal próprio, acabamos com essa dúvida.

Também aprendemos a importância da interação entre a inteligência artificial e o usuário humano, dada a grande melhoria na receptividade do sistema quando ele passou a gerar explicações em linguagem natural.

O aprendizado técnico também foi muito grande, dada a quantidade de requisitos desafiadores do módulo de IA do Sisam.



Inteligência artificial no Sistema de Seleção Aduaneira por Aprendizado de Máquina

1 Introdução

O Sistema de Seleção Aduaneira por Aprendizado de Máquina (Sisam) é uma inteligência artificial (RUSSEL; NORVING, 2013), que aprende com o histórico de declarações de importação (DIs) e tem o objetivo de ajudar a Receita Federal do Brasil (RFB) a reduzir o percentual de mercadorias verificadas no despacho aduaneiro de importação e, portanto, os custos para economia brasileira. Concomitantemente, ela ajuda a reduzir a evasão fiscal na importação e o descumprimento de exigências administrativas, como as exigências de obtenção de anuências dos Ministérios da Saúde, da Agricultura e do Exército.

Uma redução no percentual de verificação sem ganhos de precisão obviamente levaria ao aumento da evasão fiscal e do descumprimento de exigência administrativas, o oposto do objetivo. Assim, deseja-se selecionar melhor as DIs que serão liberadas automaticamente (canal verde) e as que serão distribuídas a um fiscal para verificação aduaneira (canais amarelo, vermelho e cinza).

Desde a Instrução Normativa (IN) SRF nº 680/2006, o fiscal do despacho não é mais obrigado a conferir a DI completa. Ele é obrigado apenas a conferir a parte da DI que motivou seu direcionamento para canal amarelo, vermelho ou cinza e tem a liberdade de verificar ou não qualquer outro aspecto da DI. Assim, também é importante ajudar o fiscal do despacho a escolher as mercadorias que efetivamente verificará.

O Sisam tanto aumenta a precisão da seleção de DIs para canais de conferência quanto ajuda a escolha de mercadorias individuais para verificação. Ele analisa cada item de cada adição de cada DI e, para cada um deles, calcula a probabilidade da presença de vários tipos de erro. Ele também indica possíveis valores corretos para os campos que tiverem erro e calcula a probabilidade e as consequências tributárias e não tributárias de cada um destes valores. Com isto, o Sisam consegue estimar a importância de cada uma das verificações possíveis do ponto de vista da RFB e atuar tanto decidindo automaticamente quais verificações devem ser realizadas quanto apoiar um fiscal que seja responsável por essas decisões.

O histórico de declarações de importação do Sistema Integrado de Comércio Exterior (Siscomex), acumulado desde o ano de 1997, quando ele foi implantado, contém mais uma dezena de milhões de DIs e de uma centena de milhões de mercadorias. Nesta base, as declarações verificadas (cerca de 15% do total) aparecem tanto em sua versão original quanto na versão desembaraçada pelos fiscais da RFB. Assim, é possível identificar o que mudou de uma versão para outra e saber o que estava errado na primeira versão, o que gera um grande potencial para aplicação de aprendizado de máquina (MITCHELL, 1997).

No entanto os dados presentes neste histórico são complexos e produzir um mecanismo preciso para selecionar mercadorias para verificação no despacho aduaneiro, não é uma tarefa simples. O auditor Marcos Ferreira (2003a) analisou o problema e identificou o maior obstáculo ao tratamento estatístico desta base: a presença de atributos nominais de alta cardinalidade.

Um atributo de alta cardinalidade é um atributo que pode assumir muitos valores distintos. O código da Nomenclatura Comum do Mercosul (NCM), por exemplo, pode assumir cerca de 10000 valores. O identificador do importador assume dezenas de milhares de valores na base do Siscomex. Os países envolvidos em transações com o Brasil são cerca de 200. Estes atributos, quando combinados, geram uma explosão exponencial que induz um problema que permeia toda a inteligência artificial, o superajuste. Na presença de superajuste,

a inteligência artificial se sai bem nos casos de treinamento, mas muito mal quando testada em casos novos, como se fosse um pessoa que decorou ao invés de entender um assunto.

Marcos Ferreira fez amplo levantamento bibliográfico e concluiu que a melhor estratégia disponível para lidar com o problema seria o emprego de métodos lineares, posto que eles evitam a explosão combinatória dos atributos. Ele selecionou a estratégia recomendada por Pearl (1988), o Noisy-Or. Esta pesquisa mostrou ganhos expressivos com relação à seleção parametrizada e levou Ferreira a ser agraciado com o segundo lugar no Prêmio de Criatividade e Inovação da RFB (então Prêmio Schöntag) no ano de 2003 (FERREIRA, 2003b).

Contudo métodos lineares, como o Noisy-Or, têm a óbvia deficiência de descartar as interações não lineares entre os atributos. Para um modelo linear, por exemplo, se certo importador tem um risco elevado e certa NCM é frequentemente declarada erroneamente, então a conclusão de que, quando esse mesmo importador declara esta NCM, está-se diante de uma operação cujo risco também é elevado (normalmente ainda mais elevado), é inevitável. Na prática, existe uma interação não linear entre os atributos Importador e NCM que, de tempo em tempos, torna essa conclusão falsa. Ela é causada pelo fato de que, ao ser pego cometendo um erro, o importador tende a parar de cometê-lo, embora possa tranquilamente seguir cometendo erros diferentes. Por mais que os resultados das fiscalizações mostrem que este é o caso, um modelo linear é incapaz de compreender essa exceção. Insistir em erros que são óbvios para os fiscais prejudica a credibilidade do sistema e, portanto, sua adoção, além de impedir melhorias em suas taxas de acerto.

As interações não lineares entre atributos importantes das DIs são várias. Portanto, a partir do trabalho de Marcos Ferreira, dispor de um método capaz de tratar interações não lineares entre atributos nominais de alta cardinalidade sem incorrer em superajuste tornou-se um avanço tecnológico de interesse específico da RFB. Alguns anos depois, este avanço acabou sendo obtido pelo auditor Jorge Jambeiro Filho (2007a). Ele foi apresentado na principal conferência do ramo de inteligência artificial, a *International Joint Conference of Artificial Inteligence*, realizada

na Índia em 2007 (JAMBEIRO FILHO; WAINER 2007) e publicado no principal periódico de inteligência artificial, o *Jornal of Machine Learning*, em 2008 (JAMBEIRO FILHO; WAINER, 2008).

No entanto vários aspectos técnicos importantes para a efetiva implantação de um sistema como o Sisam não estavam resolvidos, entre eles:

- tratar atributos de tipos diferentes tipos ao mesmo tempo e interações não lineares entre todos eles, incluindo: atributos nominais, como o identificador do importador; atributos hierárquicos, como a NCM, que envolve conceitos, como capítulo, posição e item; atributos contínuos, como valores de alíquotas, peso e preço e textos em linguagem natural, como nomes do fornecedor estrangeiro, do fabricante da mercadoria e, principalmente, das descrições livres de mercadorias;
- tratar múltiplas variáveis desconhecidas ao mesmo tempo para poder lidar com fatos, como o de que variáveis como o regime tributário real, o fundamento legal real e a alíquota correta podem todos divergir dos declarados e têm consequências importantes uns sobre os outros;
- ter recursos que permitam cortar o espaço de busca, compensando a explosão exponencial que decorre da presença de múltiplas variáveis desconhecidas e minimizando as perdas;
- realizar aprendizado supervisionado e não supervisionado ao mesmo tempo para aprender com os resultados das fiscalizações, mas também ser capaz de aprender com DIs não verificadas e perceber desvios do padrão que levantem suspeitas mesmo sem que qualquer fraude tenha sido detectada em um contexto parecido;
- ser capaz de aprender com novas DIs sem ter que refazer o treinamento com as DIs antigas e, assim, viabilizar atualizações diárias após já ter sido treinado com alguns milhões de DIs;
- ser capaz de aprender rapidamente e bloquear fraudes que se repitam de forma muito parecida, mesmo a partir de um único exemplo;

- ter um motor de inferência que admite intervenções *ad hoc*, para lidar com peculiaridades do ambiente da RFB e impor restrições estruturais que evitem conferir graus de liberdade excessivos aos modelos estatísticos e, portanto, aumentar sensibilidade a ruído;
- ser rápida o suficiente para levar em conta um histórico de milhões de DIs e produzir respostas em tempo real, atendendo ao fluxo constante do despacho aduaneiro;
- evitar aprender com comportamentos errados, mesmo que repetitivos, não se deixando induzir pelos importadores;
- não se perder ao tratar descrições de mercadorias que incluírem mais palavras associadas ao contexto do negócio que a mercadoria em si;
- ser capaz de lidar com classes mutantes, visto que a legislação pode determinar, por exemplo, que aquilo que era classificado em certa NCM agora se classifique em outra, criando um problema que não existe no aprendizado de máquina tradicional;
- não permitir que importadores identifiquem seu comportamento e aprendam a se situar abaixo da radar do sistema; e
- gerar explicações que permitam aos fiscais adicionar conhecimento que extrapole o escopo da inteligência artificial do Sisam ao processo de decisão e combinar esse conhecimento com as conclusões do sistema.

Cada um desses requisitos impõe restrições à arquitetura do motor de inferência probabilística. Isto faz com que o maior desafio na construção da inteligência artificial do Sisam seja lidar com vários desafios ao mesmo tempo.

É importante notar que muitas ferramentas, como, por exemplo, o Weka (WITTEN; FRANK, 1999), oferecem longas listas de recursos, mas isto não significa ter a capacidade de usar todos os recursos de uma vez para resolver um único problema.

O Sisam atende a todos os seus requisitos ao mesmo tempo, mesmo sendo imperfeito na forma com que lida com a maior parte deles

individualmente. Não havia solução para isto nem no mercado, nem no meio acadêmico.

A importância de tecnologias de mineração de dados vem crescendo no mundo e muito se ouve falar em *big data*, a expressão que denota a exploração dos bancos de dados gigantes que surgiram na internet e dentro das grandes instituições. A RFB já reconheceu a importância da aplicação dessas tecnologias e vem promovendo eventos, como o Seminário de Mineração de Dados e Inteligência Artificial,ocorrido em Bauru, em março de 2015. Além disso, vem planejando adquirir uma robusta plataforma de mineração de dados (correntemente em fase de preparação de edital de licitação). Pelo menos dois trabalhos que têm a mineração de dados em seu núcleo já foram agraciados com o Prêmio de Criatividade e Inovação da RFB.

Ferreira (2003b) investigou a aplicação de várias técnicas de inteligência artificial à seleção de declarações de importação e indicou o uso de redes bayesianas com portas do tipo Noisy-Or. Já Carvalho (2014) investigou os métodos, as técnicas e as ferramentas que propiciam a aplicação de lógica difusa no âmbito de Bancos de Dados da RFB.

Todavia não temos conhecimento de nenhum trabalho que clame ter desenvolvido tecnologia nova na área de mineração de dados e inteligência artificial no interesse específico da RFB. Neste ponto, o Sisam é único.

Além disso, o Sisam é a primeira inteligência artificial de uso generalizado na RFB. Ele está disponível *on-line* para todas as unidades aduaneiras da RFB, está integrado ao Siscomex e trata 100% das declarações de importação registradas no Brasil.

Além disso, a tecnologia desenvolvida neste trabalho não tem aplicação apenas na seleção de mercadorias para conferência aduaneira. A partir dela, vários trabalhos futuros já estão planejados pela Coordenação-Geral de Administração Aduaneira (Coana): fiscalização de mercadorias em exportação, fiscalização de remessas postais e expressas, habilitação para operação no comercio exterior, fiscalização de trânsito aduaneiro e fiscalização de bagagens acompanhadas. Na verdade, durante o processo de desenvolvimento da tecnologia do

Sisam, duas aplicações paralelas já foram desenvolvidas e disponibilizadas através do sistema Contágil (FIGUEIREDO, 2008), ambas na área de tributos internos. Trata-se das funções de emparelhamento inexato de listas de nomes usadas para comparações de folhas de pagamento e do Mecanismo de Detecção de Erros em NCMs e Códigos Fiscais de Operações e Prestações (CFOPs) em notas fiscais. Isto mostra a generalidade e a importância da pesquisa realizada.

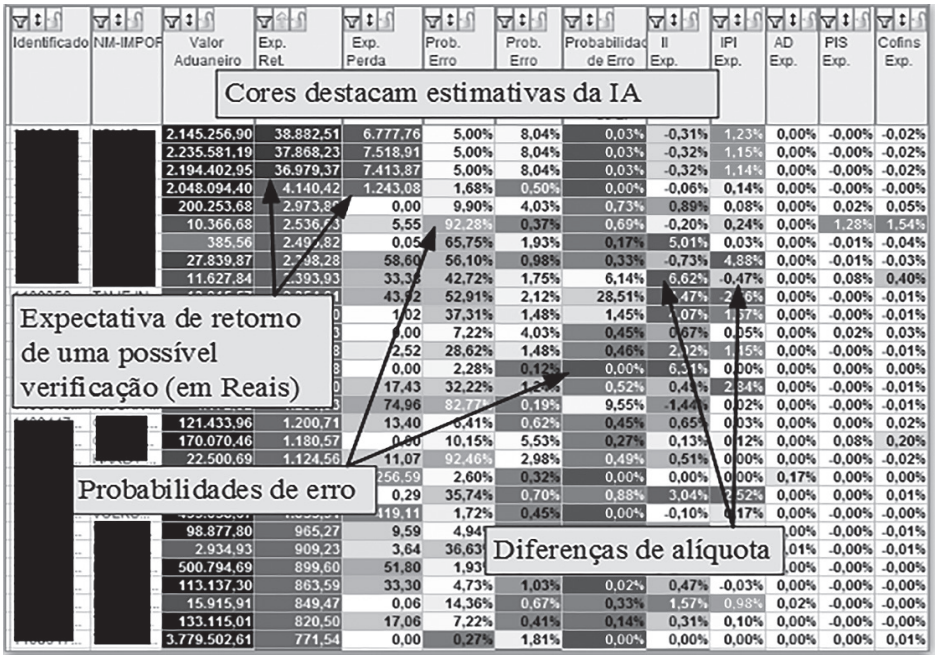
O Sisam coloca a aduana brasileira na ponta do desenvolvimento tecnológico e desperta atenção em outros países. O Canadá, por exemplo, solicitou a visita de uma equipe ao Brasil para conhecer detalhes do sistema.

2 Visão geral do sistema

O melhor ponto de partida para compreender o Sisam é sua interface com os usuários. Esta interface já foi construída dentro do Analisador Inteligente e Integrado de Transações Aduaneiras ─ Aniita (COUTINHO, 2012), que já era usado para redirecionamento de DIs e apoio ao despacho aduaneiro antes do Sisam e que apresenta, em um único local, informações oriundas de diferentes sistemas da RFB.

A espinha dorsal da interface do Sisam é uma planilha interativa com destaques coloridos. Esta planilha pode ser configurada pelo usuário para incluir quaisquer campos da declaração de importação, alertas do Aniita e resultados do Sisam não incluídos por padrão. Na figura 1, mostramos a planilha em sua configuração usual.

Figura 1: Planilha-padrão do Sisam



Para entender a planilha, é preciso conhecer a estrutura básica da DI. Cada DI fornece dados sobre várias mercadorias ao mesmo tempo. Essas mercadorias são agrupadas de acordo com alguns atributos comuns, como a NCM declarada e o fornecedor estrangeiro. Os agrupamentos de mercadorias são chamados de *adições* da DI. As mercadorias possuem alguns dados individualizados dentro de uma adição. O conjunto de dados individuais de uma mercadoria é o que chamamos de *item* da DI.

O usuário do Sisam pode escolher visualizar os dados da planilha segmentados por DI, por adição ou mesmo item. Na figura 1, os dados foram segmentados por adição, o que significa que cada linha da planilha corresponde a uma adição.

As duas primeiras colunas identificam a operação e o importador e foram cobertas por tarjas por questões de sigilo. A terceira coluna mostra o valor aduaneiro das mercadorias. Como a planilha está segmentada

por adição, isto corresponde ao valor total das mercadorias na adição identificada na linha correspondente. Poderia ser o total da DI ou o valor de um item individual se a segmentação escolhida fosse outra.

A quarta coluna mostra o primeiro valor estimado pelo Sisam. Trata-se da expectativa de retorno da verificação. Para cada item na DI, o Sisam calculou a probabilidade de cada erro possível e a probabilidade de cada valor correto caso haja erro (a NCM correta, por exemplo). Ele considerou as consequências tributárias (a alíquota da NCM correta, por exemplo) e não tributárias (uma possível exigência de licença de importação associada a NCM correta, por exemplo) de cada um desses valores possíveis e combinou tudo para chegar a este valor que corresponde em Reais ao interesse da RFB na verificação do item, da adição ou da DI representada na linha da planilha. Nos dois últimos casos, os valores da expectativa de retorno são as somas das expectativas dos itens contidos na adição ou na DI.

As consequências não tributárias foram mapeadas para Reais pela Coana e podem ser alteradas. Por exemplo, um erro de classificação fiscal sem alteração em alíquotas de impostos e sem implicações em termos de exigências administrativas pode estar mapeado para R$1.000,00. Então, se o Sisam considerar que há 10% de chance de que um erro destes esteja presente, a expectativa de retorno será acrescida de R$100,00. Informar exatamente quais são os aspectos não tributários mapeados e que valores receberam está fora do escopo deste trabalho.

Se um fiscal realizar 1.000 verificações, todas com expectativa de retorno de R$500,00 (seja por questões tributárias ou administrativas), espera-se que ele recupere para RFB o valor de R$500.000,00.

É claro que nem sempre o Sisam vai acertar essa estimativa, mas, pelo menos, se compreende exatamente o que ele está tentando prever e o que seus desenvolvedores estão tentar fazer com que ele preveja.

Não se espera que, ao fazer uma verificação com expectativa de retorno de R$500,00, o fiscal realmente vá recuperar esse valor. Em geral, se as suspeitas que pairam sobre as mercadorias não se confirmam, o retorno é zero e, se confirmam, ele é maior do que R$500,00, de modo que a média seja de R$500,00.

A expectativa de retorno orienta o fiscal para que ele aproveite bem seu tempo e sua energia. Uma mercadoria de valor médio pode ter uma expectativa de retorno alta se as probabilidades de erro forem altas. Uma mercadoria de pouco valor pode ter a expectativa de retorno alto se houver probabilidades de erro altas com consequências administrativas importantes, como uma fuga de licença de importação. Já uma mercadoria de muito valor pode merecer a atenção do fiscal ainda que as probabilidades de erro sejam relativamente baixas.

A quinta coluna mostra o que chamamos de expectativa de perda. Ela corresponde à possibilidade de que a verificação acabe apurando diferenças tributárias favoráveis ao contribuinte e que, portanto, causem prejuízo financeiro a RFB. Cabe ao fiscal decidir se deve ou não tomar providências para corrigir esses casos.

Um mesmo item pode ter expectativa de retorno e de perda elevados. Isto ocorre, por exemplo, quando há suspeita de erro de classificação fiscal com duas NCMs alternativas possíveis, uma delas com alíquota maior e outra com alíquota menor que a declarada.

A sexta coluna mostra probabilidade de que a NCM declarada para um item esteja errada. Se a planilha estiver segmentada por adição ou por DI, a probabilidade apresentada será a de que haja pelo menos um item com NCM incorreta.

A sétima coluna mostra a probabilidade de que o país origem declarado esteja incorreto.

A oitava coluna mostra a probabilidade de que a mercadoria precise de uma licença de importação que não foi obtida. Isto pode ou não ser consequência de um erro na NCM declarada.

As últimas colunas mostram expectativas de diferença entre as alíquotas efetivas declaradas e as reais para Imposto de Importação (II), Imposto sobre Produtos Industrializados (IPI), *Antidumping*, Programa de Integração Social (PIS) e Contribuição para o Financiamento da Seguridade Social (Cofins). As alíquotas efetivas são a razão entre o imposto que deve ser recolhido a título do tributo e o valor de sua base de cálculo *ad valorem*. Note que, se houver suspeita de alíquota

específica, os valores a ser recolhidos serão divididos pelo valor da base *ad valorem* para que assim possamos mostrar uma escala única.

As diferenças em alíquotas podem ser consequências de erro na NCM ou no país origem, cujas probabilidades são apresentadas explicitamente na planilha. Ao mesmo tempo, elas podem ser consequências de erros em ex-tarifários, regimes tributários e acordos tarifários que, por padrão, não têm suas probabilidades exibidas na planilha.

Para agilizar a identificação das verificações prioritárias, a planilha pode ser ordenada por qualquer coluna e os valores são acompanhados por cores.

Na figura 1, ordenamos a planilha por expectativa de retorno. Esta coluna segue um degradê que vai do branco, para expectativa de retorno igual a zero, até o preto, usado para valores muito altos, passando pelo rosa, pelo lilás e pelo roxo. Na figura 1, não conseguimos ver a cor branca nesta coluna, justamente porque, ao ordená-la, fizemos com que os menores valores não coubessem na tela.

O valor aduaneiro das mercadorias também segue o mesmo degradê. Podemos notar que os valores altos tendem a se associar a expectativas de retorno altas, mas a relação não é direta.

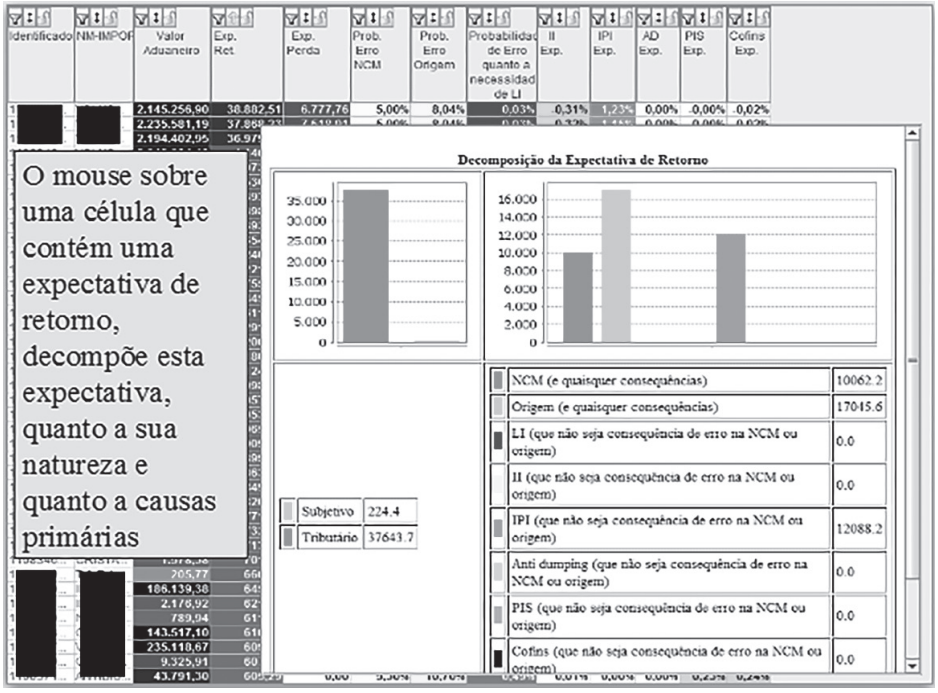
As probabilidades de erro seguem um degradê em que o branco corresponde a uma probabilidade de erro considerada média, que fixamos em 3%. À medida que a probabilidade de erro vai subindo, ela vai ficando amarela, laranja e, por fim, vermelha.

O Sisam não se limita a identificar sinais de presença de fraude. O aprendizado com o histórico de DIs também lhe permite identificar probabilidades de erro particularmente baixas. Normalmente isto decorre de um histórico de verificações que sempre confirmou o que estava declarado. Quando a probabilidade de erro vai ficando menor, ela vai ficando verde e depois azul.

As expectativas de diferenças em alíquotas seguem o mesmo degradê em tons de roxo das expectativas de retorno, para valores positivos, e um degradê em tons de azul, para valores negativos.

Ao repousar o *mouse* sobre qualquer célula da planilha, o fiscal recebe algum tipo de explicação ou informação útil. Na figura 2, repousamos o *mouse* sobre uma expectativa de retorno de R$38.882,51. O Sisam decompôs esta expectativa quanto a sua natureza que pode ser tributária, ou não tributária (aparece na figura 2 como subjetiva) e também quanto à causa primária da expectativa.

Figura 2: Decomposição da expectativa de retorno

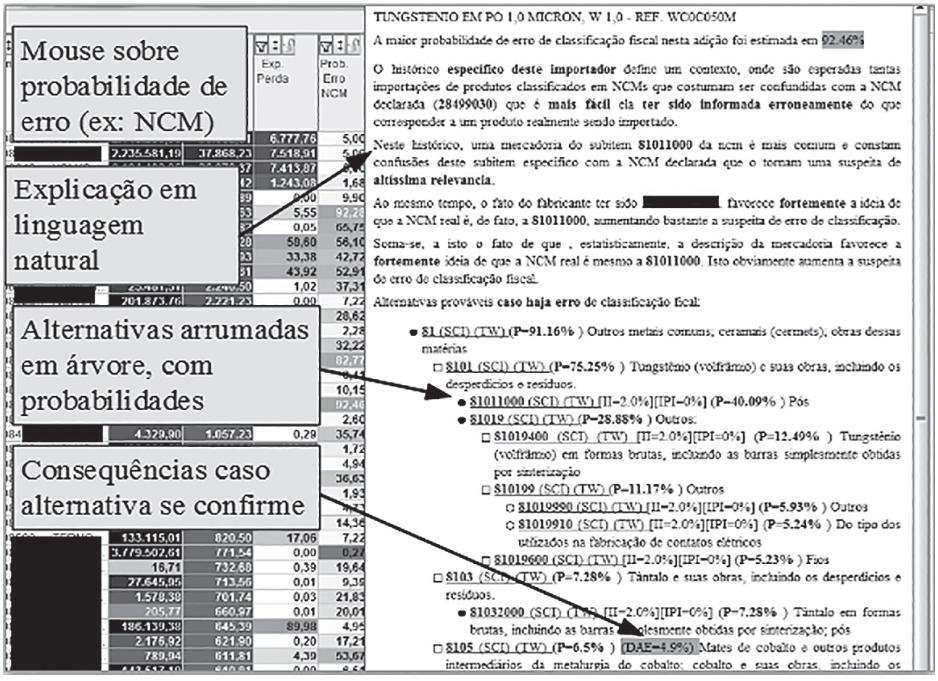


A primeira decomposição faz o fiscal saber se deve esperar encontrar diferenças nos tributos recolhidos ou descumprimento de obrigações administrativas. A segunda diz rapidamente ao fiscal onde deve procurar pelos erros. Se, por exemplo, houver uma expectativa de diferença na alíquota do II devida a uma suspeita de erro de classificação, ela estará computada na legenda correspondente a *"NCM e quaisquer consequências"*. Se a expectativa for decorrente de um erro no regime tributário do II, então ela aparecerá computada na legenda

"II que não seja consequência de erro na NCM ou no país origem". O mesmo pode ocorrer com qualquer outro tributo. O erro de origem também pode ter impacto na expectativa de retorno, normalmente por afetar acordos tarifários.

Na figura 2, repousamos o *mouse* sobre uma probabilidade de erro na NCM declarada estimada em 92,46%.

Figura 3: Justificativa da suspeita na NCM e sugestões



O Sisam exibe um texto contendo a descrição da mercadoria conforme informada pelo importador, uma justificativa em linguagem natural para a suspeita apresentada e uma árvore de sugestões de NCM. A árvore é um sumário da TEC com as posições mais prováveis considerando tudo o que o Sisam analisou. Cada código da TEC vem acompanhado da diferença de alíquota estimada caso aquele código se confirme. As necessidades de licenças de importação, quando existem, também são apontadas na árvore.

Clicando em qualquer código na árvore, o usuário pode ver a TEC completa, as soluções de consulta e, quando o Aniita é configurado corretamente, até abrir ferramentas de apoio como o TECWin. Em todos os casos, ele já é dirigido para as informações correspondentes ao código no qual clicou.

A descrição da mercadoria apresentada na figura 3 é

> "TUNGSTÊNIO EM PÔ 1,0 MICRON, W 1,0 – REF. WC0C050M".

O Sisam informa a probabilidade de erro dizendo

> A probabilidade de erro de classificação fiscal neste item foi estimada em 92.46%
>
> (Texto gerado automaticamente pelo Sisam ao analisar caso real.)

A explicação em linguagem natural é a seguinte:

> O histórico específico deste importador define um contexto, onde são esperadas tantas importações de produtos classificados em NCMs que costumam ser confundidas com a NCM declarada (28499030) que é mais fácil ela ter sido informada erroneamente do que corresponder a um produto realmente sendo importado.
>
> Neste histórico, uma mercadoria do subitem 81011000 da NCM é mais comum e constam confusões deste subitem específico com a NCM declarada que o tornam uma suspeita de altíssima relevância.
>
> Ao mesmo tempo, o fato do fabricante ter sido XXXXXX S.A. favorece fortemente a ideia de que a NCM real é, de fato, a 81011000, aumentando bastante a suspeita de erro de classificação.
>
> Soma-se, a isto o fato de que, estatisticamente, a descrição da mercadoria favorece a fortemente ideia de que a NCM real é mesmo a 81011000. Isto obviamente aumenta a suspeita de erro de classificação fiscal.
>
> (Texto gerado automaticamente pelo Sisam ao analisar caso real.)

No exemplo, a mercadoria havia sido declarada como pertencente à posição 28499030 (Carbonetos de Tungstênio). O Sisam vasculhou sua base de conhecimento e percebeu que outras NCMs costumam ser confundidas com ela. Ele analisou o histórico do importador e considerando seu Código de Atividade Econômica (Cnae) notou que muitas destas NCMs suspeitas são esperadas para ele. Na verdade, é tão esperado que esse importador importe mercadorias de NCMs que costumam ser declaradas erroneamente como se fossem 28499030, que é até mais esperado que este código apareça como consequência de erro que corretamente aplicado. Isto levou o sistema a gerar o primeiro parágrafo da explicação.

Se o Sisam tivesse encontrados confusões relevantes, mas nem tanto, ele não diria que "*é mais fácil a NCM ter sido informada erroneamente que realmente corresponder à importação"*. Ele poderia ter dito que, devido aos erros encontrados, a operação requeria *"alguma atenção"* ou mesmo *"muita atenção"* sem chegar a fazer uma afirmação tão forte quanto fez. O sistema procura regular o tom do texto de acordo com a força das evidências encontradas.

No segundo parágrafo, o Sisam informa que encontrou uma NCM em particular que é mais comum no contexto do importador e que costuma ser confundida com a NCM declarada. Esta NCM é a 81011000 (Pós de Tungstênio) e diz ao fiscal que ela é *"uma suspeita de altíssima relevância"*. Novamente, se as evidências fossem mais fracas, o Sisam teria sido mais comedido.

No terceiro parágrafo, o Sisam constatou que a NCM suspeita está no rol de mercadorias que o fabricante informado costuma vender para o Brasil. Para isto, ele olhou o histórico de outros importadores que já compraram desse mesmo fabricante. Essa informação vem a aumentar as suspeitas. Note que o Sisam introduz o parágrafo com *"Ao mesmo tempo"* uma expressão que passa a ideia de continuar na mesma direção. Se ele tivesse constatado que o fabricante não vende a NCM suspeita para outros importadores, ele comentaria isto e abriria o parágrafo de um modo compatível, usando *"Por outro lado"*, *"Em contraste" ou alguma outra expressão de oposição.*

A explicação termina com o Sisam dizendo que a descrição da mercadoria também aponta para a NCM 81011000, o que parece estar correto lendo a descrição e o texto da NCM.

Pode até ser que a NCM declarada pelo importador, na verdade, esteja correta e a descrição da mercadoria tenha deixado de incluir a informação de que se tratava de um carboneto. Porém, considerando tudo o que o Sisam achou na base de conhecimento, isto é pouco provável. Se a mercadoria for conferida, o resultado da fiscalização positiva ou negativa ensinará ao Sisam, reforçando ou mitigando suspeitas similares. Vejamos um segundo exemplo de explicação produzida pelo Sisam:

> "FC-100/F – LUMINARIA ULTRAVIOLETA 100W,230V,FAN COOLED8" PRI, 8 "SEC.
>
> A probabilidade de erro de classificação fiscal neste item foi estimada em 44.85%
>
> No histórico específico deste importador uma mercadoria do subitem 85437099 da NCM é mais comum e constam confusões deste subitem específico com a NCM declarada que o tornam uma suspeita de altíssima relevância.
>
> Vale a pena apontar o fato de que, no histórico do Sisam, este importador já teve mercadorias do subitem NCM 85437099 da NCM conferidas por fiscais 3 vezes e em todos os casos a NCM foi declarada erradamente como sendo do subitem 90275090.
>
> A influência deste fabricante (XXXXXXX CORPORATION) pesou apenas um pouco sobre a suspeita de erro de classificação, mas confirmou levemente a ideia de que a NCM real seria, de fato, a 85437099.
>
> Além disto, estatisticamente, a descrição da mercadoria favorece a ideia de que a NCM real é mesmo a 85437099, aumentando assim a suspeita de erro de classificação.
>
> (Texto gerado automaticamente pelo Sisam ao analisar caso real.)

O exemplo corresponde a uma luminária ultravioleta erroneamente classificada como um aparelho para análises físicas ou químicas.

O Sisam não encontrou uma gama variada de erros que coloquem a NCM declarada sob suspeita, como no caso anterior, mas encontrou uma NCM em particular que se confunde com a declarada, a NCM 85437099. Mais do que isto, apontou o fato de que esse importador nunca declarou uma mercadoria na 85437099 espontaneamente. Todas as três vezes que essa NCM aparece no histórico do importador se devem a reclassificações feitas por fiscais a partir da declaração da NCM 90275090. Naturalmente, pode-se concluir que se ele importar uma luminária da 85437099 novamente, poderá perfeitamente informar que está importando um instrumento de medida da 90275090. O fabricante da mercadoria e a sua descrição reforçaram as suspeitas, mas não tanto quanto no primeiro exemplo.

Quando o Sisam consegue encontrar casos concretos de erros muito similares ao que pode estar ocorrendo, isto costuma contribuir muito para convicção dos fiscais.

Convém observar que nem o Sisam precisa de muitos casos para fazer esses apontamentos (não são raras justificativas em que casos isolados são citados) nem os fiscais precisam de muitos casos para considerar a informação muito relevante. Isto vai na contramão da ideia popular de que um método estatístico precisaria se basear em muitos casos para funcionar bem. A capacidade de aprender com poucos casos é uma peculiaridade dos modelos altamente não lineares que o Sisam emprega e, portanto, um benefício da tecnologia desenvolvida especificamente para atender à RFB.

A análise mais sofisticada realizada pelo Sisam é, sem dúvida, a do erro de NCM, um erro importante o suficiente para ter uma divisão especializada na RFB e ter sido foco de um trabalho com menção honrosa no Prêmio de Criatividade de Inovação da RFB (SIFUENTES, 2014).

Todavia o Sisam também produz textos de apoio para vários outros erros. Por economia de espaço, mostraremos apenas dois outros exemplos, desta vez relacionados com erros no país origem declarado.

Exemplo 1:

> A probabilidade de erro de origem neste item foi estimada em 20.42%
>
> No contexto histórico deste importador e das rotas que envolvem este país de aquisição e procedência (ESTADOS UNIDOS), existem erros nas declarações dos países de origem que tornam a possibilidade de que um item tenha sido produzido em outro país CHINA, REPUBLICA POPULAR) uma suspeita relevante.
>
> Além disto, o fato da NCM declarada ter sido a 84433111 favorece a ideia de que o país origem real é, de fato, CHINA, REPUBLICA POPULAR e contribui para a suspeita de erro na declaração.
>
> (Texto gerado automaticamente pelo Sisam ao analisar caso real.)

Exemplo 2:

> A probabilidade de erro de origem neste item foi estimada em 18.57%
>
> A influência da NCM declarada (29291021) pesou apenas um pouco sobre a suspeita de erro de origem, favorecendo levemente a ideia de que o país origem real seria JAPAO.
>
> Ao mesmo tempo, o fato do fornecedor ter sido XXXXX AG favorece intensamente a ideia de que o país origem real é, de fato, JAPAO ao invés de INDIA, aumentando bastante a suspeita de erro na declaração de origem.
>
> (Texto gerado automaticamente pelo Sisam ao analisar caso real.)

Nos dois exemplos, notamos como o Sisam analisa a rota da mercadoria comparando o país de origem com os países de aquisição e procedência sob a ótica dos erros já capturados pelos fiscais. Também vemos que a NCM influência a análise, assim como as tendências históricas do fornecedor para o tipo de mercadoria.

O Sisam não chega a ter uma fluência poética e, com atenção, pode-se notar imperfeições em seus comentários, mas não é incomum que usuários que os veem pela primeira vez, realmente, pensem que há uma pessoa escrevendo os textos. Nenhuma pessoa poderia fazer isto para todas as mercadorias importadas no Brasil.

Quando estão trabalhando na seleção para despacho, os fiscais podem redirecionar uma DI diretamente a partir da planilha do Sisam. Quando estão trabalhando no próprio despacho aduaneiro, eles podem navegar nas DIs seguindo a planilha na ordem que acharem mais conveniente agilizando seu trabalho.

3 Estrutura básica do sistema

O Sisam é um sistema escrito na linguagem Java e, hoje, dispõe de uma plataforma composta por sete servidores alugados no Serviço Federal de Processamento de Dados (Serpro). Duas destas máquinas são usadas para desenvolvimento, testes e homologação. As outras cinco são usadas em produção, sendo três delas para executar a inteligência artificial descrita no presente trabalho e duas para balanceamento de carga, banco de dados, integração com o Siscomex e comunicação com os usuários. Todas as máquinas possuem 12 núcleos de processamento físico (24 virtuais). As máquinas que executam a IA em produção e as máquinas de teste possuem 64 GB de RAM. As duas outras máquinas possuem 24 GB de RAM.

Quando uma DI é registrada no Siscomex, ele repassa a DI ao Sisam através de um sistema de messageria. Ao ser recebida, a DI é armazenada em um banco de dados de dados MySQL e um módulo denominado *broker* a envia para alguma das três máquinas de IA. A máquina escolhida faz então a análise solicitada e o resultado é colocado de volta no banco de dados.

Quando um fiscal abre um lote de DIs no sistema Aniita, que é executado na máquina do próprio usuário, ele solicita a análise ao Sisam. Em geral, a análise já está pronta e é enviada imediatamente. Se a DI em questão for antiga ou por outro motivo não tiver sido processada ainda, o Aniita envia a DI ao Sisam que faz análise na hora. Os resultados são exibidos no Anita da forma descrita na seção 2.

Quando uma DI é desembaraçada ou retificada, o Siscomex também passa a informação ao Sisam para que ele possa aprender com o histórico de alterações.

Todo esse fluxo é complexo, segue padrões de segurança definidos pela Cotec e é otimizado para atender ao volume diário de declarações de importação. A descrição do fluxo de dados está fora do escopo deste trabalho, que é focado apenas nos aspectos de inteligência artificial.

O módulo de IA mantém uma base de conhecimento em cada máquina onde é executado. Por questão de velocidade, essa base é mantida em arquivos com formato próprio sem intermédio de qualquer sistema de banco de dados.

A base de conhecimento é inicialmente gerada de modo *offline*, a partir de uma grande quantidade de DIs já desembaraçadas. Hoje, em seu formato compactado, ela ocupa cerca de 80 GB.

Posteriormente, em momentos de baixa demanda (à noite ou nos fins de semana), uma das máquinas de produção para de analisar DIs registradas e passa a aprender com DIs desembaraçadas e retificadas. Ela gera uma base de conhecimento diferencial que é depois passada para as outras máquinas de IA e adicionada à base principal em cada uma delas. O processo não exige que o Sisam saia do ar, pois pode ser realizado em uma máquina de cada vez enquanto o *broker* redireciona as análises para outros servidores.

Apenas recentemente, a integração com o Siscomex entrou em operação. Até então, o Sisam realizava todas as análises quando o Aniita lhe enviava as DIs recuperadas nas máquinas dos fiscais. As atualizações da base de conhecimento ocorriam apenas quando um lote de DIs desembaraçadas chegava ao Sisam via apuração especial realizada pelo Serpro.

Hoje, as análises já são feitas de forma integrada e ocorrem à medida que o Siscomex envia as DIs para o Sisam. O aprendizado diário automático planejado para o Sisam ainda não está ativo e a base de conhecimento é atualizada pelo procedimento acima descrito, apenas a partir da iniciativa de operadores humanos. A completa automatização do processo deverá ocorrer nos próximos meses.

4 Estrutura de inteligência artificial do Sisam

A estrutura de aprendizado central do Sisam é um conjunto de redes bayesianas (PEARL, 1988). O mecanismo de inferência da distribuição de probabilidade de cada variável, dados os seus pais nestas redes bayesianas (BNs), foi alterado, passando de tabelas de probabilidade condicionais tradicionais para variantes das hierarquias de suavização descritas em Jambeiro Filho (2007), com o nome de *Hierarchical Pattern Bayes* (HPB), um avanço de interesse específico da RFB.

As hierarquias do HPB foram modificadas para ganho de velocidade e precisão quando a variável-alvo, e não apenas as variáveis preditoras, tem alta cardinalidade.

Nós especiais foram introduzidos nas redes para tratamento de implicações condicionais exatas entre valores de variáveis (como, por exemplo, a relação entre um regime de isenção e a alíquota de um imposto) e exploração de aspectos estruturais, como o número de subdivisões de uma NCM.

Outra necessidade especial tratada foi a de usar o histórico não verificado de um importador para fiscalizar outros importadores, mas não ele mesmo, evitando que o sistema seja levado a considerar um comportamento errado correto, por sua simples repetição. Isto também exige alterações no modo como uma BN tradicional opera.

Também foram introduzidos recursos que permitem ao sistema explorar relações hierárquicas entre valores de variáveis, como a NCM, que tem o formato de uma grande árvore e os países que foram agrupados em zonas econômicas. Estes recursos permitem que o sistema analise uma variável e suas correlações com outras variáveis, em todos os seus níveis, tirando com isto o máximo de informação do banco de dados. As hierarquias de suavização do HPB, que têm em sua base combinações progressivas de valores de atributos, passaram então a lidar também com os aspectos hierárquicos internos de um único atributo.

Essas hierarquias também foram estendidas para tratar textos livres. O aprendizado de máquina é muito usado para tratamento de textos, mas, em geral, envolve correlacioná-lo a apenas uma variável externa, chamada de "*classe*". Usando hierarquias de suavização,

o Sisam correlaciona o texto a mais de uma variável externa de uma só vez e extrai depois a influência da variável de interesse, que é a NCM. Esta manobra permite ao Sisam realizar o que se chama de *explain away* (PEARL, 1988) nas partes irrelevantes do texto, compreendendo-as como consequência do importador, do fornecedor, do fabricante ou da língua falada no país origem e não como consequência da NCM.

Também estendemos as hierarquias para tratar variáveis contínuas e recentemente sofisticamos essa extensão para ganhar velocidade, adaptação a bordas agudas, multimodalidade e conveniência assimptótica diante de *outliers*.

Todas as extensões ao mecanismo de suavização hierárquica foram feitas de modo a preservar a possibilidade de atualizações graduais do conhecimento, uma propriedade existente em BNs tradicionais que não existe em vários outros mecanismos de aprendizado de máquina.

O Sisam lida com alterações na NCM incorporando, em seus cálculos, as relações do tipo “de/para” definidas pela legislação. Se uma NCM é dividida em duas, os dados anteriores à divisão ainda são capazes de separar as duas novas NCMs das outras 10000 e uma quantidade muito menor de dados novos pode prover a separação entre os dois novos códigos, agilizando o aprendizado e evitando uma fase de erros grosseiros que tenderia a seguir as alterações.

Enquanto o Sisam faz seus cálculos, ele registra seus passos em uma estrutura que pode ser posteriormente varrida pelo mecanismo de geração de explicações em linguagem natural. Esta varredura permite a identificação dos padrões que mais afetaram os cálculos, tornando possível a escolha dos comentários que serão emitidos e a calibragem do tom da linguagem empregada. A geração dos textos baseia-se nessa varredura e em variáveis de contexto que permitem a flexão adequada de termos, quanto a gênero e número, assim como a escolha das expressões de continuidade e adversidade que dão fluência ao texto.

Dada a quantidade de variáveis desconhecidas, não foi possível tratá-las de uma vez em uma só BN. Tivemos que usar várias redes separadas e conectá-las “por fora” de um modo mais restrito que o de uma unificação geral. As BNs que tratam o erro de classificação fiscal e o

erro de origem são aplicadas primeiro e seus resultados são usados para cortar o espaço de busca das redes que tratam os outros erros (regime tributário, acordo tarifário, ex, alíquotas etc.). Além disso, criamos um mecanismo que agrupa os valores estimados como menos prováveis por análises marginais mais rápidas, em um valor denominado “outros”, com isto reduzindo a quantidade de hipóteses avaliadas pelo modelo probabilístico completo. A presença do valor “outros” também exige alterações com relação à forma com que as BNs tradicionais funcionam.

A base de conhecimento do Sisam inclui cerca de cinco bilhões de padrões (combinações de valores de atributos). As informações relativas a esses padrões vão desde a simples frequência até distribuições de probabilidade não paramétricas associadas a valores contínuos (HASTIE; TIBSHIRANI; FRIEDMAN, 2001). Esta base não cabe na memória e é gerenciada por um agressivo sistema de *swap* para disco, desenvolvido especialmente para o Sisam. Ele consegue antecipar os dados que precisarão ser carregados em memória e ativa os discos enquanto os cálculos prosseguem de forma massivamente paralela sem ter que parar para aguardá-los.

Os dados que precisam ser armazenados durante a fase de treinamento são definidos automaticamente quando o Sisam executa uma compilação dos modelos probabilísticos. A existência dessa compilação permite que cada modelo seja especificado separadamente, mas que sua execução física reaproveite tudo o que eles tiverem em comum. A compilação também define as posições em que os modelos serão colocados na memória e, para máxima eficiência, gera grandes blocos de dados definidos de acordo com a ordem em que o motor de inferência varre o espaço de hipóteses. Isto reduz a necessidade de *swap*, aumentando a eficiência de um sistema de cache de três níveis: associativo central, associativo por *thread* (para minimizar ações de sincronização) e posicional (reduzindo cálculos com funções de espalhamento e comparações).

Os dados não são mantidos na memória em objetos Java tradicionais e, sim, em grandes vetores de *bytes*, que são interpretados pelo sistema como modelos estatísticos. Isto gera uma economia de espaço da ordem de 14 vezes e reduz o número de objetivo vivos no *heap* da Java Virtual

Machine em muitas ordens de grandeza, evitando conhecidos problemas de desempenho causados pelo coletor de lixo da linguagem Java quando se possui *heaps* da ordem de vários *gigabytes*.

Com já descrito, o aprendizado do Sisam envolve a identificação das diferenças entre as versões registradas e desembaraçadas das DIs. Esta identificação tem que ser feita antes que tentemos treinar as BNs que descrevemos nesta seção e não é trivial. Os itens das DIs não têm identificação única e, quando são corrigidos, frequentemente são movidos para adições diferentes daquela em que se encontravam na versão original da DI. Isto é muitas vezes necessário, porque aquele item passa a não ter os mesmos atributos que os demais na adição, o que não é admitido pelo Siscomex. Não é raro que sejam até movidos para outras DIs (que felizmente recebem o mesmo número de conhecimento de carga), posto que o Siscomex não admitiu, por muito tempo, a criação de mais adições em uma DI. O Sisam tem que achar o item, onde quer que ele esteja, para compará-lo com a versão inicial e descobrir o que mudou. Esta busca é feita por um mecanismo que foi adotado pelo Contágil para comparação de nomes de pessoas em folhas de pagamento. O mecanismo alinha a DI original com todas as versões finais das DIs com mesmo conhecimento de carga encontrando os pares dos itens.

Além de atuar como apoio às decisões dos fiscais, o Sisam tem a capacidade de selecionar DIs para canais de conferência de forma automática, ocupando parte do espaço da seleção parametrizada do Siscomex. Esse recurso ainda não está em produção, mas está implementado e atua usando a teoria da decisão que recomenda a seleção dos casos com maior expectativa de retorno e a teoria dos jogos que recomenda ser imprevisível para o adversário, fazendo sorteios ponderados pela expectativa de retorno. Isto não permite que ninguém se sinta confortável, mantendo-se abaixo do radar do Sisam.

5 Medidas de desempenho

Nesta seção, exibiremos dois tipos de medida de desempenho. Na primeira, previsões realizadas pelo módulo de inteligência artificial do Sisam, atuando de forma isolada, a respeito de mercadorias já verificadas por fiscais, são comparadas com os resultados realmente obtidos com

estas verificações. No segundo, apresentaremos dados relativos ao uso do Sisam em produção, como apoio a decisões de Auditores-Fiscais da Receita Federal do Brasil (AFRFBs). Também incluíremos medidas que mostram que o Sisam é capaz de atender ao volume diário de DIs registradas no Brasil.

Medidas de acurácia com isolamento da inteligência artificial

Para realizar medidas que representem o desempenho da inteligência artificial do Sisam como uma entidade isolada, selecionamos uma amostra de testes contendo 624.517 itens, cada um correspondente a uma mercadoria. Os itens foram oriundos de 187.417 adições presentes em 36.291 DIs, todas desembaraçadas em canais amarelo, vermelho ou cinza (não teríamos como saber os resultados reais se incluíssemos o canal verde). O Sisam foi treinado com 98.798.545 itens oriundos de 27.909.669 adições em 5.509.000 DIs, sendo 88% delas desembaraçadas em canal verde (menos informativas) e 12% desembaraçadas em canais amarelo vermelho ou cinza (mais informativas).

O módulo de IA do Sisam foi isolado do sistema como um todo e ele foi usado *offline* para analisar todos os itens na amostra de testes. A base de DIs usada nos testes não foi de modo algum incluída na base de treinamento.

Desempenho das previsões de erros de classificação fiscal

Para avaliar o desempenho do sistema, medimos a taxa de recuperação para cada percentual de seleção, S, com S variando de 0% a 100%. Para cada valor de S, consideramos que fossem selecionados os $S \cdot N$ itens com maior probabilidade de erro, onde N é o número total de itens na amostra de testes, e fossem liberados sem verificação todos os demais. A taxa de recuperação é dada por

$$Recuperação = \frac{Soma\ de\ itens\ selecionados\ com\ erro}{Total\ de\ itens\ com\ erro};$$

O número total de erros de classificação em nossa base de testes é de 6.007, o que corresponde a pouco menos de 1% do total de itens.

Na figura 4 e na tabela 1, mostramos as taxas de recuperação para erros de classificação fiscal. Para entender a figura, basta notar que, quando a taxa de seleção é de 0%, obviamente nenhum erro é capturado. Quando a taxa é de 100%, também obviamente, todos os erros o são. O importante então é que a curva suba rapidamente mostrando que, com taxas de seleção pequenas, conseguimos capturar um percentual de erros alto, ou seja, é importante que tenhamos taxas de recuperação altas com taxas de seleção baixas.

Figura 4: Curva de recuperação para erros de classificação fiscal

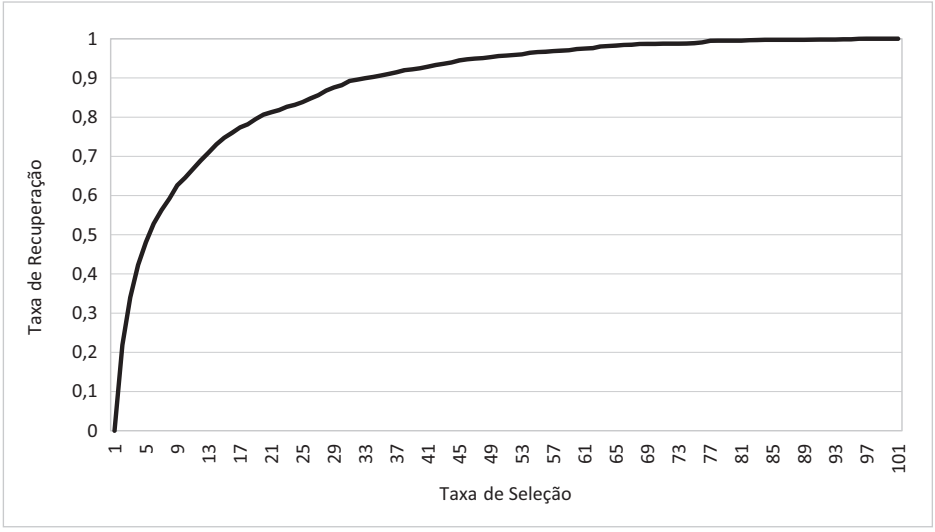


Tabela 1: Taxas de recuperação para erros de classificação

Taxa de Seleção	1%	2%	5%	10%	20%	50%	75%
Taxa de Recuperação	22%	34%	52%	66%	81%	96%	99%

Observando a tabela 1, nota-se que, se ao invés de ter verificado os mais de 600 mil itens em nossa amostra de testes, os fiscais tivessem verificado apenas 2% disto, desde que estes itens fossem os 2% com maior probabilidade de erro de classificação fiscal de acordo com o Sisam, 34% de todos os erros que foram capturados por eles ainda o

seriam. Uma alavancagem de 17 vezes com relação à taxa de seleção. Ao mesmo tempo, se eles tivessem verificado 10% do que verificaram, teriam capturado 66% do total de erros na amostra. Se, por outro lado, eles tivessem verificado 75% de tudo o que verificaram, seria economizado 25% do esforço com análises de classificação fiscal, mas apenas 1% dos erros deixariam de ser capturados por causa disto.

A observação da figura 4 e da tabela 1 nos permite afirmar tranquilamente que o sistema traz benefícios significativos para qualquer taxa de seleção escolhida.

Naturalmente, se os fiscais tivessem conferido apenas 10% dos itens que conferiram, eles poderiam ter conferido vários outros itens que foram liberados no canal verde. Se estes itens também tivessem probabilidades de erro equivalentes, aos 10% que eles continuariam conferindo, a eficiência na captura de erros de classificação fiscal seria multiplicada por 6,6.

A taxa de seleção adequada varia de acordo com a política da RFB e com o contexto. Para fiscais que atuam na seleção de DIs para o despacho e que tratam todas as DIs registradas na unidade diariamente (em alguns casos mais de 10.000 itens por dia), taxas de cerca de 1% são relevantes, visto que eles não costumam ter tempo de redirecionar mais casos do que isto. Nessa taxa, o ganho de eficiência é de 22 vezes, pois aí temos uma concentração de casos altamente promissores.

Para fiscais que atuam no próprio despacho aduaneiro, trabalham apenas com as DIs que foram distribuídas para eles, e que, no passado, eram obrigados a conferir 100% destas DIs, taxas de seleção bem mais altas são razoáveis (note que aqui estamos falando do percentual de itens de uma DI que serão verificados). Hoje, os fiscais do despacho só precisam conferir a parte da DI que motivou sua seleção e, ao mesmo tempo, têm a liberdade de conferir tudo o que acharem pertinente. É bastante razoável, por exemplo, selecionar 20% dos itens em uma DI e conferir a classificação fiscal ou mesmo verificar fisicamente apenas esses casos. Fazendo isto, eles ainda estariam capturando mais de 80% de todos os erros de classificação fiscal e se liberando para realizar outras tarefas. Também é razoável tentar conferir a DI inteira, mas na

ordem favorável indicada pelo Sisam. Se, nesse caso, o fiscal acabar só conseguindo conferir as classificações fiscais de 75% dos itens, apenas 1% dos erros terão escapado.

Na tabela 2, mostramos que o Sisam também tem uma boa capacidade de indicar a classificação fiscal correta. Vemos que 40% das vezes a posição com maior probabilidade de acordo com o Sisam é, de fato, a posição correta e que 65% das vezes a posição correta está entre as cinco primeiras sugestões do Sisam. Este resultado é bastante favorável, considerando que são 10.000 as NCMs possíveis, que vários produtos distintos cabem em cada NCM e que o processo de classificação fiscal segue várias regras não triviais.

Tabela 2: Posição da NCM correta na lista de sugestões

Posição	1	2	3	4	5
Percentual para todos os itens	40%	52%	57%	61%	65%
Percentual para itens entre os 2% com maior probabilidade de erro	56%	67%	70%	75%	77%

Contamos como corretos apenas os casos em que o Sisam acertou a previsão do subitem correto da NCM. Mesmo quando ele não consegue acertar o subitem, o Sisam ainda pode acertar a posição da NCM, ou pelo menos o capítulo, agilizando o trabalho do fiscal.

Na última linha da tabela 2, vemos também que, quando o Sisam estima que a probabilidade de erro é alta, a frequência com que ele consegue acertar a NCM correta também é maior. Isto é esperado, porque os mesmos indícios que fazem o Sisam desconfiar da presença de erro também o fazem identificar a NCM correta. Essa tendência é vantajosa, porque os fiscais tendem a verificar os casos que têm probabilidade de erro alta e, nesses casos, deseja-se que encontrem rapidamente as NCMs adequadas.

Desempenho das previsões de erros de origem

Nossas medidas de desempenho para previsões de erro no país origem declarado são totalmente análogas às medidas relativas a erros de classificação fiscal. O total de erros de origem na base de teste foi

de 1.890. Na figura 5 e na tabela 3, mostramos as taxas de recuperação para erros na declaração do país origem.

Figura 5: Curva de recuperação para erro na declaração do país origem

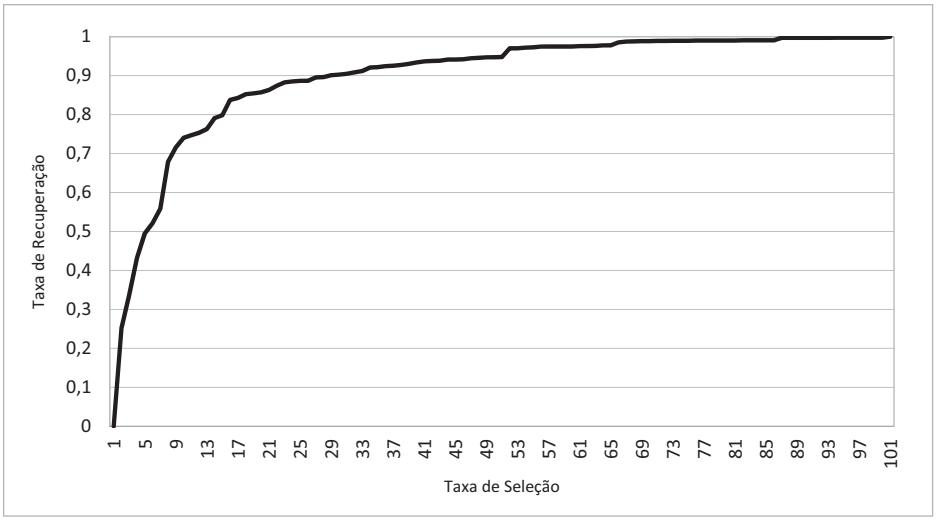


Tabela 3: taxas de recuperação para erros na declaração do país origem

Taxa de Seleção	1%	2%	5%	10%	20%	50%	75%
Taxa de Recuperação	25%	34%	52%	75%	86%	95%	99%

Na tabela 4, mostramos as posições em que o país origem correto costuma aparecer entre as sugestões do Sisam.

Tabela 4: Posição do país origem correto na lista de sugestões

Posição	**1**	**2**	**3**	**4**	**5**
Percentual para todos os itens	53%	64%	72%	82%	83%
Percentual para itens entre os 2% com maior probabilidade de erro	74%	86%	90%	90%	90%

De forma geral, as taxas de acerto para erros de origem são superiores às taxas para erro de classificação fiscal. O Sisam é mais sofisticado no trato do erro de NCM do que no do país origem. Apesar

disso, o Sisam ainda acerta o país com mais facilidade, porque o problema é inerentemente mais fácil. São apenas cerca de 200 países ao invés de 10.000 e a complexidade das trocas que costumam ocorrer é menor.

Desempenho na previsão de erros de licenciamento

Na figura 6 e na tabela 5, mostramos as taxas de recuperação para erros de licenciamento. Isto inclui a falta de Licença de Importação (LI) quando necessária ou a presença de LI quando isto não precisava ocorrer. Em geral, os erros são do primeiro tipo, mas o segundo caso também existe. Em nossa base de testes, havia 986 erros de licenciamento.

Com apenas 1% de seleção, 51% dos erros de LI são capturados. Uma enorme alavancagem. Grande parte destes erros decorre de erros de classificação fiscal em que uma mercadoria tem sua NCM alterada de uma posição que não requer LI para uma que o faz. Porém algumas posições da NCM requerem LI para algumas mercadorias e outras não, permitindo que ocorra falta de LI mesmo sem erro de classificação. Os resultados na figura 6 e na tabela 5 incluem todos os erros de LI.

Embora as taxas sejam de forma geral muito boas, o fato de que observamos uma quase estabilização da recuperação para taxas de seleção entre 10% e 15% com uma subida repentina logo a seguir mostra uma deficiência em nosso mecanismo. Uma de duas coisas tem que estar acontecendo: ou estamos subestimando a probabilidade de erro de LI em um grupo relevante de mercadorias, fazendo com que elas sejam movidas para o fim do *ranking* ou estamos superestimando a probabilidade de erro em algum grupo que está sendo movido em massa para frente no *ranking*, empurrando para trás os demais itens.

Figura 6: Curva de recuperação para erros de licenciamento

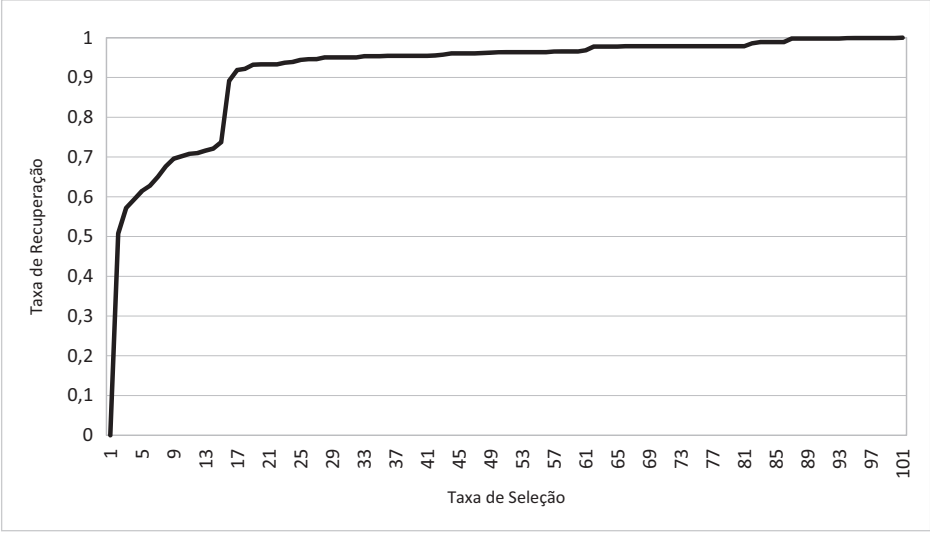


Tabela 5: Taxas de recuperação para erros de licenciamento

Taxa de Seleção	1%	2%	5%	10%	20%	50%	75%
Taxa de Recuperação	51%	57%	63%	71%	93%	96%	98%

Sabemos que alterações na legislação que fazem com que LIs passem a ser exigidas ou deixem de sê-lo são relativamente comuns. A velocidade com que o Sisam se adapta a mudanças na legislação ainda é pequena e isso pode explicar vários erros nas previsões relativas a licenciamento. Pretendemos tornar essa adaptação mais rápida e abordaremos o assunto na seção 7.

Desempenho em outros tipos erro

Na tabela 6 e na figura 7, mostramos taxas de recuperação para erros em regimes tributários, na existência de redução na base de cálculo do PIS/Cofins e nos acordos tarifários informados. Na tabela 6, consta também a quantidade de erros de cada tipo encontrados em nossa base de testes.

O nível de esforço realizado para melhorar o desempenho do Sisam nesses erros foi, até agora, menor que o esforço realizado para tratar o erro de classificação fiscal. Apesar disso, várias taxas de recuperação são muito altas.

A explicação para isto é que erros em regimes tributários quase sempre só ocorrem quando algum regime tributário especial é solicitado (embora o Sisam também aponte casos em que um regime especial deveria ter sido pedido e não o foi). O mesmo vale para acordos e para redução da base de cálculo do PIS/Cofins. Isto torna esses erros mais facilmente previsíveis.

Figura 7: Curvas de recuperação para vários tipos de erro

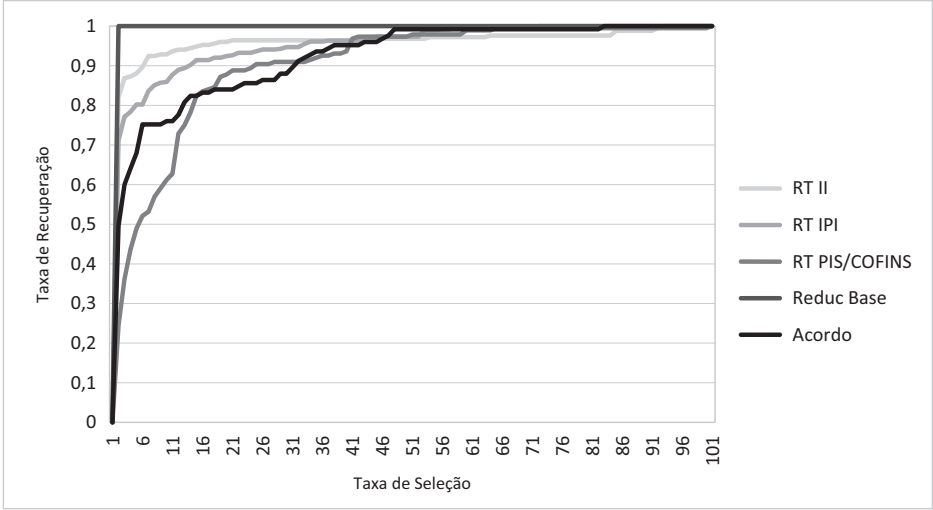


Tabela 6: Taxas de recuperação para vários tipos de erro

	Total	1%	2%	5%	10%	20%	50%	75%
RT II	**252**	82%	87%	90%	94%	96%	97%	98%
RT IPI	**490**	71%	77%	80%	88%	93%	97%	100%
RT PIS/Cofins	**188**	24%	36%	52%	62%	89%	98%	99%
Redução base PIS/Cofins	**13**	100%	100%	100%	100%	100%	100%	100%
Acordo tarifário	**125**	50%	60%	75%	76%	84%	99%	99%

Planejamos investir mais no tratamento desses erros. Quando fizermos isto, também realizaremos medidas de desempenho mais precisas, que diferenciem as taxas de acerto para cada regime tributário solicitado, assim como para os tipos de acordo tarifário. Também selecionaremos uma base de testes que contenha um percentual maior de pedidos especiais. Nos testes que fizemos, só havia, por exemplo, 13 casos de erro na redução da base de cálculo de PIS e Cofins deformando os resultados.

O mais importante nas medidas apresentadas na tabela 6 e na figura 7 é que o Sisam já é capaz de contribuir significativamente para captura de todos estes tipos de erro.

Desempenho na previsão da diferença na alíquota efetiva

Definimos a alíquota efetiva pela equação

$$Alíquota\ efetiva = \frac{Total\ de\ tributos\ recolhidos}{Valor\ aduaneiro\ da\ mercadoria}.$$

Assim, ela já inclui II, IPI, PIS, Cofins e *Antidumping*. Os efeitos de regimes tributários, acordos tarifários, ex-tarifários e da redução na base de cálculo do PIS e da Cofins também já estão embutidos na alíquota efetiva.

Com se vê na equação que a define, a alíquota efetiva é um conceito *ad valorem*, mas os valores pagos como consequência de alíquotas específicas também são incluídos.

Na tabela 7 e na figura 8, mostramos taxas de recuperação para diferenças em alíquotas. O conceito de recuperação nesses casos é ligeiramente diferente para refletir, não apenas a existência ou não de diferenças, mas também seus valores:

$$Recuperação = \frac{Soma\ das\ diferenças\ em\ itens\ selecionados}{Soma\ das\ diferenças\ sem\ todos\ os\ itens}.$$

Para taxas de seleção pequenas, temos uma grande alavancagem. Selecionando apenas 1% das mercadorias para conferência, conseguimos

recuperar 27% do total da soma das diferenças em alíquotas observadas. Conseguimos, portanto, multiplicar a eficiência por 27.

As mercadorias selecionadas, quando a taxa de seleção é pequena, são aquelas com alta probabilidade de erro e consequências tarifárias grandes. O Sisam consegue apontar esses casos com bastante sucesso, frequentemente como consequência do tratamento da NCM, do país origem, do Regime Tributário (RT), do acordo e do ex.

Figura 8: Curva de recuperação para diferenças em alíquotas

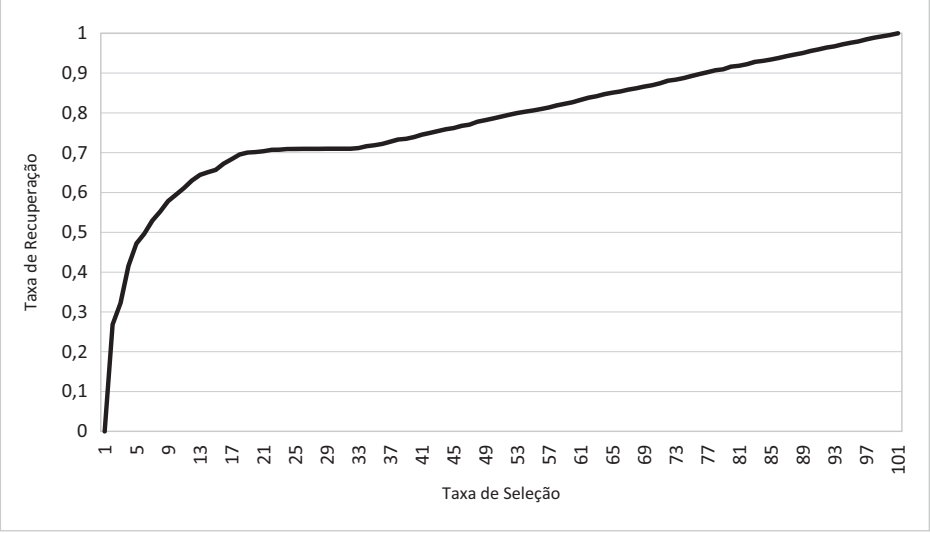


Tabela 7: Taxas de recuperação para diferenças em alíquotas

Taxa de Seleção	1%	2%	5%	10%	20%	50%	75%
Taxa de Recuperação	27%	32%	50%	61%	70%	79%	90%

Para taxas de seleção mais altas, percebemos uma queda no desempenho das previsões. Atribuímos essa queda à lentidão na adaptação a mudanças na legislação que alteram o valor da alíquotas. Como veremos na seção 7, já temos planos para melhorar esse desempenho.

Medidas de adequação ao volume diário de DIs

Na tabela 8, mostramos a quantidade de DIs, adições e itens e os tempos de processamento observados em nossos testes, usando apenas uma máquina com 12 núcleos de processamento e 64 GB de RAM.

Tabela 8: Volume e tempo de processamento

Número total de DIs	Número total de adições	Número total de itens	Tempo total	Tempo por DI	Tempo por adição	Tempo por item
36.291	183.417	624.517	47665 s	1,3 s	0,26 s	0.08 s

Em um dia útil, cerca de 10.000 DIs são registradas e, em raros dias de pico, este número pode chegar perto de 20.000. Isto corresponde a cerca de 100.000 adições e 340.000 itens.

As DIs são enviadas pelo Siscomex ao Sisam que tem que as analisar tão rapidamente quanto elas vão chegando. As DIs concentram-se no horário comercial e, entre 16:00 e 17:00, chegam ao momento de pico. Nesse horário, 13% das DIs do dia são registradas. Assim, em uma hora, 2.600 DIs podem precisar ser analisadas. Como o Sisam analisa as DIs item por item, é importante considerar que isso corresponde, em média, a 44.200 itens.

A quantidade de itens em uma DI varia bastante. Muitas DIs têm apenas um item, outras (raras) mais de 30.000. Se as maiores DIs fossem todas registradas de uma vez, não haveria como dar conta delas. O melhor que podemos fazer é manter folga na capacidade de processamento requerida para tratar DIs de tamanho médio.

Pela tabela 8, o Sisam é capaz de processar 45.000 itens em uma hora usando um única máquina. Como ele dispõe de três máquinas em produção, consegue atingir uma folga bastante satisfatória.

O tempo consumido pela comunicação com o Siscomex e com as máquinas dos usuários também são satisfatórios, mas seus detalhes estão fora do escopo deste trabalho. Na prática, o Sisam em produção consegue analisar cerca de duas DIs por segundo, o que é muito próximo do que ocorre quando a IA é aplicada isoladamente.

Medidas com o sistema sendo operado por AFRFBs

Nesta seção, mostramos medidas associadas ao uso do Sisam por fiscais tanto na tarefa de selecionar DIs para o despacho aduaneiro quanto no próprio despacho aduaneiro.

Seleção para despacho aduaneiro

Para avaliar o desempenho dos trabalhos de redirecionamento de DIs do canal verde para os canais amarelo, vermelho e cinza realizado pelas unidades locais, a Coana fez um levantamento usando uma amostra de 7.201 DIs redirecionadas por fiscais nos meses de abril e maio de 2015.

Todo redirecionamento é acompanhado por uma justificativa textual que fica registrada no Siscomex. Os fiscais haviam sido instruídos a incluir a palavra SISAM quando a motivação para o redirecionamento tivesse partido deste sistema. Outros motivos de redirecionamento também são identificados por expressões textuais. Na tabela 9, temos a lista dos motivos mais comuns. O Sisam é a motivação mais frequente para redirecionamento, seguido por suspeitas associadas a NCM cujas justificativas não incluíram a palavra SISAM. Acreditamos que boa parte desses redirecionamentos tenham tido a influência do Sisam que não foi mencionado apenas por esquecimento, mas não vamos contabilizar isso.

Tabela 9: Motivo de redirecionamento

Sisam	NCM	Peso	Descrição incompleta	Vínculo não declarado	LI posterior ao embarque	Cobertura cambial	Outros
31,62%	16,78%	12,80%	5,93%	5,46%	2,49%	0,97%	23,96%

Fonte: Coordenação-Geral de Administração Aduaneira.

Os percentuais de redirecionamentos com resultado positivo (medidos pela presença de retificação na DI) estão na tabela 10.

Tabela 10: Percentuais de redirecionamento com resultado por canal

Amarelo	Vermelho	Cinza
61%	75%	85%

Fonte: Coordenação-Geral de Administração Aduaneira.

A maior parte dos redirecionamentos com base no Sisam tem o erro de classificação fiscal como principal suspeita, o que nem sempre se pode confirmar no canal amarelo. Nos canais vermelho e cinza, temos mais de 75% de resultados positivos. A seleção parametrizada do Siscomex tem uma média 30%.

Hoje, não existem dados no DW Aduaneiro que nos permitam saber exatamente o que foi retificado nas DIs, mas sabe-se que alguns tipos de retificação são muito mais fáceis de ser obtidas. Algumas delas são espontâneas como a frequente alteração no representante legal da empresa. Alterações em descrições que podem ou não ser relevantes também são muito comuns. Juntamente com outras pequenas mudanças que não representam detecções de infração, essas alterações estabelecem um piso para o percentual de retificações tornando muito mais fácil chegar até 30% do que a 75%. Também se deve considerar que o Siscomex escolhe primeiro podendo selecionar todos os casos óbvios. Os redirecionamentos são feitos a partir dos conjuntos de DIs que restaram. Nelas os erros tendem a ser mais sutis.

Assim, pode-se dizer que os redirecionamentos feitos com ajuda do Sisam realmente são muito mais precisos que a seleção parametrizada do Siscomex.

Por outro lado, não podemos clamar que isso seja um resultado do sistema em si. Trata-se de um resultado conseguido pelos fiscais usando o sistema. O que realmente mostra a utilidade do Sisam é o fato de os fiscais optarem por realizar mais de 30% de todo o seu trabalho de redirecionamento a partir das informações que ele produziu. Isto só ocorre porque o Sisam identifica suspeitas relevantes, faz sugestões de correção plausíveis, apresenta essas sugestões tempestivamente e explica os motivos destas suspeitas e sugestões de forma que sejam compreendidas rapidamente por fiscais que tem que fazer seleções entre milhares de mercadorias todos os dias.

Despacho aduaneiro

Para realizar um levantamento da utilidade do Sisam no despacho aduaneiro, fizemos o sistema Aniita perguntar aos fiscais do despacho se o Sisam o ajudou. A pergunta é feita quando o fiscal encerra o trabalho

com um conjunto de DIs e as remove do Aniita. Para não atrapalhar o fiscal, a pergunta é feita apenas uma vez para cada grupo de DIs que o fiscal tenha resolvido remover.

O fiscal pode responder que o Sisam não o ajudou, que o ajudou em alguns casos, que o ajudou em vários casos e que o ajudou em todos os casos.

Quando o fiscal responde que o Sisam o ajudou, isto fica gravado. Infelizmente não foram registradas quantas vezes a pergunta foi feita e não sabemos quantas vezes os fiscais responderam que não foram ajudados pelo Sisam. Coletamos, no entanto, 3.844 declarações de que o Sisam havia sido útil, distribuídas segundo a tabela 11.

O fato de 56% corresponderem a declarações de que o Sisam foi útil em vários casos ou em todos os casos é muito favorável. Em particular, o fato de que 32% das DIs foram incluídas em grupos em que o fiscal declarou que o Sisam foi útil em todos os casos é realmente surpreendente, principalmente considerando que 94% da DIs não haviam sido selecionadas pelo Sisam, mas sim pela parametrizada do Siscomex (80%) ou por outro critério de redirecionamento.

Tabela 11: Utilidade do Sisam no despacho

Útil em alguns casos	1.680	44%
Útil em vários casos	947	24%
Útil em todos os casos	1.217	32%

Fonte: Banco de dados do sistema Aniita.

Isto confirma nossa afirmação prévia de que o Sisam não precisa ter sido o critério de seleção para ajudar no despacho.

6 Trabalhos derivados

A tecnologia do Sisam já é empregada em duas funções do sistema Contágil, ambas na área de tributos internos.

A primeira é o mecanismo de casamentos inexatos que no Sisam é usado para fazer o alinhamento entre as versões desembaraçada e registrada das DIs (veja seção 4).

No Contágil, o mecanismo é usado para emparelhar listas de nomes de pessoas na fiscalização de folhas de pagamento, onde é comum a falta do CPF.

A segunda aplicação da tecnologia da Sisam é o Mecanismo de Detecção de Erros em NCMs e CFOPs em Notas Fiscais (MDECNF), que está descrito no manual do Contágil. O objetivo aqui é detectar créditos indevidos de PIS e Cofins.

Esses créditos normalmente são devidos quando a empresa adquire matérias-primas, mas não quando compra mercadorias para consumo próprio ou revenda. O Código Fiscal da Operação (CFOP) indica o uso ao qual se prestará a mercadoria adquirida e, portanto, define se a empresa tem ou não o direto a se creditar. A detecção de CFOPs errados é, portanto, fundamental.

A partir de um conjunto de notas fiscais, o MDECNF aprende correlações entre a atividade econômica da empresa, indicada pelo Cnae, o tipo da mercadoria, indicado pela NCM, e o uso que a empresa faz da mercadoria, indicado pelo CFOP.

Como para disfarçar um erro no CFOP é comum que a empresa informe uma NCM errada, logo se torna importante também detectar erros nas NCMs.

O histórico específico de cada contribuinte, as descrições das mercadorias e a natureza dos terceiros com quem a empresa negocia são levados em consideração nessa estimativa.

O MDECNF não conta com a estrutura genérica de compilação de modelos probabilísticos e geração para bases de conhecimento que veio a ser desenvolvida para o Sisam posteriormente, o que torna sua atualização mais trabalhosa. Ele também não conta com servidores centrais e precisa levar a base completa para a máquina do usuário, nos forçando a manter essa base pequena. Ela tem cerca de 160 *megabytes* quando compactada e é, portanto, da ordem de 500 vezes menor que a base do Sisam. Além disso, não há um sistema de atualização automático, o que prejudica seu desempenho.

Esperamos que, no futuro, o MDECNF passe a ter todos os recursos que o Sisam possui e passe a ter um desempenho tão bom quanto ele.

Para fins do presente trabalho, o importante é que o MDECNF prova que o interesse na tecnologia do Sisam não está de modo algum restrito à aduana e pode vir a ser útil em todas as frentes de trabalho da RFB.

7 Trabalhos futuros

Nos próximos anos, a Coana não apenas pretende continuar melhorando a fiscalização na importação, mas também otimizar todos os demais procedimentos de fiscalização na área aduaneira, entre eles a fiscalização de mercadorias em exportação, a fiscalização de remessas postais e expressas, a habilitação para operação no comércio exterior, a fiscalização de trânsito aduaneiro e a fiscalização de bagagens acompanhadas.

O Sisam faz parte dos planos da Coana para todas essas áreas e, de fato, a experiência com a importação mostra que ele poderá realizar contribuições importantes em todas elas.

É importante ressaltar que o Sisam não é o único recurso planejado para nenhuma das melhorias pretendidas. Da mesma forma que, na importação, ele será integrado às ferramentas já existentes e a outras que ainda serão construídas. Em particular, o Sisam deverá ser integrado aos novos módulos do Aniita que serão construídos para várias novas áreas.

O Aniita otimiza a aplicação da inteligência dos fiscais de forma direta. O Sisam aprende com os fiscais e com os importadores e adiciona inteligência de máquina aos processos. Assim, a inteligência artificial soma-se à inteligência humana, podendo ser decisiva ou coadjuvante, mas, ainda assim, positiva.

Melhorias na importação

Na área de importação, pretendemos melhorar a sensibilidade do Sisam ao tempo, fazendo com que ele se adapte mais rapidamente a mudanças na legislação.

O tempo costuma ser tratado com o uso de séries temporais (HOLT, 2004), mas é difícil conciliar essas séries com todos os requisitos que

o Sisam já atende. Nossa abordagem para lidar com o problema será tratar o tempo como mais uma variável contínua nas redes bayesianas do Sisam. O tempo não deixa de ser uma variável contínua, portanto a abordagem faz sentido. Porém simplificações que costumam ser razoáveis para outras variáveis, como a suposição de normalidade ou mesmo a suposição mais fraca de monomodalidade, são completamente irrazoáveis para ele. As outras variáveis oscilam com o tempo e matematicamente podemos transformar isto em uma oscilação do tempo em função de outras variáveis. Essa inversão permite o encaixe do tempo na estrutura que já possuímos e que resolve tantos outros problemas. Portando, se formos capazes de modelar variáveis contínuas de um modo que contemple oscilações, podemos chegar a nosso objetivo. O mecanismo de tratamento de variáveis contínuas que acabamos de desenvolver e que descrevemos na seção 4 será fundamental nesta tarefa.

Além da melhoria no tratamento do tempo, pretendemos tratar melhor variáveis como preço, quantidade e peso, que também são contínuas. Elas têm interesse explanatório na detecção dos erros que o Sisam já trata e são, em si, variáveis-alvo, principalmente no que tange a sub e superfaturamento.

Outra melhoria planejada é a previsão de erros em descrições de mercadorias. Hoje, o Sisam já considera a possibilidade de erro nas descrições das mercadorias ao tratar o erro de classificação fiscal. Entretanto ele não sugere explicitamente que a descrição esteja errada, algo que deverá passar a fazer. O Sisam não se limitará a dizer que há suspeita de algum erro na descrição, mas também informará que palavras ele acredita que não deveriam fazer parte do texto. Isto é possível porque, quando um fiscal solicita uma alteração em uma descrição, a alteração fica registrada. Hoje, o Sisam já alinha os textos inicial e final descobrindo que parte do texto inicial foi substituída por que parte no texto final (isto faz parte do alinhamento de itens). Usando mais intensamente essa informação, o Sisam poderá passar a dizer algo como,

> na descrição da mercadoria, onde está escrito *parafuso de alumínio*, considere a possibilidade que devesse estar escrito *parafuso de aço inoxidável*.

O ponto até o qual esta melhoria será implementada dependerá dos custos em termos de memória e processamento que forem percebidos quando ela começar a ser testada.

Também pretendemos tratar melhor os códigos de produtos que aparecem em meio às descrições usando uma versão modificada de tratamento por n-gramas (GUSFIELD, 1997). Com isso, o Sisam poderá deduzir que, se uma HP810 é uma impressora e uma HP820 também é uma impressora, uma HP830 também deve ser uma impressora, mesmo nunca tendo visto esse código antes.

Outro ponto importante é tornar ativa a seleção automática feita pelo Sisam já que, hoje, ele está operando apenas como apoio à decisão humana. Esta decisão automática é menos precisa que a decisão tomada por um fiscal, mas permite um percentual de seleção maior. Este percentual substituiria parte da parametriza, que é bem menos precisa que o Sisam. Além disso, aspectos da teoria dos jogos que os fiscais dificilmente levam em conta seriam considerados, o que levaria a seleções que, mesmo sem trazer o máximo resultado imediato, induziriam o comportamento espontâneo de modo mais eficiente (veja seção 4).

Também é possível fazer com que o Sisam passe a interagir diretamente com o contribuinte, dando-lhe a oportunidade de retificar certas declarações antes que elas sejam submetidas a um fiscal. Seria um procedimento similar ao que já ocorre na malha fiscal do Imposto de Renda de Pessoa Física (IRPF) e permitiria uma quantidade maior de correções sem custo com mão de obra.

Usar dados oriundos de outras bases da RFB que não o Siscomex também trará ganhos importantes. Índices relativos à saúde e à estrutura econômica das empresas deverão ser bastante úteis, assim como as relações entre indivíduos disponíveis no grafo de relacionamentos do Contágil.

Outro ponto importante é fazer o Sisam ajudar na detecção de erros mais raros e graves como contrafação, interposição fraudulenta e tráfico de drogas.

Como as empresas envolvidas nessas atividades, após ser descobertas, param de atuar, o histórico individual para aprendizado

de máquina é menor quando não nulo. Porém a análise da estrutura da empresa e das relações entre indivíduos envolvidos deverão ser os ganchos que tornarão possíveis as abordagens estatísticas do Sisam também nestas áreas.

Sisam em outras áreas da aduana brasileira

Embora já estejam entre nossos objetivos detalhes das extensões do Sisam, para trânsito aduaneiro, bagagem e habilitação ainda não estão definidos.

A aplicação do Sisam à exportação é natural dadas as similaridades com a importação. O mecanismo de detecção de erros de classificação fiscal, por exemplo, deverá ser totalmente aproveitado. Porém há muita preocupação com erros que o Sisam ainda não trata como sub e superfaturamento, contrafação e drogas. Assim, os novos recursos mencionados como melhorias na fiscalização de importações serão ainda mais importantes aqui.

O tratamento de remessas postais e expressas também requer um bom mecanismo de avaliação de preço, porém com uma dificuldade a mais. Nessas operações, a NCM não precisa ser informada, o que prejudica estatísticas sobre os valores das mercadorias. Elas têm que ser feitas apenas a partir dos textos livres. Os textos serão considerados de forma direta, mas uma abordagem planejada para obter um ponto de partida mais estruturado é usar o banco de dados de declarações de importação normais para deduzir a NCM da remessa postal e fazer a análise de preço a partir dela. Para isso, o módulo de análise de DIs precisará ganhar o recurso de atuar como suporte *on-line* a um módulo de uma área diferente, dando início a uma integração de diferentes processos de aprendizado de máquina que, em um mundo ideal, contaminaria toda a RFB.

8 Conclusão

Apresentamos o Sistema de Seleção Aduaneira por aprendizado de Máquina, uma inteligência artificial que aprende com o histórico de declarações de importação e aumenta a eficiência do despacho aduaneiro.

Mostramos, pela lista de seus requisitos e pela descrição de sua arquitetura, que não se poderia esperar encontrar um sistema equivalente no mercado, nem tampouco construí-lo apenas aplicando teorias acadêmicas já existentes.

Também mostramos, tanto com testes em que a inteligência artificial atua isoladamente, quanto com testes em que ela interage com fiscais, que os ganhos resultantes são palpáveis.

Por fim, uma lista de trabalhos futuros relevantes associados à tecnologia do Sisam e a existência de duas aplicações derivadas desta tecnologia já em produção nos permitem afirmar que o Sisam vai além da aplicação de técnicas de mineração de dados e abre um caminho para RFB em sua pesquisa e desenvolvimento. Isto coloca a instituição em uma posição compatível com o potencial e com a complexidade inerentes a seu gigantesco e riquíssimo ambiente de informações.

Referências

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**. Dissertação (Mestrado)—Instituto Tecnológico de Aeronáutica, São José dos Campos, 2003a.

_____. **Seleção probabilística**: melhorando a eficiência da conferência aduaneira. Prêmio de Criatividade e Inovação Auditor-Fiscal José Antônio Schöntag, 2003b.

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. New York: Cambridge U. Press, 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), p. 5-10, 2004.

JAMBEIRO FILHO, J. **Tratamento Bayesiano de Interações entre Atributos de Alta Cardinalidade**. Tese (Doutorado)—Instituto de Computação da Universidade Estadual de Campinas (IC/Unicamp), Campinas, 2007.

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.

_____. HPB: a model for handling BN nodes with high cardinality parents. **Journal of Machine Learning Research (JMLR)**, 9, p. 2141-2170, 2008.

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