Preço de Referência

Manual Técnico:

Precificação NF-e

Sub-Secretário do Tesouro: Bruno Queiroz Jatene

Sub-Secretário Adjunto do Tesouro: Guilherme Correa Petry

Chefe da Divisão de Estudos Econômicos e de Qualidade do Gasto: Marcos Antônio Bosio

Membros da Seção de Qualidade do Gasto:

Eduardo Akira Yonekura (chefe adjunto da seção)

Evelise Dalmoro

Rafael Rodrigues Viero

Valmira de Barros Jordan Filippon

Israel Campos Fama (chefe da seção)

ABRIL DE 2020

Sumário

1. SEFAZ RS E TESOURO DO ESTADO

O projeto Preços de Referência NF-e é a principal ação do Programa de Qualidade do Gasto do RS, coordenado pelo Tesouro do Estado/SEFAZ RS. A SEFAZ RS- A Secretaria da Fazenda do Estado do Rio Grande do Sul - é responsável pela arrecadação dos tributos estaduais, pela gestão financeira e pelo controle da execução orçamentária da administração estadual. Dentre suas principais competências estão: administração tributária, administração financeira, administração orçamentária, programação financeira e liberação de recursos orçamentários, administração da dívida pública, contabilidade pública e societária, auditoria da administração pública, política de estímulos fiscais, avaliação dos convênios e ajustes realizados pela administração com a união, os estados e os municípios, identificação e análise de fontes de recursos, administração financeira da folha de pagamento de pessoal do estado, definição de limites globais para orçamentação e programação de liberação de recursos orçamentários e financeiros, compatíveis com as estimativas e a arrecadação da receita pública, administração do serviço público de loterias do Estado e tecnologia da informação e certificação digital.

Com sede em Porto Alegre, a SEFAZ RS conta com unidades em todo o Estado e com cerca de 3 mil pessoas atuando em suas repartições. Sua estrutura atual foi estabelecida no decreto nº 47.590, de 23 de novembro de 2010. É formada por três órgãos de execução (ou subsecretarias), cada um com atribuições específicas: a Receita Estadual (administração tributária estadual e pela administração das demais receitas públicas estaduais), o Tesouro do Estado (administração financeira estadual) e Contadoria e Auditoria-Geral do Estado – CAGE (sistema de controle interno do Estado).

O Tesouro do Estado é responsável por gerir as finanças gaúchas e por zelar pela aplicação dos recursos estaduais (o que justifica estar sob seu comando iniciativas voltadas à eficiência do gasto). Diretamente, o Tesouro do Estado atende a administração estadual, órgãos e gestores públicos, servidores, fornecedores. De maneira indireta, atende a toda a população gaúcha, uma vez que atua para que a sociedade possa contar com mais e melhores serviços. Sendo assim, presta serviços a:

- Administração estadual, órgãos e gestores públicos - Programação e execução orçamentária e financeira, planejamento de toda a despesa, na elaboração de projetos voltados à eficiência do gasto, atendimento aos gestores de pessoal e departamentos de recursos humanos (folha de pagamento).
- Servidores estaduais - Gestão e processamento da folha salarial e disponibilização de diversos serviços referentes a pagamentos do quadro gaúcho.
- Fornecedores - Responsável pelo pagamento dos fornecedores de bens e serviços públicos de todo o Estado.
- Sociedade - Além de alguns serviços voltados aos cidadãos, o Tesouro desenvolve projetos e ações de políticas públicas de interesse a toda a sociedade gaúcha, buscando sempre disponibilizar informações voltadas à transparência e ao maior entendimento das finanças do RS.

Dentre aquelas já citadas como de responsabilidade da SEFAZ/RS, são atribuições específicas do Tesouro do Estado, além do planejamento, programação financeira e pagamentos: pagamento de pessoal, gestão da dívida pública, precatórios e requisições de pequeno valor, elaboração de estudos, cenários, relatórios gerenciais**, Programa da Qualidade do Gasto**, proposição de soluções para os temas fiscais, ampliação das fontes de financiamento e participação em discussões nacionais sobre finanças públicas.

2. PROGRAMA DE QUALIDADE DO GASTO

A consolidação de um programa que contemplasse ações voltadas à eficiência do gasto acontece no Rio Grande do Sul em um momento crítico para as finanças públicas em todo o país. Nos últimos anos, as administrações públicas e, em especial os estados brasileiros, vivem a pior crise fiscal e financeira de suas histórias. O cenário de estruturas historicamente deficitárias acrescido de fatores como endividamento público, déficits previdenciários e o crescente aumento das despesas são causas centrais do problema. No caso específico do RS, dos últimos 45 anos, em apenas sete o Estado arrecadou mais do que gastou.

Neste cenário, olhar para a eficiência do gasto e passar a desenvolver ações que amenizem a crise não é uma opção. No RS, o empenho tem sido grande e vem estando à altura da urgência, considerando que o RS encontra-se entre os entes federados com as piores situações do país.

Inicialmente, mesmo antes de um programa formalmente estruturado e a partir de conceito um ainda incipiente de qualidade do gasto público, alguns projetos já começavam a ser desenhados na SEFAZ RS/Tesouro do Estado (ainda que desarticulados entre si), dando provas do amadurecimento não somente da discussão e da relevância do tema como também do potencial de resultados que começavam a ser vislumbrados.

Em 2010, com a criação da Divisão de Estudos Econômicos e Fiscais e de Qualidade do Gasto (DEQG), as iniciativas que já vinham sendo construídas foram acolhidas e passaram ter maior peso institucional. O amadurecimento das ações levou à consolidação e formalização, em março de 2013, do Programa de Qualidade do Gasto. Foi instituído por meio de decreto estadual sob a coordenação do Tesouro do Estado, voltado a "racionalizar o gasto público por meio da promoção e da integração de ações voltadas à gestão eficiente do gasto" (decreto 50.183, de 25 de março de 2013).

O propósito do programa é desenvolver projetos e soluções para a melhor gestão da aplicação dos recursos públicos e consequente melhoria da prestação dos serviços. Desde que foi instituído, já alcançou uma economia de R$ 983 milhões (até 2019) em áreas essenciais para a oferta de serviços públicos aos cidadãos e mapeou expressiva economia potencial. O foco de atuação de suas ações é a eficiência e a economicidade no uso dos recursos.

Em sua concepção, foram definidas como premissas do programa a disseminação das melhores práticas de eficiência do gasto público e o envolvimento e integração dos servidores, órgãos e entidades da administração pública estadual. Entre os objetivos pretendidos em sua implantação estavam aprimorar a sistemática de utilização de preços de referência para as compras de bens e contratação de serviços (ou seja, já em sua concepção o preço de referência vem sendo uma diretriz

importante), implementar metodologia de reestruturação e aperfeiçoamento dos processos de trabalho nos órgãos e entidades do RS, sistematizar modelo de gestão do gasto público com o fim de monitorar as despesas e, por fim, capacitar os gestores para ações voltadas à gestão eficiente do gasto público.

Diante dos esforços de estruturação e da curva de maturidade de suas iniciativas, o RS conta hoje com um programa sólido, com foco e eixo de atuação bem definidos, inserido com destaque no planejamento estratégico da instituição e do próprio governo do Estado. Cinco anos após a implantação oficial do programa, o RS experencia resultados concretos e expressivos, e os impactos de seus projetos transformaram o RS em uma das principais referências nacionais no assunto. Não apenas pelo pioneirismo de algumas ações, mas principalmente pelos resultados atingidos. As oportunidades de economia são significativas, e lançam mão de conceitos de inteligência de negócios, incorporação de metodologias próprias, tecnologias avançadas e novos conceitos de gestão de projetos.

No centro dessas ações e referência na aplicação de algumas dessas tecnologias está o projeto Preço de Referência Nota Fiscal Eletrônica, a principal frente do Programa de Qualidade do Gasto do RS e um dos principais projetos do Tesouro do Estado. Trata-se de uma inovação que pode impactar as compras públicas do Brasil inteiro e que permite um alto potencial de economia.

Atualmente, o Programa de Qualidade do Gasto do RS apresenta quatro pilares sobre os quais são desenvolvidas suas ações:

1. Preços de Referência
2. Gestão da Despesa (Gestão Matricial da Despesa)
3. Capacitação e disseminação
4. Redesenho de Processos

Nas ações do pilar Gestão da Despesa, vale destacar os expressivos resultados com a redução do desperdício e com a fortalecimento da cultura de monitoramento da despesa. Em 2018, o projeto acumulou uma economia de cerca de R$ 2 milhões, com destaque para os R$ 500 mil em economia de energia elétrica registrada em 500 escolas gaúchas e para R$ 400 mil economizados com a revisão dos contratos em presídios.

3. NOTA FISCAL ELETRÔNICA

O Sistema de Nota Fiscal Eletrônica (NF-e) é a fonte para a busca de preços que irá compor o cálculo do Preço de Referência NF-e.

Adotada no Brasil compulsoriamente a partir de 2008 (mas desde 2006 com adesões voluntárias, incluindo a do RS), a Nota Fiscal Eletrônica foi um projeto desenvolvido de forma integrada, pelas secretarias da Fazenda dos estados e pela Receita Federal. Consiste em um documento de existência exclusivamente digital, emitido e armazenado eletronicamente, com o intuito de documentar, para fins fiscais, uma operação de circulação de mercadorias ou uma prestação de serviços, ocorrida entre as partes. Sua validade jurídica é garantida pela assinatura digital do remetente, e a autorização de uso fornecida pelo fisco, antes da ocorrência do fato gerador. Sua emissão é obrigatória a todas as empresas que pagam o Imposto sobre Circulação de Mercadorias e Serviços – ICMS - e o Imposto sobre Produtos Industrializados – IPI (Portal NF-e/Ministério da Fazenda).

Resumidamente, o processo consiste em a empresa emissora de NF-e gerar um arquivo eletrônico com informações fiscais da operação, assinado digitalmente (garantia de integridade e de autoria). Este arquivo (a NF-e propriamente dita) é transmitido para a secretaria de Fazenda do estado de jurisdição do contribuinte, onde ocorre uma pré-validação do documento. Após o recebimento, a respectiva SEFAZ disponibiliza consulta pela Internet ao destinatário e a outros legítimos interessados, que precisam deter a chave de acesso do documento eletrônico. O arquivo é também transmitido para a Receita Federal, o repositório nacional das NF-e emitidas em todo o território nacional (e, no caso de operação interestadual, para a SEFAZ do Estado de destino da operação). Para o trânsito da mercadoria, a DANFE (Documento Auxiliar da Nota Fiscal Eletrônica) deve ser impressa para acompanhar a mercadoria e auxiliar na consulta da NF-e.

A NF-e foi um dos três subprojetos que integraram a implantação do SPED (Sistema Público de Escrituração Digital). Os outros dois foram a Escrituração Contábil Digital e a Escrituração Fiscal Digital. O SPED é o instrumento que unifica as atividades de recepção, validação, armazenamento e autenticação de livros e documentos que integram a escrituração comercial e fiscal dos empresários e das sociedades empresárias, mediante fluxo único, computadorizado, de informações. É a modernização das exigências atuais da legislação fiscal e comercial, sendo os livros contábeis e fiscais escriturados na forma digital e sem a necessidade de serem impressos. Graças ao certificado digital (assinatura de documentos eletrônicos), esses documentos são validados em meio eletrônico e assim fiscalizados de forma mais eficiente pela Receita Federal (Portal NF-e/Ministério da Fazenda).

A exemplo do que já vinha acontecendo à época na Espanha (apontada como a pioneira no mundo), Chile (pioneiro na América Latina) e México, a implantação da NF-e no Brasil visava instaurar um modelo nacional de documento fiscal eletrônico que basicamente substituísse a emissão em papel, simplificando as obrigações acessórias dos contribuintes e permitindo, ao mesmo tempo, o acompanhamento em tempo real das operações comerciais pelo fisco.

A implantação constituiu grande avanço na gestão fiscal brasileira, possibilitando muitos benefícios a contribuintes e a administrações tributárias. Além das facilidades aos emissores, possibilitou simplificar e ampliar atividades de fiscalização, maior controle da arrecadação e combate à sonegação.

Conforme especialistas, estudiosos e acadêmicos, a chegada da NF-e, especificamente, e do SPED, de forma geral, promoveram uma verdadeira quebra de paradigmas tanto pelas inovações tecnológicas envolvidas quanto por um novo comportamento adotado, com novas políticas e processos adotados nas administrações públicas e das empresas.

Além disso, o projeto da NF-e no Brasil foi considerado modelo e inovador. Não apenas pela modernização dos processos e da tecnologia aplicada como também pelo envolvimento dos contribuintes já em seu desenvolvimento, fazendo com que o resultado final alcançado tenha atendido também suas necessidades. O resultado foi a geração de benefícios não apenas às administrações tributárias e sim a todas as partes envolvidas no processo de arrecadação – incluindo a sociedade em geral, que passou a contar com um sistema mais transparente e eficaz de arrecadação e fiscalização.

Importante ressaltar o pioneirismo do RS no processo de implantação da Nota Fiscal Eletrônica, Estado que desde a fase inicial do projeto esteve à frente nas discussões, na elaboração de legislação e no desenvolvimento das tecnologias e dos sistemas necessários. Foi o primeiro estado a colocar no ar a SEFAZ Virtual (plataforma para processar e autorizar a NF-e) e seguiu na vanguarda na implantação dos demais documentos fiscais eletrônicos do projeto. Esse posicionamento tornou o RS referência na implantação da nota fiscal eletrônica, o que lhe rendeu prêmios e a visita de diversas comitivas de estados e países, que buscavam conhecer a experiência realizada pelos gaúchos.

4. PREÇOS DE REFERÊNCIA

4.1. INFORMAÇÕES GERAIS

O projeto **Preço de Referência NF-e** pode ser definido como uma frente que desenvolve técnicas para calcular preços de mercado para bens adquiridos pelo setor público.

Resumidamente, a precificação por meio da NF-e consiste na identificação de transações (vendas) de um determinado produto na base de dados gerada pelo Sistema da Nota Fiscal Eletrônica. A partir das transações efetuadas com destino a estabelecimentos (CNPJs) localizados no Estado, calcula-se o preço de referência de mercado. É essencial ressaltar que esta base não contém, via de regra, transações feitas a pessoas físicas.

4.2. FRENTES DO PROJETO

Desde sua concepção inicial, em 2012, a Precificação NF-e foi aplicada em três frentes de atuação na administração estadual do RS:

1. Secretaria Estadual da Saúde (SES): subsidiar compras de medicamentos e insumos
2. Instituto de Assistência à Saúde dos Servidores do RS (IPE-Saúde): remunerar credenciados.
3. Central de Licitações do RS (CELIC): subsidiar compras de bens pelo estado

4.3. DISTINÇÃO METODOLÓGICA

A elaboração dos preços de referência com base na nota fiscal eletrônica segue um processo básico de trabalho, que divide-se nas seguintes etapas (detalhadas mais adiante):

Entretanto, apesar das etapas serem as mesmas, o processo de formação de preços de referência voltado a medicamentos e o voltado aos demais itens que não são desta categoria apresentam metodologias totalmente distintas. As transações com medicamentos são identificadas por meio do seu GTIN, enquanto as transações de outros produtos são identificadas por meio de análise do texto de descrição da NFe. O motivo é explicado na sequência.

O código GTIN nada mais é do que o código de barras e segue um padrão internacional de mercado para identificação de produtos. Atualmente, o campo GTIN na NF-e não possui qualquer validação prévia. Ou seja, o contribuinte é livre até para deixa-lo em branco, por exemplo. Nos testes realizados ao longo do projeto Precificação NF-e, foi possível encontrar casos nos quais um mesmo GTIN foi utilizado para identificar mais de 300 produtos completamente diferentes, quando na verdade, cada código deveria corresponder a um único produto.

Sendo assim, no caso específico de **medicamentos,** sua regulação por parte da ANVISA, bem como as informações constantes no GUIA BRASÍNDICE (publicação que atualiza, quinzenalmente, a relação de medicamentos comercializados no Brasil e que apresenta dados como código numérico do medicamento, código alfabético da apresentação e código numérico do laboratório fabricante), considerados os fatores externos, tornam **o código GTIN um parâmetro suficientemente confiável para a busca de transações**. Por esta razão, a metodologia para medicamentos utiliza o GTIN na identificação das transações.

No entanto, em produtos que não são medicamentos, o GTIN mostrou-se insuficiente para a correta identificação de transações no banco de dados NF-e, e por isso, o processo segue outro caminho metodológico. Nesse caso, é necessário que a pesquisa aconteça a partir da própria descrição do produto. O que exige que a busca de transações na base seja feita por mineração de texto.

Outro ponto sensível para que ocorra separação metodológica nas duas categorias é o fracionamento (ex.: se a NF-e trata da venda uma garrafa ou de um pacote de seis garrafas). O fracionamento é importante pois é a partir dessa etapa que será possível identificar o valor unitário do item descrito na NF-e, assim, aplicar com precisão os métodos de precificação.

A tabela do GUIA BRASÍNDICE, por ser uma referência externa à NF-e, pôde ser utilizada como base de comparação para o algoritmo de fracionamento desenvolvido para os medicamentos. No caso de produtos em geral, não existe tal referência. De forma que o algoritmo desenvolvido para fracionar os medicamentos não funciona para outros produtos. Por este motivo, a metodologia se diferencia também na etapa do fracionamento.

4.4. PROCESSO APLICÁVEL A MEDICAMENTOS

Conforme mencionado anteriormente, o processo de precificação de medicamentos apresenta um comportamento particular e homogêneo graças a fatores que dizem respeito ao preenchimento da nota fiscal eletrônica e a fatores externos, como a utilização dos dados do Guia Brasíndice.

Com a possibilidade de cruzamento dos dados da NF-e com os dados do Guia Brasíndice, a utilização do código GTIN torna-se um parâmetro confiável para a busca de transações.

Dessa forma, o processo básico para execução da metodologia para medicamentos é:

ANÁLISE NA BASE MEDICAMENTOS – GTIN X BRASÍNDICE

Nesta primeira etapa, são geradas duas bases de dados: uma partir da busca/filtro ao GTIN do medicamento pesquisado e outra a partir do Guia Brasíndice. É realizado então o cruzamento dessas informações e gerada uma nova base, preliminar, de onde partirão as etapas seguintes.

Este confronto dos dois dados acontece já a partir do fracionamento, como veremos a seguir.

FRACIONAMENTO MEDICAMENTOS

A quantidade de fracionamento de um medicamento é o número de comprimidos, cápsulas, ampolas, etc contidos na embalagem. Para precificar corretamente os medicamentos, é necessário identificar se o medicamento está com preço, descrito na compra, por unidade de apresentação (“unidade maior”) ou por unidade de fracionamento (“unidade menor”). Para tanto, o algoritmo faz comparações entre os valores unitários das notas de um mesmo GTIN, bem como entre os valores unitários das notas com os valores de referência obtidos do Brasíndice (Preços Fábrica - PF e Preço Máximo de venda ao consumidor - PMC).

Após a identificação da unidade de venda de cada nota, os preços são fracionados, divindindo-os pela quantidade de fracionamento de cada GTIN, que é também uma informação extraída do Brasíndice. Depois disso, os valores unitários fracionados são utilizados na etapa de precificação.

PRECIFICAÇÃO MEDICAMENTOS

Uma vez identificadas as transações de interesse, a remoção de *outliers* (ou valor atípico, que apresenta um grande afastamento das demais da série, ou que está "fora" dela, ou que é inconsistente) foi feita pela aplicação da regra de Tukey, que é baseada nos quartis. A distância interquartílica, *IQR*, como $Q_3 - Q_1$ (ou seja, o preço de transação que representa o terceiro quartil, menos aquele que representa o primeiro quartil).

Segundo a regra de Tukey, um *outlier* seria qualquer valor a uma distância maior que 1,5 vezes a distância interquartílica. Em termos algébricos, fica bem simples: são outliers valores abaixo de $Q_1 - 1{,}5IQR$ ou acima de $Q_3 + 1{,}5IQR$. Apesar de ser um tratamento simples, a mera aplicação da regra de Tukey tem mostrado bons resultados nos casos analisados.

Removidos os outliers, são avaliados os dados restantes a fim de inferir se é possível garantir um intervalo de confiança de 90%. Por meio de um método bootstrap (método de reamostragem usado para aproximar distribuição na amostra de um levantamento estatístico), é avaliado se os dados resultantes de todo o pós-processamento atingem um tamanho mínimo de amostra, que seja capaz de garantir o intervalo de confiança referido.

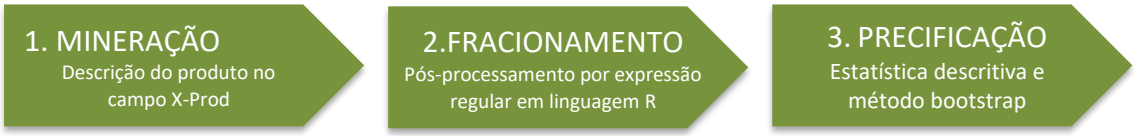
Calculamos, então, a mediana: que é o nosso PREÇO DE REFERÊNCIA.

4.5. PROCESSO ALICÁVEL A NÃO-MEDICAMENTOS

Conforme o que já apresentado anteriormente, a busca no banco de dados da Nota Fiscal Eletrônica com base nos parâmetros do código de barras - GTIN se mostrou sensível quando a pesquisa não se trata de medicamentos, por não ser uma informação precisa ou obrigatória no campo da NF-e.

Diante desse contexto, foi estabelecida como premissa do projeto Precificação NF-e que a ferramenta de pesquisa fosse capaz de receber como entrada a descrição textual do produto a ser licitado. O que exige que a busca de transações na base NF-e seja feita por mineração de texto.

Sendo assim, o processo básico para execução da metodologia para produtos não medicamentos é:



ANÁLISE NA BASE NÃO MEDICAMENTOS – MINERAÇÃO

Na primeira etapa, ao receber a demanda, as únicas informações que a equipe técnica dispõe para a elaboração do preço de referência são a descrição do produto e a quantidade desejada. O processo de mineração de texto que será apresentado neste manual consiste em comparar a descrição do produto, conforme escrito na NF-e, inserido no campo X-PROD, e decidir se condiz com as características do produto que se deseja precificar.

Na NF-e, a descrição do produto é um campo de texto de 120 caracteres, de preenchimento livre pelo contribuinte. Ou seja, não há qualquer regra ou padrão para seu uso.

Como exemplo, será usado como referência a precificação do produto ***"leite de vaca integral em pó***". O primeiro passo, claro, seria identificar transações com tal produto na base NF-e. É natural que se pense na seguinte sequência:

1. Remover os verbetes "***de***" e "***em".***
2. Realizar a busca das transações cujas descrições de produto contenham, obrigatoriamente, as seguintes palavras: ***"leite", "vaca", "integral"*** e ***"pó"***.

Apesar de não haver qualquer erro lógico nesta abordagem, a busca exemplificada resulta em uma quantidade ínfima de transações. No entanto, se retirarmos da busca apenas a expressão ***"vaca"***, a quantidade de transações retornadas é enorme, condizente com o esperado.

Este exemplo foi apresentado para ilustrar o fato de ser muito comum o licitante descrever o produto de forma distinta daquela feita pelos contribuintes ao preencherem suas notas fiscais. Neste caso, ***"leite de vaca integral em pó***" é uma nomenclatura presente no catálogo de produtos do órgão que deseja licitar este produto. Porém, ao longo do trabalho, verificou-se que praticamente nenhum contribuinte preenche a expressão ***"vaca"*** na descrição do produto ao preencher a NF-e. Ficou evidente que órgãos públicos (licitantes) e contribuintes (emissores da NF-e) podem utilizar diversas formas para descrever um mesmo produto (não seguindo um padrão e apresentando uma infinidade de variações no preenchimento).

Para que a mineração de texto seja eficaz, deve ser capaz de perceber tais diferenças e selecionar adequadamente as palavras que devem compor as buscas (*queries*) que serão feitas no banco de dados NF-e.

Uma vez que os órgãos públicos trabalham com catálogos que normalmente apresentam milhares de produtos, a solução de mineração de texto deve ser capaz de, automaticamente, detectar quais palavras devem compor a pesquisa. Nas seções seguintes será apresentado como isso foi feito pela SEFAZ-RS no projeto Precificação NF-e.

CADEIAS DE MARKOV

Nesta abordagem, modelamos a descrição do produto como uma Cadeia de Markov. Além disso, é levado em conta a hipótese de que as palavras se agrupam de acordo com um processo de Markov de primeira ordem.

Pode-se descrever uma cadeia de Markov como um conjunto de *estados*, $S = \{s_1, s_2, \dots, s_r\}$. O processo pode ter início em qualquer estado e mover-se para qualquer outro. A este movimento é dado nome de "passo". Ainda, se o processo se encontra no estado s_i, a probabilidade de que, no próximo passo, vá para o estado s_j é p_{ij}. Uma vez que, por hipótese, o processo é de primeira ordem, esta probabilidade depende apenas do estado atual; **não importando os estados anteriores**. Na sequência, ficará claro o impacto desta simplificação na implementação do modelo (será apresentado também como obter a matriz de transição de estados.

TRANSFORMANDO TEXTOS EM CADEIAS DE MARKOV

Para que se possa representar descrições de produtos por cadeias de Markov, optou-se por um modelo no qual cada palavra corresponda a um estado possível no conjunto de estados "*S"*. A fim de facilitar a compreensão, o exemplo apresentado anteriormente seguirá sendo utilizado: ***"leite de vaca integral em pó***".

É usual que se faça um pré-processamento para remover algumas expressões e a utilização de acentos gráficos. As preposições "***de***" e "***em***" são algumas delas. A descrição pré-processada fica: "***leite vaca integral po"***. A partir de agora, cada uma dessas palavras é um estado possível de ser atingido no processo de formação da descrição do produto.

É necessário ressaltar um ponto importante que chamou a atenção no projeto: uma hipótese adicional, e essencial, ao trabalho. Foi verificado que, apesar das variadas maneiras de se descrever os produtos, tanto os órgãos públicos quanto os contribuintes costumam utilizar a mesma palavra inicial. Ou seja, considerando o exemplo utilizado, ambos iniciariam sua descrição com a palavra "***leite***". Só depois adicionariam demais detalhes: animal, tipo, apresentação etc. Assim, foi decidido travar o estado inicial da cadeia de Markov sempre na palavra inicial usada pelo órgão público e, a partir daí, inferir como os contribuintes detalham aquele produto.

A Figura 1 mostra uma representação gráfica da cadeia de Markov:

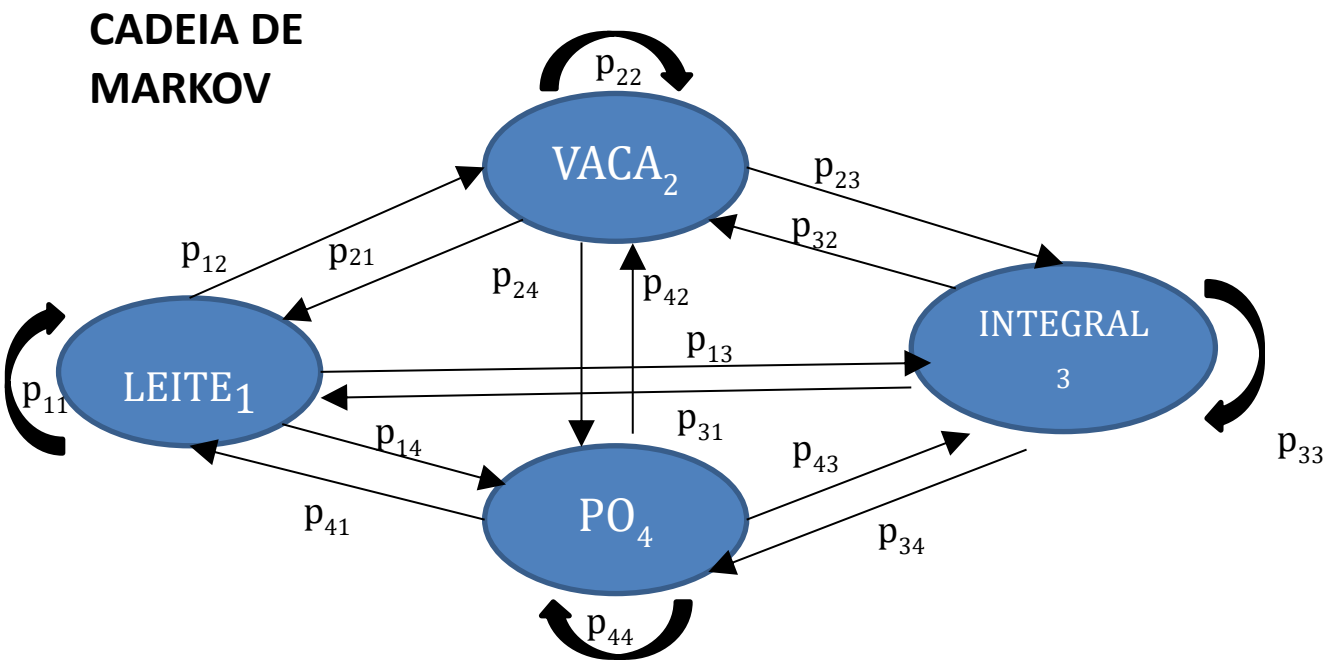


Figura 1: Representação gráfica da cadeia de Markov.

Figura 1:

As setas indicam as transições de estados, acompanhadas das respectivas probabilidades. Por exemplo, p_{12} é a probabilidade de ocorrer a palavra "***vaca***", dado que ocorreu a palavra "***leite***". Já p_{22} é a probabilidade de ocorrer a palavra "***vaca***", dado que a última palavra foi "***vaca***".

A matriz de transição de estados agrupa todas essas probabilidades.

$$M=\begin{bmatrix} p_{11} & \cdots & p_{14} \\ \vdots & \ddots & \vdots \\ p_{41} & \cdots & p_{44} \end{bmatrix}$$

Cada palavra que compõe a descrição do produto (como escrita pelo órgão licitante) é um estado possível na cadeia de Markov. Assim, as palavras ***leite, vaca, integral*** e ***po*** serão denominadas por estados s_1, s_2, s_3 e s_4.

Falta apenas definir o estado da cadeia de Markov, o que é feito por meio de um vetor-probabilidade que, na verdade, representa a probabilidade de cada um dos estados da cadeia. Este vetor pode ser escrito da seguinte forma:

$$u = (p_1, p_2, p_3, p_4)$$

Onde, p_1 representa a probabilidade de s_1, p_2 a probabilidade de s_2 e assim por diante.

O próximo estado da cadeia (leia-se o mais provável de ocorrer) é dado por, $u \cdot M$:

$$u = (p_1, p_2, p_3, p_4) \cdot \begin{bmatrix} p_{11} & \cdots & p_{14} \\ \vdots & \ddots & \vdots \\ p_{41} & \cdots & p_{44} \end{bmatrix}$$

Destaca-se que a presente explanação objetiva apenas ilustrar o raciocínio por trás da representação de textos por meio de cadeias de Markov.

ALGORITMO TRADUTOR-MARKOV

Após demonstrar de que forma são utilizados alguns conceitos básicos de cadeias de Markov para modelar a descrição de produtos na NF-e, é apresentado o próximo passo. Com base em tais conceitos, foi desenvolvido o algoritmo TRADUTOR-MARKOV.

O primeiro termo do nome deve-se ao próprio objetivo do algoritmo: para cada produto procurado por um órgão público, encontrar sua descrição mais provável nas notas fiscais eletrônicas. Devido ao fato do campo descrição da NF-e estar limitado a 120 caracteres, ao preenchê-lo, o contribuinte tem uma limitação que os órgãos públicos, por exemplo, não têm ao montar seus catálogos de produtos. Por este motivo, estabelece-se o entendimento de que o contribuinte descreve os produtos "em um idioma ligeiramente diferente" daquele usado pelos órgãos públicos. Por essa razão, a utilização de um algoritmo "tradutor" de descrições de produtos (deixar ambos no mesmo idioma).

O segundo termo faz referência ao uso de algumas propriedades das cadeias de Markov.

PASSO A PASSO

1. DESCRIÇÃO DO PRODUTO A SER LICITADO

Nesta fase, a equipe técnica da SEFAZ RS recebe do órgão público uma lista com os produtos que se deseja licitar. Novamente para facilitar, seguirá sendo utilizado o exemplo apresentado anteriormente: ***"leite de vaca integral em pó"***, na descrição feita pelo próprio órgão.

Já que o objetivo é precificar o produto com base em transações realizadas, é necessário, antes de mais nada, poder identificar as transações deste produto no banco NF-e. Mas, será que os contribuintes costumam incluir a palavra VACA nas descrições dos leites integrais em pó? Quais palavras devemos utilizar na busca feita no banco NF-e, de forma a não restringir desnecessariamente a quantidade de transações identificadas? Ao conjunto destas palavras, nós demos o nome de TRADUÇÃO DO PRODUTO. É isso que o algoritmo TRADUTOR-MARKOV objetiva fazer: encontrar este conjunto de palavras para cada produto a ser precificado.

2. MONTANDO UM DICIONÁRIO

Para que o algoritmo possa traduzir as descrições dos produtos desejados, é necessário primeiramente criar um modelo do "idioma NF-e" para estes produtos em particular.

A primeira coisa a fazer é retirar da descrição original do produto palavras muito comuns e que, isoladamente, não sejam relevantes. No exemplo utilizado, novamente são retiradas as expressões "***de***" e "***em***". Restando, então: LEITE, VACA, INTEGRAL e PÓ.

Fazemos isso nesta fase por conta da forma como será feita a pesquisa no banco de dados: utilizando cláusula **OU**. Isso significa que buscaremos, no banco de dados NF-e, todas as descrições que tenham pelo menos uma das palavras LEITE, VACA, INTEGRAL ou PÓ.

Claro que é possível ter, por exemplo, LEITE DE CABRA, CARNE DE VACA, ARROZ INTEGRAL, ACHOCOLATADO EM PÓ. O que vai gerar uma quantidade considerável de dados (isso será tratado na sequência). No entanto, se mantidas as expressões "DE" e "EM" na busca ao banco de dados, a quantidade de resultados absolutamente fora de contexto aumentaria consideravelmente.

Nesta fase da análise, a busca ao banco de dados é limitada ao período de três meses de NF-e. Além disso, o único dado da NF-e a ser trazido como resultado, neste momento, é a descrição do produto.

Para a implementação desta fase, vai ser necessário rodar o script MONTANDO_DICIONARIO.R (disponível para download). Como resultado, ele grava o arquivo "palavras_busca_banco.csv" com o

conjunto de palavras que devem ser pesquisadas no banco de dados. É fundamental que esta pesquisa seja realizada com cláusula "**OU**".

Para possibilitar a demonstração do passo a passo de uma utilização real do algoritmo, foi gravado o resultado da consulta ao banco, no arquivo "resultado_primeira_query.csv"; também disponível para download.

3. EXECUTANDO O ALGORITMO TRADUTOR-MARKOV

Nesta seção, será demonstrado o passo a passo da execução do algoritmo TRADUTOR-MARKOV. Uma vez já formado o dicionário pela primeira busca ao banco de dados, é possível montar a matriz de transição de estados e ilustrar como o algoritmo "toma suas decisões".

Como já explicado, leva-se em conta a hipótese de que a primeira palavra da sequência seja a mesma em ambas descrições: tanto do órgão público quanto do contribuinte que preencheu a NF-e. Esta hipótese equivale a dizer que o primeiro estado da cadeia de Markov do exemplo é a palavra "***LEITE***" (estado s_1 no vetor *u*). Suponha a seguinte situação: estamos diante de uma NF-e e só podemos ver o início da descrição do produto. Já sabemos que a primeira palavra é "***leite***". Dentre as palavras que compõem o dicionário, qual a mais provável?

Para encontrar a resposta, devemos lembrar que o vetor "*u*" representa a probabilidade da cadeia encontrar-se em um determinado estado. Sabemos, ainda, que o estado atual é "LEITE", s_1. Sendo assim, sua probabilidade de ocorrência é igual a 1. Ele é o primeiro estado por hipótese. Forçosamente, os demais estados têm probabilidade ocorrência zero. O que resulta no seguinte vetor - probabilidade:

$$u = (p_1, p_2, p_3, p_4) \rightarrow u = (1,0,0,0)$$

Com a matriz de transição de estados calculada pelo algoritmo, podemos determinar o próximo estado da cadeia:

$$(1,0,0,0) \cdot \begin{bmatrix} 0 & 0.003220612 & 0.4946860 & 0.5020934 \\ 0.65217391 & 0 & 0.1739130 & 0.1739130 \\ 0.09090909 & 0 & 0 & 0.09090909 \\ 0.24019608 & 0 & 0.7598039 & 0 \end{bmatrix}$$

O vetor-probabilidade passa a ser, então:

$$u = (p_1, p_2, p_3, p_4) \rightarrow u = (0,0.003220612,0.494686,0.5020934)$$

O vetor indica como próximo estado mais provável o s_4 (probabilidade pouco acima de 0.5). Esse estado representa a palavra "***pó"***. O algoritmo seleciona, portanto, esta palavra como a próxima mais provável a compor a descrição do produto em uma NF-e preenchida pelo contribuinte.

Neste ponto, o algoritmo realiza duas ações:

- Coloca como primeiro estado do próximo passo, aquele de maior probabilidade. Ou seja, s_4 (PÓ);
- Retira do modelo a palavra "***leite***". Na iteração seguinte, o estado inicial será a palavra selecionada agora: "***pó***".

Tem-se, agora, apenas três estados: VACA, INTEGRAL e PÓ. E sua numeração mudou para, respectivamente, s_1, s_2 e s_3.

É preciso estar atento apenas para o fato de que, agora, o estado inicial é a palavra "***pó***". É como se o processo inteiro fosse iniciado novamente.

O vetor - probabilidade fica preenchido da seguinte forma:

$$u = (p_1, p_2, p_3) \rightarrow u = (0,0,1)$$

Agora o estado inicial (PÓ) é, por hipótese, s_3. Por isso, p_3=1.

Vamos, agora, calcular o próximo estado da cadeia.

$$(0,0,1) \cdot \begin{bmatrix} 0 & 0.333333 & 0.6666667 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

O vetor-probabilidade passa a ser, então:

$$u = (p_1, p_2, p_3) \rightarrow u = (0,1,0)$$

O estado s_2 representa a palavra "***integral"***; próximo estado da cadeia.

Novamente, o algoritmo realiza duas tarefas:

- Coloca como primeiro estado do próximo passo aquele de maior probabilidade. Ou seja, s2 (INTEGRAL);
- Retira do modelo a palavra "***pó"***. Na iteração seguinte, o estado inicial será a palavra selecionada agora: "***integral"***.

De forma análoga às outras iterações, tem-se apenas dois estados: VACA e INTEGRAL. E sua numeração mudou para, respectivamente, s_1 e s_2.

Deve-se atentar, novamente, para o fato de que o estado inicial mudou para a palavra INTEGRAL. O vetor - probabilidade fica preenchido da seguinte forma:

$$u = (p_1, p_2) \rightarrow u = (0,1)$$

Agora o estado inicial (***integral***) é, por hipótese, s_2. Por isso, p_2=1. Sendo assim, calcula-se o próximo estado da cadeia.

$$(0,1) \cdot \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Aqui, se tem uma situação diferente. O vetor-probabilidade resulta em:

$$u = (p_1, p_2) \rightarrow u = (0,0)$$

A interpretação deste resultado é que, dado o estado inicial, não haverá um próximo. Ou seja, não ocorrerá a transição do estado "***integral"*** para o estado "***vaca"***. Neste ponto, o algoritmo descarta a palavra "***vaca"*** e encerra sua execução salvando a tradução no arquivo "tokens_segunda_query.csv". Assim, a descrição final do produto ficou ***leite integral po***.

Neste momento, será necessário fazer uma nova consulta ao banco de dados NF-e, mas haverá diferenças cruciais em relação à primeira, sendo elas:

- Todas as palavras devem estar presentes nas notas fiscais. Ou seja, a busca deve ser feita de tal sorte que só sejam retornadas notas que contenham em sua descrição as palavras ""LEITE"" **E** "INTEGRAL" **E** "PO";
- O analista deve selecionar, além do campo (coluna) descrição do produto, aqueles relevantes para a precificação, como: preço, unidade, quantidade etc.

RESULTADOS DO ALGORITMO

O resultado desta segunda consulta ao banco de dados NF-e está salvo no arquivo "resultado_segunda_query.csv". Importante ressaltar que os dados ali presentes são os suficientes apenas para identificar o produto: descrição, quantidade e valores comercializados, data e GTIN (quando informado). Desta forma, fica garantida a preservação do sigilo fiscal.

É comparada a quantidade de NF-e identificadas utilizando a descrição do órgão público e a descrição proposta pelo algoritmo TRADUTOR-MARKOV:

- Descrição do órgão público: **3 transações**.
- Descrição proposta pelo algoritmo: **132.204 transações**.

Obviamente, a análise do resultado é bem mais complexa que contar as transações identificadas. Como o passo a passo deixou evidente, o algoritmo retira palavras que, em princípio, não são relevantes. Aqui é importante registrar duas observações:

OBSERVAÇÃO 1

Devido à forma como foi concebido, o algoritmo TRADUTOR-MARKOV nunca resultará em uma quantidade menor de transações quando comparado à descrição original. A razão é simples: ele até pode manter todas as palavras, mas jamais acrescentará. Assim, quando retira alguma palavra da descrição e faz a busca no banco de dados, está é menos restritiva e resultará, pelo menos, na mesma quantidade de transações;

OBSERVAÇÃO 2

O algoritmo faz o julgamento de relevância baseado puramente na matriz de transição de estados. É um cálculo puramente estatístico: se a probabilidade de ocorrência conjunta de dois estados quaisquer em uma mesma nota fiscal é alta, ela é relevante. Caso contrário, não. Segundo este critério, saber se o leite é de vaca ou de cabra não é relevante. Claro que isso é um absurdo, já que o órgão público deixou claro no edital que deseja comprar LEITE DE VACA INTEGRAL EM PÓ. Neste caso específico, uma análise mais detida do produto mostrou que o resultado era confiável, mesmo não tendo expressamente a palavra VACA (no caso desse exemplo, conclui-se que os contribuintes costumam escrever o nome do animal apenas quando não quiserem se referir a leite de vaca.

FRACIONAMENTO NÃO MEDICAMENTOS

Uma vez identificadas as NF-e para pesquisa, o analista pode começar a se debruçar sobre a definição do preço de referência. Antes disso, entretanto, o primeiro e, talvez, maior desafio, seja o fracionamento: selecionar a apresentação correta. No exemplo aqui utilizado, o leite em pó pode ser encontrado em diversas embalagens: caixa, saco ou lata. Além, claro, da própria massa. No exemplo, o órgão público queria na apresentação de 1 kg.

O algoritmo TRADUTOR-MARKOV não trata o fracionamento. Esta parte do pós-processamento é feita por um algoritmo específico, também criado por nossa equipe.

PRECIFICAÇÃO NÃO MEDICAMENTOS

É idêntica à metodologia já explicada para os medicamentos.