MATRIZ ORIGEM-DESTINO (O-D) DA CARGA AÉREA DOMÉSTICA ESTIMADA A PARTIR DE DADOS DOS DOCUMENTOS DE CONHECIMENTO DE TRANSPORTE ELETRÔNICO (CT-E)

Anderson Schmitt
Rafael Cardoso Cunha
Universidade Federal de Santa Catarina (UFSC)
Laboratório de Transportes e Logística (LabTrans)
Karla Andrea Rodrigues Dos Santos
Marcelo Leme Vilela
Secretaria Nacional de Aviação Civil (SAC/MInfra)
Departamento de Planejamento e Gestão (DPG/SAC)
Letícia Pinto da Silva
Amir Mattar Valente
Universidade Federal de Santa Catarina (UFSC)
Laboratório de Transportes e Logística (LabTrans)

RESUMO
Conhecer e qualificar a carga transportada por modo aéreo é uma importante etapa para o planejamento deste modo de transporte. Normalmente a coleta desse tipo de informação acontece por meio de observações e entrevistas com os agentes do setor. Em geral, essa coleta é custosa e demorada e, em alguns casos, há receio dos atores do setor em fornecer os dados de forma desagregada. Este trabalho apresenta a construção da matriz Origem-Destino (O-D) de carga aérea doméstica a partir de informações contidas nos documentos de Conhecimento de Transporte Eletrônico (CT-e). A metodologia apresenta etapas de coleta de dados, limpeza e tratamento dos dados dos documentos CT-es, incluindo um modelo de classificação dos tipos de carga utilizando regressão logística. A matriz O-D gerada é um avanço significativo na identificação e qualificação da demanda da carga aérea. As informações contidas na matriz incluem o tipo de carga, a real origem e o real destino da carga transportada por modo aéreo, tornando possível gerar mapas de linhas de desejo e áreas de influência dos aeroportos para carga.

ABSTRACT
Identifying and qualifying the air-cargo is a relevant planning step. The collection of this data usually happens through observations and interviews with the sector players. The process can be costly and time-consuming and, in some cases, the sector players can be uneasy to provide disaggregated data. This study developes the domestic air cargo origin-destination (O-D) matrix based on the Domestic Electronic Air Waybill (CT-e) information. The methodology comprises data collection and treatment of the CT-es documents, incorporating an air-cargo type classification model using logistic regression. The generated O-D matrix is a significant contribution that allows the identification of domestic air-cargo demand. The information contained in the O-D matrix includes the type of cargo, the real origin and the real destination of air-cargo, making it possible to generate maps and analyze the cargo desire lines and air-cargo catchment areas.

1. INTRODUÇÃO

Uma rede de transporte eficiente, que maximize a mobilidade das cargas e minimize o custo e tempo de deslocamento destas, é essencial para o desenvolvimento econômico do país. Por conta disto, há especial atenção das ações governamentais voltadas a uma nova visão da matriz de transporte de cargas, buscando equilibrar a participação dos modos de transporte. O transporte aéreo possui sua parcela de participação nesta matriz e, embora não seja alvo do transporte de grandes volumes de *commodities*, é adequado para o transporte de bens de alto valor agregado, principalmente os manufaturados e os produtos perecíveis (Brasil, 2018).

Para promover uma maior participação do transporte aéreo na matriz de transporte de carga são necessários estudos e levantamentos para identificação de potencialidades e demandas não atendidas (Brasil, 2018). Historicamente, dados estatísticos da demanda de carga aérea são

coletados e divulgados pela Agência Nacional de Aviação Civil (ANAC) em seu sítio online, no entanto a informação presente nos dados limita-se apenas ao volume de carga transportada entre aeroportos. Faz-se necessário, portanto, a obtenção de dados que informem a real origem e o real destino das cargas, bem como suas características. Normalmente essas informações são coletadas por meio de observações, entrevistas e dados dos agentes do setor. No setor de carga aérea, as companhias aéreas e os agentes de carga são os recebedores e os principais detentores desse tipo de dado e, por operarem em um ambiente competitivo, muitas vezes têm receio em fornecê-los de forma desagregada.

Conforme apresentado por Bruton (1975), a coleta, o tratamento e a disponibilização de bases de dados é um dos principais desafios enfrentados pelos planejadores de transporte, pois demanda esforço e tempo, além de recursos humanos e financeiros consideráveis. Tristão e Coelho (2018) dizem que, no início de um projeto de planejamento, é custosa a construção de uma base de dados para o fornecimento de informações voltadas à sua execução. As bases de dados necessárias normalmente envolvem diferentes fontes e estão disponíveis em formatos variados. Tavasszy e De Jong (2014) apresentam uma discussão sobre a escassez de dados. Os autores relacionam as fontes de dados mais adequadas conforme o tipo de modelagem em transportes a ser realizada. Entre essas, são citados dados provenientes de Identificação por Radiofrequência (RFID) e dos documentos administrativos que os emissores da carga devem legalmente preencher, o que, por sua vez, remetem aos documentos fiscais relacionados com o transporte da carga.

Santos (2015) aponta que, no caso do Brasil, há um significativo *big data* de fluxo de cargas, pois praticamente todas as transações comerciais são, por lei, amparadas por documentos fiscais digitais estruturados de modo padronizado. Essas informações são armazenadas, utilizadas e atualizadas continuamente sob uma mesma base de dados controladas pelos estados e pela Receita Federal do Brasil (RFB). Como resultado, atualmente existem vários documentos em formato digital com registros de dados relacionados ao transporte de cargas, a citar: a Nota Fiscal Eletrônica (NF-e), o Conhecimento de Transporte Eletrônico (CT-e) e o Manifesto Eletrônico de Documentos Fiscais (MDF-e). Estes são públicos, atualizados diariamente, podem interpretar a movimentação da carga com alta abrangência, e serem acessados se mantido o sigilo fiscal.

Recentemente estudos e métodos veem sendo conduzidos e aplicados nas análises dos documentos fiscais eletrônicos para a obtenção de dados de demanda da carga. Os trabalhos de Santos (2015) e Pipicano (2018) trouxeram importantes avanços nas análises das NF-e com aplicação em contexto urbano. Moreira e Brasil (2019) também aplicam a metodologia em contexto urbano e ampliam a análise para os documentos de MDF-e e CT-e. Por sua vez, Campos Júnior e Silva (2019) analisaram os documentos de MDF-e para a construção de uma matriz O-D do número de viagens de caminhões com aplicação limitada apenas ao Distrito Federal (DF). De modo geral, os trabalhos até então apresentam abrangência geográfica limitada, não abordam o transporte inter-regional de carga e focam esforços na identificação da demanda do transporte rodoviário.

O presente trabalho, por sua vez, objetiva estimar a matriz O-D inter-regional de carga aérea e identificar a real origem e destino da carga que utiliza o modo aéreo. Para isso utilizou-se uma metodologia que extrai, analisa e cruza dados dos documentos de CT-e armazenados pelas Secretarias de Fazendas (SEFAZ) de diferentes estados, os quais são organizados e

armazenados em todos os estados brasileiros.

2. CONHECIMENTO DE TRANSPORTE ELETRÔNICO (CT-E)

O CT-e surgiu a partir do Projeto da Nota Fiscal Eletrônica, que possuia como objetivo a implementação de um modelo nacional de documento fiscal eletrônico para substituir a emissão em papel. O CT-e é um modelo de documento fiscal emitido e armazenado eletronicamente, com validade jurídica garantida pela assinatura digital do emitente e pela autorização de uso fornecida pela administração tributária do domicílio do contribuinte. Atualmente, o CT-e é obrigatório para empresas que trabalham com transporte rodoviário, dutoviário, aéreo, ferroviário e aquaviário e pode ser utilizado para o transporte multimodal (Brasil, 2016).

O fluxo para a criação do CT-e é constituído pela geração de um arquivo eletrônico que possui informações fiscais de serviço de transporte, que deve ser assinado digitalmente para garantir a integridade dos dados e deverá ser transmitido, pela internet, para SEFAZ de jurisdição do contribuinte emitente. Esta, por sua vez, deve fazer uma pré-validação do arquivo e autorizar seu uso (Brasil, 2016).

As informações dos CT-es devem ficar disponíveis para consultas *on-line*, para o tomador do serviço e para outros que detenham a chave de acesso do documento eletrônico. Os documentos são, também, transmitidos pela SEFAZ para a RFB e para as SEFAZ de origem e destino da prestação, caso sejam diferentes da SEFAZ de circunscrição do emissor (Brasil, 2016).

2.1. Sigilo dos dados

De acordo com a Lei nº 12.527 de 18 de novembro de 2011, qualquer pessoa interessada poderá apresentar pedido de acesso às informações aos órgãos públicos integrantes dos poderes Executivo, Legislativo e Judiciário e do Ministério Público. Assim, os órgãos devem conceder o acesso à informação disponível, com exceção de informações total ou parcialmente sigilosas (Brasil, 2011). O conceito de sigilo das informações também está de acordo com a Lei nº 13.709 de 14 de agosto de 2018, que dispõe sobre o tratamento e a omissão de dados pessoais, inclusive nos meios digitais, por pessoa natural ou jurídica de direito público ou privado. (Brasil, 2018).

Tendo em vista as premissas supracitadas, para a utilização dos dados contidos nos documentos de CT-e, são necessários cuidados para garantir a privacidade dos usuários, como a omissão de informações que identificam empresas, tais como a razão social, o nome fantasia, o Cadastro Nacional da Pessoa Jurídica (CNPJ) ou inscrição estadual. Nesse contexto, os órgãos, as empresas ou as pessoas que coletam informações podem fornecê-las não de forma individual, mas agregadas e anonimizadas, ou seja, não se referindo aos indivíduos específicos.

3. METODOLOGIA E APLICAÇÃO

Para possibilitar a análise e a transformação dos dados dos CT-e na Matriz O-D de carga aérea doméstica realizou-se um fluxo de trabalho para a extração, a limpeza e o tratamento dos dados dos CT-es. Esse fluxo foi baseado no modelo Crisp-DM (do inglês, *Cross-Industry Standard Process for Data Mining*) aplicado em Pipicano (2018), que possui seis etapas:

- **Etapa 1 - Entendimento do assunto e definição dos objetivos**: Foram realizadas pesquisas para embasar teoricamente e definir objetivos, como os dados necessários, a forma de coleta e a avaliação da possibilidade de utilização de dados provenientes dos documentos CT-e.
- **Etapa 2 - Entendimento dos dados:** A partir do embasamento teórico assimilado na Etapa 1, foram selecionados os dados necessários a serem solicitados às SEFAZ. Além disso, a Etapa 2 engloba a familiarização e a avaliação da qualidade dos dados recebidos.
- **Etapa 3 – Preparação dos dados:** Normalmente os dados não estão prontos para uso. Nesta etapa aplica-se o tratamento básico nos dados recebidos, com o objetivo de padronizá-los para posterior aplicação de modelos de análise.
- **Etapa 4 - Modelagem e mineração dos dados:** Nesta etapa são aplicadas as técnicas de mineração de dados. Iniciou-se pela padronização de todos os tipos e as unidades de medida para, então, realizar a classificação dos tipos de carga em grupos e a formatação das matrizes O-D estaduais. Por fim, todas as bases foram agrupadas para formar a Matriz de Carga Aérea doméstica.
- **Etapa 5 - Avaliação dos modelos da etapa anterior:** Na etapa 5 são realizadas avaliações dos resultados obtidos na agregação dos dados e eventuais correções. Nesta etapa já são adiantadas análises dos resultados, com o objetivo de confrontá-los com dados esperados ou já conhecidos, como a base de dados estatísticos disponibilizada pela ANAC. Em caso de inconsistências, retorna-se à Etapa 4 para a avaliação e modificação do modelo.
- **Etapa 6 - Divulgação e implantação:** A última etapa se refere à divulgação dos resultados, à verificação da utilidade dos resultados e à utilização nas tarefas de tomada de decisão. Com a matriz O-D desenvolvida, foram realizadas análises dos resultados, dos fluxos de carga e do perfil da carga aérea doméstica.

3.1. Coleta e entendimento dos dados

Os dados dos documentos CT-e foram solicitados às SEFAZ por meio de ofício. Foram solicitados os dados de 14 estados: Amazonas, Ceará, Espírito Santo, Pará, Santa Catarina, Paraná, Rio de Janeiro, Rio grande do Norte, Rio Grande do Sul, São Paulo, Distrito Federal, Pernambuco, Minas Gerais e Bahia. Para a escolha dos estados foi considerado o volume de carga movimentada, conforme dados estatísticos da ANAC para o ano de 2018, sendo que esses estados movimentaram mais de 90% de toda a carga doméstica por modo aéreo.

Foram solicitados dados exclusivamente dos CT-e aéreos, de todos os meses do ano de 2018, os municípios de início e término da prestação de serviço, os aeroportos de origem, de passagem e destino da carga e as informações do valor da carga, produto predominante, quantidade e tipo de medida. Conforme apresentado no Manual de Orientações e padrões técnicos do CT-e (Brasil, 2016), nem todos os dados solicitados são de preenchimento obrigatório e/ou possuem formato padronizado de preenchimento, destacam-se com essa característica o tipo de medida e o produto predominante da carga transportada.

No processo de validação dos dados foram analisadas diversas características daqueles recebidos e, em alguns casos, foram solicitados novos dados às SEFAZ. Os itens que foram avaliados para determinar a qualidade dos dados recebidos foram:

- **Completude de informações:** verificou-se se a base de dados possui todas as informações necessárias para a análise, com todas as colunas solicitadas e as informações de todos

meses do ano. Conforme observado no processo de obtenção dos dados, essas características variavam devido à forma de armazenamento dos dados e também pelas características dos softwares de extração utilizados pelas SEFAZ.

- **Estrutura da base:** no geral, as bases recebidas apresentaram colunas separadas por ponto, vírgula ou ponto e vírgula. No entanto, algumas bases possuíam distintas separações de colunas, essas diferenças podem ser explicadas devido aos softwares utilizados para extrair os dados de cada SEFAZ estadual.
- **Número de linhas:** o número de linhas indica, em geral, a desagregação da base, ou seja, quanto maior o número de linhas maior a quantidade de informação. Dessa forma, foi feita a comparação do número de linhas entre os dados dos estados coletados, considerando também na comparação o volume de carga movimentada por cada estado, conforme dados estatísticos disponibilizados pela ANAC.
- **Quantidade de produtos predominantes únicos:** a diversificação de produtos predominantes mostrou-se essencial para uma análise satisfatória do perfil da carga.

3.2. Preparação dos dados

Cada SEFAZ disponibilizou os dados em uma estrutura diferente. No entanto, em geral, foram recebidos dois conjuntos de dados, apresentados nas Tabelas 1 e 2.

O conjunto de dados A (Tabela 1) apresenta uma coluna com o identificador único que relaciona as informações das duas tabelas de dados. A origem e o destino da carga são informados pelo código do IBGE dos munícipios, campo obrigatório e padronizado. O aeroporto de origem e destino é informado pelo código IATA, campo obrigatório e de preenchimento livre. O valor da carga é apresentado em formato numérico de preenchimento livre. O produto predominante, por sua vez, possui formato textual de preenchimento livre.

Tabela 1: Estrutura do conjunto de dados A

ID	PER_REF	CD_MUNIC ORIGEM	CD_MUNIC DESTINO	AERO ORIGEM	AERO DESTINO	VL CARGA	PROD PRED
Identificador numérico	Mês	Código IBGE	Código IBGE	Código IATA	Código IATA	Valor da carga	Produto predominante
45879	201807	4205407	3304557	FLN	SDU	2800	Carne de Siri

Tabela 2: Estrutura do conjunto de dados B

ID	PER_REF	NM TIPO_MEDIDA	QT_MEDIDA	CD_UNID
Identificador numérico	Mês	Tipo de medida	Quantidade	Unidade de medida
45879	201807	VOLUMES	2	3
45879	201807	PESO BRUTO	61	1
45879	201807	PESO CUBADO	61	1
45879	201807	CUBAGEM	0.5	0

Relativo ao conjunto de dados B, cada CT-e pode apresentar mais de um tipo de medida, conforme exemplo da Tabela 2. O campo de tipo de medida é em formato texto de entrada livre, a quantidade é em formato numérico de entrada livre, a unidade de medida é em formato numérico padronizado conforme Tabela 3.

Tabela 3: Tipos de unidade medidas existentes no CT-e

ID	PER_REF
0	m^3
1	kg
2	t
3	Unidade
4	Litros

Fonte: Brasil (2016).

Em alguns casos o conjunto de dados enviado pela SEFAZ era único, e apresenta, em uma mesma tabela, tanto as informações gerais do conjunto A quanto as informações das medidas da carga do conjunto B. Nessa configuração cada CT-e possui mais de uma linha na tabela.

Todas os dados recebidos das SEFAZ foram tratados e padronizados para a estrutura da Tabela 1 e 2. Além disso, outros procedimentos de padronização prévios foram realizados em todos os CT-es recebidos, sendo:

- Padronização do período de referência, indicando somente o número do mês
- Padronização das siglas do aeroporto de origem e destino para o padrão IATA, os que foram possíveis de serem identificados

3.3. Modelagem e mineração dos dados

Essa etapa possui como objetivos a padronização das unidades de medida da carga em cada CT-e e seleciona aquela a ser utilizada para compor a Matriz O-D; a classificação dos tipos de carga em grupos, facilitando o agrupamento e a análise posterior da Matriz O-D; a formatação das matrizes O-D estaduais e a consolidação em uma Matriz O/D de âmbito nacional.

3.3.1. Padronização das unidades de medida

Conforme apresentado no exemplo da Tabela 2, cada documento CT-e pode possuir mais de uma unidade de medida. Com o objetivo de escolher aquela utilizada na Matriz O-D, elaborou-se um procedimento de ordem de preferência entre os valores contidos em cada documento. A ordem de preferência da unidade utilizada foi peso bruto (kg), seguido pelo peso real (kg), peso taxado (kg) e peso cubado (kg).

3.3.2. Modelo para classificação dos tipos de carga

Devido à característica de texto aberto da informação de produto predominante (Tabela 1), a base de dados dos documentos de CT-e apresentou mais de 200 mil valores distintos, incluindo erros ortográficos. Para facilitar a análise, aplicou-se um modelo (Figura 1) para classificar os diversos tipos de carga presentes nos dados em grupos padronizados (Tabela 4).

Em razão do volume de dados únicos, a substituição direta torna-se inviável. Assim construiu-se um modelo para classificar o grupo de carga de cada registro. Para essa etapa, foi utilizado um modelo de classificação/predição por regressão logística. Esse modelo busca estimar a probabilidade de um objeto assumir um determinado valor em função de informações já conhecidas. O modelo é gerado a partir de um conjunto de dados de treinamento rotulados (arquivo treino), ou seja, trata-se de uma tabela, em que há a informação da classe à qual o objeto pertence para cada vetor de entrada (Aborisade e Anwar, 2018; Pranckevičius e

Marcinkevičius, 2017; Castro e Ferrari, 2016).

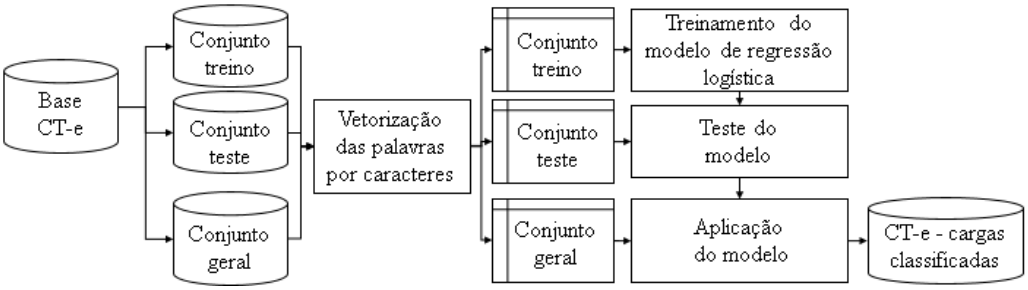

Figura 1: Fluxo para classificação dos tipos de carga

Tabela 4: Grupos de carga utilizados

Grupo de carga
Máquinas e eletrônicos
Produtos industrializados
Carga postal
Material de companhia aérea (COMAT)
Produtos perecíveis
Cargas perigosas e de alto risco
Produtos alimentícios
Medicamentos
Minérios, metais, pedras preciosas e outros artefatos
Veículos
Animais vivos

Na atual aplicação, as palavras do arquivo treino foram extraídas aleatoriamente dos documentos de CT-es. Foram selecionados e classificados manualmente mais de 6 mil itens contendo todos os grupos propostos e suas respectivas classificações. Tanto no treinamento do modelo de regressão logística, quanto para a classificação, foram aplicados os seguintes tratamentos das informações (Pranckevičius e Marcinkevičius, 2017):

- Remoção de palavras que não são consideradas importantes (conhecidas como *Stop Words*) ou que não trazem informações sobre o tipo da carga, como “e”, “para”, “peças”, “conjunto”, “amostra”, entre outras.
- Padronização de todas as letras para minúsculas.
- Remoção de acentuações e números.

Para a vetorização das palavras foram testados o modelo de separação por palavras (conhecido como *Bag of Words*) e o modelo de separação por caracteres (conhecido como *Char N-gram*) (Gómez-Adorno et al, 2018). Neste último o produto predominante é representado como um conjunto dos caracteres que formam suas palavras, conforme exemplo da Tabela 5.

Tabela 5: Exemplo de separação por caracteres - *Char N-gram*

Produto predominantes	Variáveis do modelo										
agulha descartavelp	agu	des	agul	artr	vel	cart	des	elp	esc	agu	lha
agulh descartevel	agu	des	agul	arte		cart	des		esc	ulh	evel
agulha descartavel	agu	des	agul	arta	vel	cart	des		esc	ulh	lha

Uma forma de avaliar o desempenho do modelo, ou seja, determinar o quanto ele é eficiente e

preciso para cada conjunto de dados, é medindo a sua habilidade preditiva através de métricas de desempenho (Castro e Ferrari, 2016). Para isso, o modelo foi aplicado em um conjunto de teste já classificado. Assim, a avaliação do desempenho na classificação deste oferece uma estimativa de capacidade de generalização do modelo, ou seja, sua capacidade de responder corretamente aos dados não usados no processo de treinamento.

O modelo final de regressão logística *Char N-gram* apresentou resultados entre 89% e 93% de assertividade, enquanto o modelo *Bag of Words* classificou corretamente entre 58% a 63% das cargas. Devido ao campo produto predominante do CT-e possuir características de texto aberto, há muitos casos de ocorrência de erros ortográficos, o que favoreceu o modelo de vetorização *Char N-gram*.

3.3.3. Consolidação matrizes estaduais em uma única matriz doméstica

A etapa de modelagem e mineração dos dados foi aplicada separadamente em cada base de dados estadual dos CT-e. A característica do CT-e de ser transmitido entre os estados de origem e destino do serviço de transporte causa uma possível intersecção e duplicação dos dados, o que impede o agrupamento simples das bases estaduais.

Para consolidar os pares O-D onde essa intersecção acontece, aplicou-se uma metodologia que usa critérios de qualidade para automatizar o procedimento de seleção. A metodologia empregada na escolha entre pares duplicados foi o *Analytic Hierarchy Process* (AHP), criado por Thomas L. Saaty (1980). Esse escolhe o melhor conjunto de dados entre os disponíveis, com a utilização dos critérios de qualidade:

- Número de tipos de carga (Tabela 4) únicos no par O-D (C1)
- Percentual de carga identificada do par O-D (C2)
- Quantidade de documentos de CT-e que compõem o par O-D (C3)

Após aplicação do método AHP, a síntese final dos pesos de cada critério é calculada, resultando na Equação 1 do Valor de Escolha (VE) do par.

$$VE = 0{,}114C1 + 0{,}479C2 + 0{,}405C3 \quad (1)$$

A Equação 1 é aplicada em cada conjunto de par duplicado. O par que possui o maior VE no conjunto avaliado é selecionado para compor a matriz O-D final.

4. RESULTADOS

A Figura 2 exibe a soma em kg da carga enviada e recebida em cada estado da Federação e Distrito Federal, obtidos nos documentos de CT-e. Para fins de comparação, a Figura 2 também apresenta os valores de carga embarcada e desembarcada provenientes da base de dados estatísticos da ANAC.

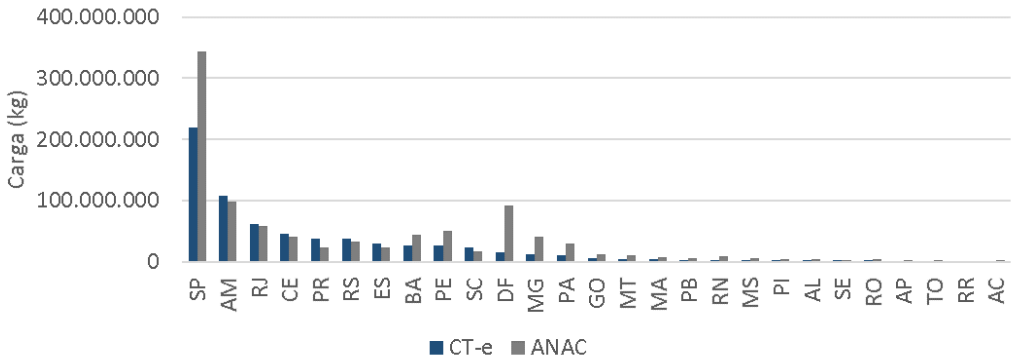

Figura 2: Carga enviada e recebida por região e estado brasileiro (2018)

Por sua vez, a Figura 3 evidencia a sazonalidade da movimentação de carga no ano de 2018 e compara o volume de carga proveniente dos CT-es e da ANAC.

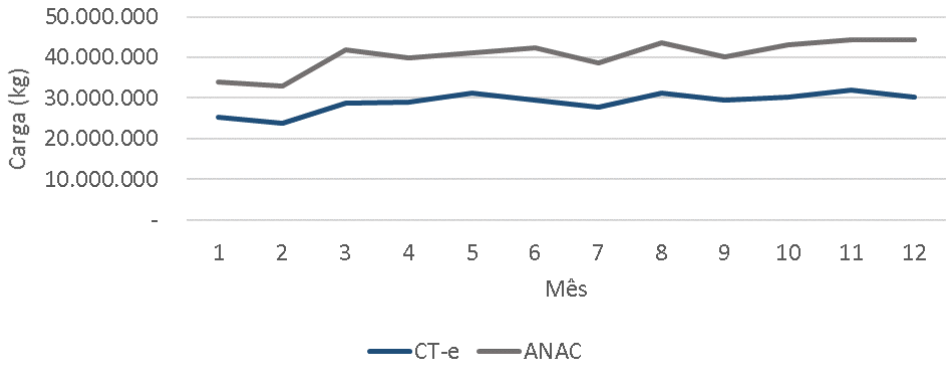

Figura 3: Sazonalidade da movimentação de carga – ano 2018 (CT-e vs. ANAC)

Os valores apresentados nas Figuras 2 e 3, indicam alta correlação geográfica e temporal entre os volumes de carga da matriz O-D (proveniente dos documentos de CTe) e os dados estatísticos da ANAC. Nota-se que o volume de carga obtido dos dados CT-e, em geral, é menor do que o volume dos dados estatísticos da ANAC. Esse resultado está de acordo com o esperado, pois os dados estatísticos da ANAC apresentam informações operacionais, ou seja, informam o volume total de carga transportada em cada voo/rota, enquanto que os dados contidos no CT-e apontam a real origem e destino da carga, sem registrar as conexões realizadas pela carga. Destaca-se a maior diferença no volume de carga no estado de SP e no DF, estados que possuem aeroportos utilizados como *Hub* de carga pelas companhias aéreas.

4.1. Tipo de carga aérea doméstica movimentada

A Tabela 6 apresenta a participação de cada tipo de carga aérea movimentada domesticamente.

Tabela 6: Participação dos tipos de carga aérea na movimentação doméstica - 2018

Grupo de carga	%
Produtos industrializados	22,3
Carga postal	20,4
Máquinas e eletrônicos	15,1
Veículos e suas partes	13,9
Medicamentos	12,1

Grupo de carga	%
Produtos perecíveis	7,1
Produtos alimentícios	5,0
Cargas perigosas e de alto risco	2,2
Minérios, metais, pedras preciosas e outros artefatos	0,9
Material de companhia aérea (COMAT)	0,8
Animais vivos	0,2

A classe de Produtos industrializados apresentou a movimentação mais expressiva, com 22,3 % do total movimentado no País. Também se destacaram as classes de Carga postal/documentos (20,4 %), Máquinas e eletrônicos (15,1 %), Veículos e suas partes (13,9 %) e Medicamentos (12,1 %). Entre os Produtos industrializados e Máquinas e eletrônicos, foi possível identificar um volume expressivo de cargas provenientes de *e-commerce*, como calçados, roupas, acessórios, utensílios domésticos, equipamentos de informática e eletrodomésticos. A Figura 4 apresenta as linhas de desejo dos principais tipos de carga identificados.

A matriz O-D gerada confirma a concentração dos maiores volumes transportados em poucas regiões, principalmente do Sudeste brasileiro. Os 20 maiores pares O-D identificados representam mais de 50% de toda a carga doméstica movimentada, destacando-se o par São Paulo (SP) – Manaus (AM) como o maior par O-D doméstico. Dos maiores pares, apenas dois não possuem origem e destino localizado na Região Sudeste, a saber, o par Fortaleza (CE) – Manaus (AM) e Salvador (BA) – Fortaleza (CE).

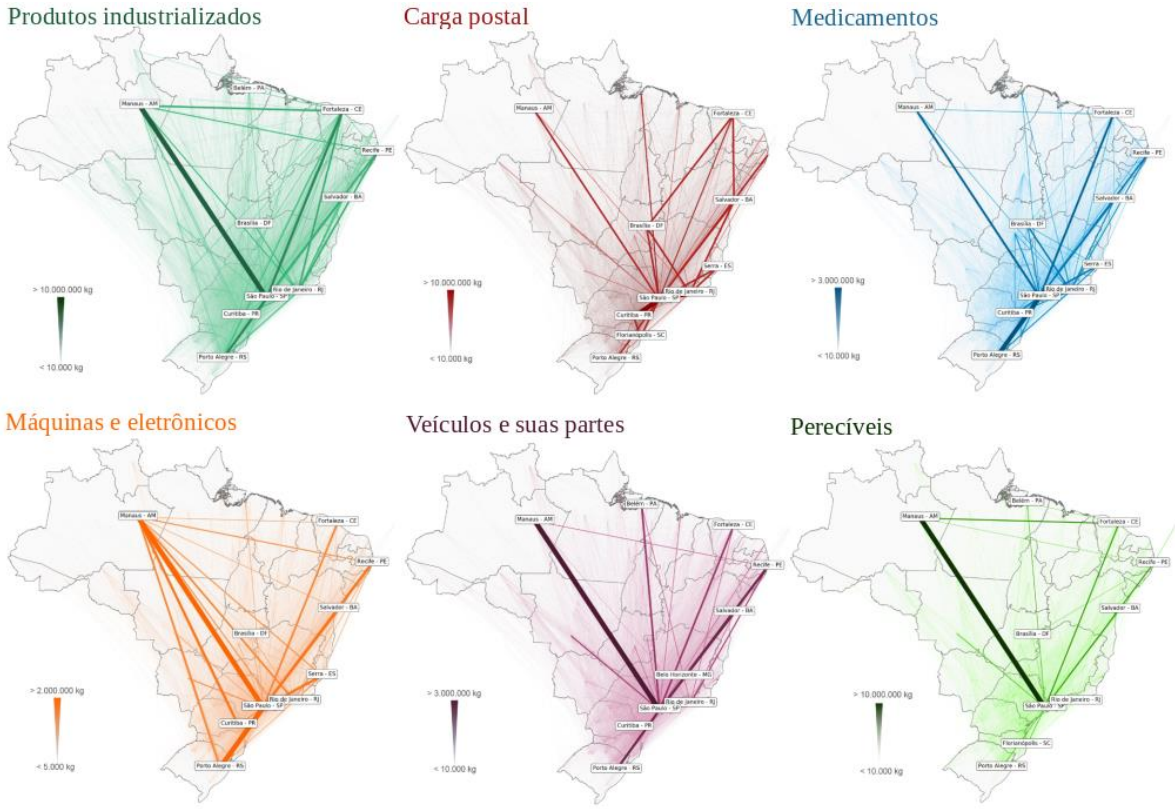

Figura 4: Linhas de desejo dos principais tipos de carga

4.2. Área de influência dos aeroportos para carga

A partir da matriz O-D é possível também identificar a área de influência dos aeroportos para captação e distribuição de carga. Para cada aeroporto, são identificados os municípios de origem e destino das cargas e os volumes de carga gerados e recebidos a partir deles. Dessa forma, é possível delimitar as áreas de influência dos aeroportos mediante os deslocamentos de carga observados. A Figura 5 apresenta as áreas de influência dos aeroportos de Campinas (SBKP) e Fortaleza (SBFZ).

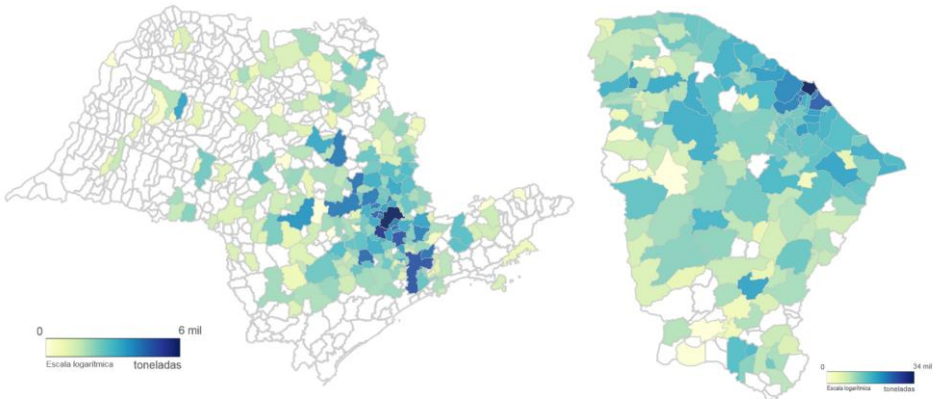

Figura 5: Área de influência para carga doméstica dos aeroportos de Campinas - SBKP (esquerda) e Fortaleza - SBFZ (direita)

8. CONSIDERAÇÕES FINAIS

A matriz O-D da carga aérea doméstica gerada a partir de dados dos documentos de CT-e apresenta um avanço significativo na identificação e qualificação da demanda desse modo de transporte. Com ela é possível identificar a real origem e o real destino, além do tipo de carga transportada, informações importantes para o planejamento setorial da carga aérea e para os atores do setor.

Durante o processo de análise dos dados, as principais dificuldades encontradas para a obtenção da matriz O-D referem-se à qualidade do preenchimento dos documentos de CT-e. De modo geral, observaram-se erros de digitação e inconsistências em diversos campos, principalmente, nos destinados à identificação do produto predominante e dos tipos de medida. Em muitos casos, o produto predominante foi preenchido como o número da nota fiscal, ou com um termo geral, sem constar a definição real do tipo de carga. Com relação aos tipos de medida, não há uma padronização de nomenclaturas e, muitas vezes, os pesos e os volumes são cadastrados incorretamente, com algarismos a mais, por exemplo. Em um contexto semelhante ao observado neste trabalho, Pipicano (2018) apontou que os dados das NF-e apresentaram problemas de consistência nas informações. Também, segundo Santos (2015), na análise de 408.000 NF-e cerca de 30% delas mostraram inconsistências nos campos relacionados à carga. Essa inconsistência possivelmente ocorre devido ao fato de os emissores não serem fiscalizados pela ausência desses dados, visto que não seria possível inspecionar todas as cargas e comparar com o registro dos documentos fiscais.

Embora os dados contidos nos documentos do CT-e não sejam perfeitos, a base de dados apresenta informações de toda a população e não somente de uma amostra, pois toda a carga transportada por modo aéreo deve estar amparada por um documento CT-e aéreo. Assim, com um tratamento adequado dos dados, incluindo a identificação e o tratamento dos erros

encontrados, a base de dados apresenta-se como uma completa e valiosa fonte de informações.

Como aperfeiçoamento deste trabalho, pretende-se realizar um estudo de demanda reprimida com a identificação da origem e destino de cargas rodoviárias. Algumas dessas cargas poderiam ser transportadas pelo modo aéreo devido às características de segurança e rapidez deste modo. A identificação das cargas rodoviárias pode ser feita utilizando-se a mesma fonte (nesse caso específico, dados de CT-es rodoviários) e a mesma metodologia e técnicas apresentadas no atual trabalho. Esse estudo apresentaria informações importantes para entender a dinâmica do setor e contribuiria com o objetivo de equilibrar a participação dos modos de transporte na matriz de carga nacional.

Por fim, levando-se em consideração o apresentado e de acordo com Santos (2015), recomenda-se o aprofundamento de iniciativas integradas entre os órgãos fiscais e de planejamento. Essas iniciativas devem buscar a estruturação das bases de dados fiscais, incluindo principalmente mecanismos que permitam disponibilizá-las em escala suficiente para que pesquisadores e gestores públicos possam utilizá-las em estudos de planejamento.

Agradecimentos
Os autores agradecem às Secretarias de Fazenda estaduais, que disponibilizaram os dados para o trabalho, à Secretaria Nacional de Aviação Civil (SAC/MInfra) e à Fundação de Amparo à Pesquisa Universitária (FAPEU).

REFERÊNCIAS BIBLIOGRÁFICAS

Aborisade, M. A. e Anwar, M. (2018). Classification for Authorship of Tweets by Comparing Logistic Regression and Naive Bayes Classifiers. IEEE International Conference on Information Reuse and Integration for Data Science (IRI), Salt Lake City.

Brasil (2011). Presidência da República. Lei nº 12.527, de 18 de novembro de 2011. Diário Oficial da União, Brasília.

Brasil (2018). Presidência da República. Lei nº 13.709, de 14 de agosto de 2018. Diário Oficial da União, Brasília.

Brasil. SAC (2018). *Plano aeroviário nacional 2018-2038*. Ministério dos Transportes, Portos e Aviação Civil, Brasília.

Brasil (2016). Projeto Conhecimento de Transporte Eletrônico. Manual de Orientações do Contribuinte: padrões técnicos de comunicação. Encontro Nacional de Coordenadores e Administradores Tributários Estaduais (ENCAT), Brasília.

Bruton, M. J (1979). Introdução ao Planejamento dos Transportes. Rio de Janeiro: Interciência da Universidade de São Paulo (EDUSP).

Campos Júnior, N. F. R. e Silva, A. R. (2019). Geração da matriz origem-destino para o transporte rodoviário de carga usando o manifesto eletrônico de documentos fiscais. Anais do 33º Congresso de Pesquisa e Ensino em Transportes, ANPET, Balneário Camboriú, v. 1, p. 2511–2522.

Castro, L. N., Ferrari, D. G (2016). Introdução à mineração de dados conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva. ISBN 978-85-472-0099-2.

Gómez-Adorno H. et al (2018). Document embeddings learned on various types of n-grams for cross-topic authorship attribution. Computing Vol. 100, 741–756. Springer.

Moreira, C. M. e Brasil, J. C. (2019). Metodologia da pesquisa origem e destino de cargas com dados fiscais da região metropolitana de Belo Horizonte. Anais do 33º Congresso de Pesquisa e Ensino em Transportes, ANPET, Balneário Camboriú, v. 1, p. 2269–2278.

Pipicano, E. F. M (2018). Metodologia para o planejamento do transporte urbano de carga usando dados de documentos fiscais eletrônicos. 2018. Tese (Doutorado em Transportes) – Faculdade de Tecnologia, Universidade de Brasília, Brasília.

Pranckevičius T. e Marcinkevičius V. (2017). Comparison of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic J. Modern Computing, Vol. 5 (2017), No. 2, 221-232.*

Saaty, T. L (1980). The Analytic Hierarchy Process. McGraw-Hill, New York.

Santos, E. M. (2015) Uso de dados de documentos fiscais eletrônicos para o planejamento de transporte urbano de cargas. 2015. Tese (Doutorado em Transportes) – Faculdade de Tecnologia, Universidade de Brasília, Brasília.

Tavasszy, L. e De Jong, G. (2014). Data Availability and Model For in Modelling Freight Transport. Elsevier. ISBN r978-0-12-410400-6.

Tristão, L e Coelho A. H. (2018). Proposta de uma base de dados integrada para apoiar no planejamento do setor de aviação civil brasileiro. Anais do 32º Congresso de Pesquisa e Ensino em Transportes, ANPET, Gramado, v. 1, p. 1037–1047.

Anderson Schmitt (anderson_schmitt@yahoo.com.br)
Rafael Cardoso Cunha (ccunha.rafael@gmail.com)
Letícia Pinto da Silva (leticiadasilvalps@gmail.com)
Amir Mattar Valente (amir.valente@ufsc.br)
Laboratório de Transportes e Logística (LabTrans), Departamento de Engenharia Civil, Universidade Federal de Santa Catarina (UFSC), Rua João Pio Duarte da Silva, 205, 88040-900, Florianópolis, SC, Brasil
Karla Andrea Rodrigues Dos Santos (karla.santos@infraestrutura.gov.br)
Marcelo Leme Vilela (marcelo.vilela@infraestrutura.gov.br)
Departamento de Planejamento e Gestão (DPG), Secretaria Nacional de Aviação Civil (SAC), Ministério da Infraestrutura (MInfra).