

Base de Dados de Defesa Comercial Brasileira (2010–2025): Metodologia de Construção

Combinando dados administrativos do MDIC com fluxos de importação do ComexStat via pacote `comexTL`

Alan Leal

Economista, Theta Lab

Sumário Executivo

Este documento descreve a metodologia de construção da base de dados de defesa comercial brasileira – cobrindo medidas antidumping, compensatórias e de salvaguarda vigentes de janeiro de 2010 em diante. A base combina dois insumos principais: (1) uma planilha administrativa obtida junto ao MDIC via Lei de Acesso à Informação (LAI, protocolo 52016.000348/2026-79), contendo registros de investigações e medidas aplicadas; e (2) dados de fluxo de importação do ComexStat, acessados pelo pacote R **comexTL** (Theta Lab). O resultado é um painel enriquecido com variáveis de contexto e efeito para cada investigação, pronto para análises econométricas e de política comercial.

ORIGEM DOS DADOS

Pedido de Acesso à Informação (LAI)

A base primária desta pesquisa foi obtida mediante pedido formal ao Ministério do Desenvolvimento, Indústria, Comércio e Serviços (MDIC) pela Lei de Acesso à Informação (Lei n.º 12.527/2011). O pedido (protocolo **52016.000348/2026-79**, cadastrado em 08/02/2026) solicitou uma base em formato Excel (.xlsx) com os seguintes campos:

- Produto investigado;
- Tipo de medida (direito antidumping definitivo ou outro tipo);
- Origem dos produtos sancionados;
- Prazo de vigência (início e fim) em colunas separadas;
- Código NCM do(s) produto(s) sancionado(s);
- Valor do direito aplicado por produto e país.

O pedido foi registrado como Acesso à Informação Pública, com esfera federal, destinado ao MDIC. A situação foi concluída dentro do prazo estabelecido (02/03/2026).

A planilha devolvida pelo MDIC (*Planilha Base de dados de defesa comercial (57968866).xlsx*) contém registros de investigações abertas a partir de 2010, incluindo processos encerrados com e sem aplicação de direito, investigações em curso e revisões de medidas vigentes.

Dados de Fluxo de Importação — comexTL

Os dados de importação utilizados para enriquecer a base provêm do ComexStat (MDIC), plataforma oficial de estatísticas de comércio exterior brasileiro. O acesso foi feito exclusivamente pelo pacote R **comexTL** (Theta Lab, Alan Leal), disponível em github.com/alanleal-econ/comexTL. O pacote baixa arquivos Parquet hospedados em infraestrutura Cloudflare R2, com cobertura anual e sistema de cache local para otimizar consultas repetidas. A função utilizada foi `comex_stat_geralimps()`, que retorna importações agregadas por NCM, país e ano, com as métricas VL_FOB e KG_LIQUIDO.

ESTRUTURA DA BASE ORIGINAL (MDIC/LAI)

A planilha recebida do MDIC foi carregada ignorando as duas primeiras linhas (título e linha em branco), sendo a linha 3 o cabeçalho. As colunas foram renomeadas para identificadores limpos e consistentes:

Campo Original (MDIC)	Nome Interno	Descrição
Produto	PRODUTO	Nome do produto investigado
Origem	ORIGEM	País ou grupo investigado
Tipo da investigacao	TIPO_INV	Antidumping, compensatório, salvaguarda
Subtipo da investigacao	SUBTIPO_INV	Original, revisão, extensão etc.
NCM	NCM	Código(s) NCM – pode haver múltiplos
Setor	SETOR	Setor industrial
Status da investigacao	STATUS_INV	Status corrente da investigação
Status da medida	STATUS_MEDIDA	Vigente, suspensa, revogada etc.
Data da abertura	DATA_ABERTURA	Data de início da investigação
Data encerramento / início vigência	DATA_VIGENCIA	Âncora temporal principal
Direito aplicado	DIREITO_APLICADO	Valor ou faixa do direito
Término da vigencia	DATA_TERMINO_VIGENCIA	Data de expiração da medida

PRÉ-PROCESSAMENTO E NORMALIZAÇÃO

Variáveis derivadas da base administrativa

A partir das datas e campos de status, foram calculadas as seguintes variáveis auxiliares:

Variável	Lógica de construção
ANO_ABERTURA	year(DATA_ABERTURA)
ANO_VIGENCIA	year(DATA_VIGENCIA)
DIAS_INVESTIGACAO	DATA_VIGENCIA – DATA_ABERTURA (inteiro)
MEDIDA_APLICADA	TRUE se STATUS_INV contém “aplicação de direito”, “compromisso de preço”, “estendida” ou “alterada”
TEVE_PROVISORIO	TRUE se DATA_APLIC_PROVISORIO não é NA
CHINA_FLAG	TRUE se ORIGEM == "china" (case-insensitive)
TODOS_FLAG	TRUE se ORIGEM == "todos" (salvaguardas globais)

Normalização de NCMs múltiplos

Alguns registros da base do MDIC contêm múltiplos códigos NCM em uma única célula, separados por ponto-e-vírgula (;), quebra de linha (\n) ou dois ou mais espaços consecutivos. O script expande essas células criando uma linha por NCM (unnesting), garantindo granularidade adequada para os joins com o ComexStat. Após a expansão:

- Pontos são removidos do código NCM (ex.: 7208.51.00 → 72085100);
- O código é completado com zeros à esquerda até 8 dígitos;
- Entradas vazias, com menos de 8 dígitos ou com letras são descartadas;
- São derivadas as colunas SH6_NUM (6 dígitos iniciais) e SH2_NUM (2 dígitos iniciais).

Mapeamento de países

O campo ORIGEM da base administrativa contém nomes textuais de países, enquanto o ComexStat utiliza códigos numéricos CO_PAIS. O mapeamento foi feito em duas etapas:

- **Match exato (case-insensitive):** compara ORIGEM com NO_PAIS e NO_PAIS_ING da tabela PAIS.csv carregada via comexTL (load_reference_tables("3"));
- **Match parcial (grep1):** se o match exato falhar, busca o nome como substring nos campos de nome do país;
- **Dicionário de correções manuais:** casos residuais com grafias alternativas ou países extintos são tratados por lookup estático.

Registros com TODOS_FLAG == TRUE (salvaguardas globais) não recebem CO_PAIS e ficam com NA nas variáveis de fluxo de importação por país. Países extintos (URSS, Tchecoslováquia) e investigações anteriores a 1997 também resultam em NA.

ENRIQUECIMENTO COM DADOS DE IMPORTAÇÃO (COMEXTL)

Para cada linha da base normalizada, são calculadas variáveis de importação em quatro janelas temporais, ancoradas nos anos-chave de cada investigação:

Janela	Âncora	Variáveis geradas
T0 – Abertura	ANO_ABERTURA	IMP_PAIS_TO, IMP_TOTAL_TO, MSHARE_TO
V0 – Vigência	ANO_VIGENCIA	IMP_PAIS_VO, IMP_TOTAL_VO, MSHARE_VO
V1 – 1 ano após	ANO_VIGENCIA + 1	IMP_PAIS_V1, IMP_TOTAL_V1, MSHARE_V1
V3 – 3 anos após	ANO_VIGENCIA + 3	IMP_PAIS_V3, IMP_TOTAL_V3, MSHARE_V3

Para cada janela, IMP_PAIS é o valor FOB (US\$) importado do país investigado no código SH6 correspondente ao NCM da investigação; IMP_TOTAL é o total importado do SH6 por todos os países; e MSHARE é a participação percentual do país no total ($IMP_PAIS / IMP_TOTAL \times 100$). Adicionalmente, são calculadas as variações percentuais:

- VAR_IMP_VO_V1: variação de IMP_PAIS entre V0 e V1 (%);
- VAR_IMP_VO_V3: variação de IMP_PAIS entre V0 e V3 (%).

Processo de download e cache

Os dados de importação foram baixados com a seguinte chamada ao comexTL:

```
df_imp_raw <- comex_stat_geralimps(
  anos      = anos_disponiveis,
  agregar_por = c("CO_ANO", "CO_PAIS"),
  metricas  = c("VL_FOB", "KG_LIQUIDO")
)
```

O parâmetro `anos_disponiveis` foi definido como o conjunto de anos únicos presentes nas colunas ANO_ABERTURA, ANO_VIGENCIA, ANO_VIGENCIA+1 e ANO_VIGENCIA+3 da base normalizada. O `comexTL` armazena os arquivos Parquet baixados em cache de sessão, evitando downloads duplicados para o mesmo ano.

ESTRUTURA DA BASE ENRIQUECIDA

O dataset final (*defesa_comercial_enriquecida.xlsx*) é organizado em uma linha por combinação investigação × NCM, com os seguintes grupos de colunas:

Grupo	Colunas	N
Identificação	PRODUTO, ORIGEM, TIPO_INV, SUBTIPO_INV, NCM, NCM_LIMPO, SH6_NUM, SH2_NUM, SETOR	9
Status	STATUS_INV, STATUS_MEDIDA, MEDIDA_APLICADA, TEVE_PROVISORIO, CHINA_FLAG, TODOS_FLAG	6
Datas e anos-chave	DATA_ABERTURA, ANO_ABERTURA, DATA_VIGENCIA, ANO_VIGENCIA, DATA_TERMINO_VIGENCIA, DIAS_INVESTIGACAO	6
Administrativo	Circulares, datas de determinação preliminar, direito provisório, resolução de encerramento etc.	12
Código de país	CO_PAIS_FINAL	1
Fluxo – T0	IMP_PAIS_TO, IMP_TOTAL_TO, MSHARE_TO	3
Fluxo – V0	IMP_PAIS_VO, IMP_TOTAL_VO, MSHARE_VO	3
Fluxo – V1	IMP_PAIS_V1, IMP_TOTAL_V1, MSHARE_V1	3
Fluxo – V3	IMP_PAIS_V3, IMP_TOTAL_V3, MSHARE_V3	3
Variações	VAR_IMP_VO_V1, VAR_IMP_VO_V3	2

LIMITAÇÕES E NOTAS METODOLÓGICAS

- **Cobertura temporal do ComexStat:** dados de importação estão disponíveis a partir de 1997. Investigações anteriores a esse ano resultam em NA nas variáveis de fluxo.
 - **Países extintos:** URSS, Tchecoslováquia e outras entidades sem código CO_PAIS ativo no ComexStat não têm dados de importação.
 - **Salvaguardas globais:** quando ORIGEM == "Todos", não há país específico para cruzar com o ComexStat; as variáveis IMP_PAIS e MSHARE ficam como NA.
 - **Granularidade SH6 vs. NCM:** os dados de importação são agregados ao nível SH6 (6 primeiros dígitos do NCM). A agregação pode incluir NCMs adicionais não relacionados à investigação.
 - **Ano V3 no futuro:** investigações encerradas após 2022 terão NA em IMP_PAIS_V3 e MSHARE_V3, pois os dados de 2025+ ainda não estão disponíveis no ComexStat.
 - **Múltiplos NCMs:** investigações com vários NCMs geram múltiplas linhas – uma por NCM. Análises em nível de investigação devem agregar ou selecionar o NCM principal.
-

Fontes e Referências

MDIC – Secretaria de Comércio Exterior (SECEX). Base de dados de medidas de defesa comercial (antidumping, compensatórias e salvaguardas), obtida via LAI, protocolo 52016.000348/2026-79. Ministério do Desenvolvimento, Indústria, Comércio e Serviços, 2026.

Leal, A. (2024–2026). *comexTL: Dados do Comércio Exterior Brasileiro no R.* Theta Lab. Disponível em: github.com/alanleal-econ/comexTL.

MDIC – ComexStat. Estatísticas de Comércio Exterior. Dados oficiais de importação e exportação do Brasil. Disponível em: comexstat.mdic.gov.br.

As opiniões expressas são de responsabilidade do autor. **Contato:** alan.leal@thetalab.com.br