



# PRF Accident Clustering

Uma análise end-to-end de 18 anos de acidentes  
(2007-2024).

Vinícius Santos Monteiro - ICMC/USP - 2025



# O Problema e o Desafio

Os dados da PRF são públicos e de livre e fácil acesso.



gov.br | Ministério da Justiça e Segurança Pública | Órgãos do Governo | Acesso à Informação | Legislação | Acessibilidade | Entrar com gov.br

Polícia Rodoviária Federal

O que você procura?

Acesso à Informação > Dados Abertos > Dados Abertos da PRF

## Dados Abertos da PRF

Publicado em 19/04/2022 06h50 | Atualizado em 06/10/2025 15h43

Compartilhe: f x in

Dados Abertos são dados institucionais, disponibilizados em formato legível por máquina e sem restrição de licenças, patentes ou mecanismos de acesso, que qualquer pessoa pode livremente usá-los, reutilizá-los e redistribuí-los.

Os dados classificados como abertos podem ser utilizados de várias formas, seja pelo próprio governo ou pela sociedade, como, por exemplo, no desenvolvimento de aplicativos, que exibem informações de forma gráfica e interativa.

A elaboração do Plano de Dados Abertos da PRF vem ao encontro do disposto na **Lei de Acesso à Informação (LAI)**, na **Instrução Normativa SLTI nº 4 de abril de 2012** (que institui a Infraestrutura Nacional de Dados Abertos), no **Decreto nº 8.777, de 11 de maio de 2016** (que institui a Política de Dados Abertos no Executivo Federal), bem como dos compromissos assumidos pelo Brasil no âmbito do Plano de Ação Nacional de Governo Aberto.



Base de dados:	BAT: Boletim de Acidente de Trânsito
Informações:	Unidade responsável: DIOP. Frequência de atualização: Mensal
Referência	Link
Documento CSV de Acidentes 2025 (Agrupados por ocorrência)	Baixar planilha
Documento CSV de Acidentes 2025 (Agrupados por pessoa)	Baixar planilha
Documento CSV de Acidentes 2025 (Agrupados por pessoa - Todas as causas e tipos de acidentes)	Baixar planilha
Documento CSV de Acidentes 2024 (Agrupados por ocorrência)	Baixar planilha
Documento CSV de Acidentes 2024 (Agrupados por pessoa)	Baixar planilha
Documento CSV de Acidentes 2024 (Agrupados por pessoa - Todas as causas e tipos de acidentes)	Baixar planilha

**PRF - Dados Abertos:**

<https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf>



# O Problema e o Desafio

Os dados da PRF são públicos, mas muito fragmentados.



**2007 - 2016**

**26 variáveis**

**DICIONÁRIO DE VARIÁVEIS**  
(DADOS DO BR-BRASIL – 2007 A 2016)

ID VARIÁVEL	NOME DA VARIÁVEL	DESCRIÇÃO
1	<i>id</i>	Variável com valores numéricos, representando o identificador do acidente.
2	<i>data_inversa</i>	Data da ocorrência no formato dd/mm/aaaa.
3	<i>dia_semana</i>	Dia da semana da ocorrência. Ex.: Segunda, Terça, etc.
4	<i>horario</i>	Horário da ocorrência no formato hh:mm:ss.
5	<i>uf</i>	Unidade da Federação. Ex.: MG, PE, DF, etc.
6	<i>br</i>	Variável com valores numéricos,

**PRF - Dados Abertos:**

<https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf>



# O Problema e o Desafio

Os dados da PRF são públicos, mas muito fragmentados.



**2017 - Atual**

**30 variáveis**

**+ Latitude, longitude,  
delegacia...**

**PRF - Dados Abertos:**

<https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf>

**DICIONÁRIO DE VARIÁVEIS**  
(DADOS DO BAT – A PARTIR DE 2017)

ID VARIÁVEL	NOME DA VARIÁVEL	DESCRIÇÃO
1	<i>id</i>	Variável com valores numéricos, representando o identificador do acidente.
2	<i>data_inversa</i>	Data da ocorrência no formato dd/mm/aaaa.
3	<i>dia_semana</i>	Dia da semana da ocorrência. Ex.: Segunda, Terça, etc.
4	<i>horário</i>	Horário da ocorrência no formato hh:mm:ss.
5	<i>uf</i>	Unidade da Federação. Ex.: MG, PE, DF, etc.
6	<i>br</i>	Variável com valores numéricos,

# O Problema e o Desafio

Os dados da PRF são públicos, mas muito fragmentados.

**2007**

	A	B	C	D	E	F	G
1	id	data_inversa	dia_semar	horario	uf	br	km
2	10	11/06/2007	Segunda	15:30:00	MG	381	623.2
3	10	11/06/2007	Segunda	15:30:00	MG	381	623.2
4	1032898	13/08/2007	Segunda	14:25:00	MG	40	585.5
5	1051130	12/02/2007	Segunda	02:10:00	MA	135	11
6	1066824	20/11/2007	Terça	05:30:00	CE	222	30.8
7	1069918	16/12/2007	Domingo	17			

**2024**

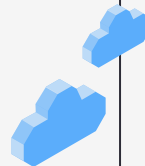
	A	B	C	D	E	F	G
1	id	data_inversa	dia_semana	horario	uf	br	km
2	571789	2024-01-01	segunda-feira	03:56:00	ES	101	38
3	571804	2024-01-01	segunda-feira	04:50:00	PI	343	185
4	571806	2024-01-01	segunda-feira	04:30:00	BA	116	459,7
5	571818	2024-01-01	segunda-feira	06:30:00	SE	101	18
6	571819	2024-01-01	segunda-feira	05:00:00	MT	364	240
7	571820	2024-01-01	segunda-feira	11:50:00	MG	251	447

**2016**

	A	B	C	D	E	F	G
1	id	data_inversa	dia_semana	horario	uf	br	km
2	36727	10/06/16	Sexta	18:30:00	RJ	101	66
3	83425846	01/01/16	Sexta	01:30:00	SC	101	135,5
4	83425850	01/01/16	Sexta	01:00:00	PR	277	2
5	83425852	01/01/16	Sexta	01:45:00	PR	476	357
6	83425853	01/01/16	Sexta	02:00:00	SE	101	94,2
7	83425855	01/01/16	Sexta	02:20:00	SC	282	3

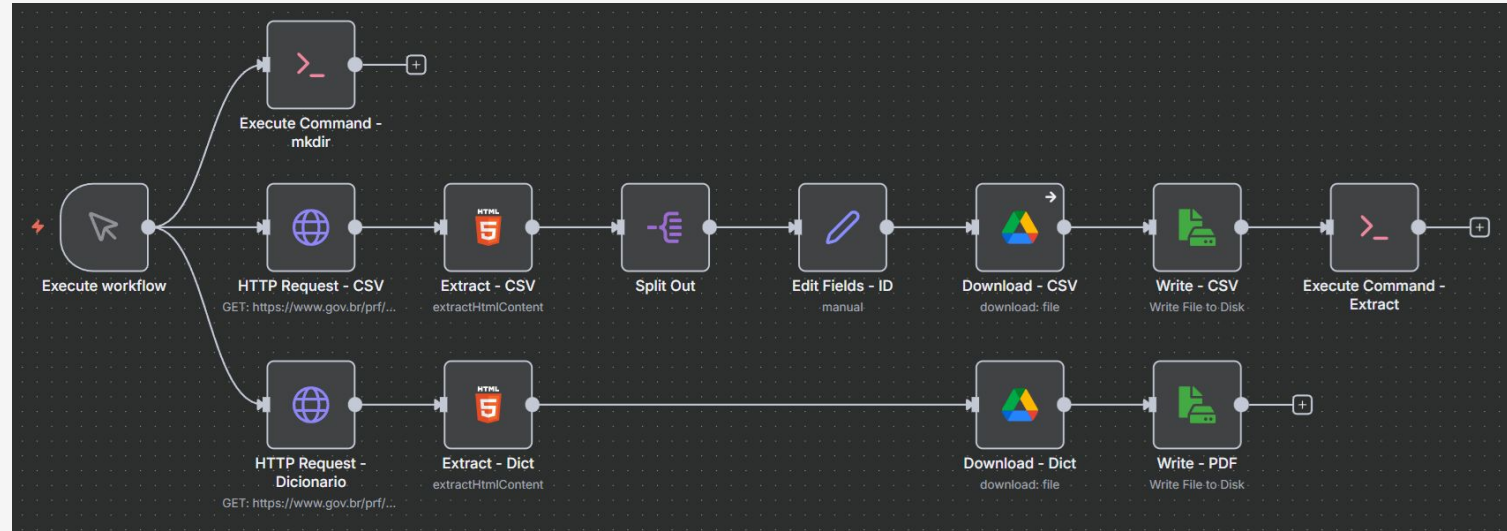
**PRF - Dados Abertos:**

<https://www.gov.br/prf/pt-br/aceso-a-informacao/dados-abertos/dados-abertos-da-prf>



# Arquitetura da Solução (Engenharia)

Extração automática do portal do governo, descompressão e organização em Data Lake



**Docker:**

<https://www.docker.com/>

**n8n:**

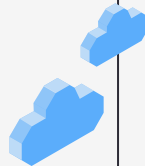
<https://n8n.io/>



# Arquitetura da Solução (Harmonização)

Script para limpar texto, gerar nulos, criar features de engenharia...

```
mapping = {  
  
    # ** SONO  
    'Dormindo': 'Condutor Dormindo',  
  
    # ** ÁLCOOL/SUBSTÂNCIAS  
    'Ingestão De Álcool': 'Ingestão De Álcool/Substâncias',  
    'Ingestão De Álcool Pelo Condutor': 'Ingestão De Álcool/Substâncias',  
    'Ingestão De Substâncias Psicoativas': 'Ingestão De Álcool/Substâncias',  
    'Ingestão De Substâncias Psicoativas Pelo Condutor': 'Ingestão De Álcool/Substâncias',  
  
    'Pedestre - Ingestão De Álcool/ Substâncias Psicoativas': 'Ingestão De Álcool/Substâncias (Pedestre)',  
    'Ingestão De Álcool E/Ou Substâncias Psicoativas Pelo Pedestre': 'Ingestão De Álcool/Substâncias (Pedestre)',  
    'Ingestão De Álcool Ou De Substâncias Psicoativas Pelo Pedestre': 'Ingestão De Álcool/Substâncias (Pedestre)',  
  
    # ** DISTÂNCIA  
    'Não Guardar Distância De Segurança': 'Não Manter Distância De Segurança',  
    'Condutor Deixou De Manter Distância Do Veículo Da Frente': 'Não Manter Distância De Segurança',  
  
    # ** DEFEITO MECÂNICO (para não pulverizar o cluster)  
    'Problema Com O Freio': 'Defeito Mecânico',  
    'Problema Na Suspensão': 'Defeito Mecânico',  
    'Defeito Mecânico No Veículo': 'Defeito Mecânico',  
    'Defeito Mecânico Em Veículo': 'Defeito Mecânico',  
    'Demais Falhas Mecânicas Ou Elétricas': 'Defeito Mecânico',  
    'Avarias E/Ou Desgaste Excessivo No Pneu': 'Defeito Mecânico',  
}
```





# Arquitetura da Solução (Harmonização)

Script para limpar texto, gerar nulos, criar features de engenharia...

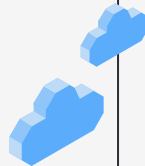
```
weekday_mapping = {  
    'Segunda': 'Segunda-feira',  
    'Terça': 'Terça-feira',  
    'Quarta': 'Quarta-feira',  
    'Quinta': 'Quinta-feira',  
    'Sexta': 'Sexta-feira',  
    'Sábado': 'Sábado',  
    'Domingo': 'Domingo',  
}
```

```
weather_mapping = {  
    'Ignorada': 'Ignorado',  
    'Ceu Claro': 'Céu Claro',  
    'Nevoeiro/neblina': 'Nevoeiro/Neblina',  
}
```

```
mapping = {  
    ** Atropelamento  
    tropelamento De Pessoa': 'Atropelamento (Pessoa/Animal)',  
    tropelamento De Animal': 'Atropelamento (Pessoa/Animal)',  
    tropelamento De Pedestre': 'Atropelamento (Pessoa/Animal)',  
    ** Colisão  
    olisão Com Objeto Fixo': 'Colisão (Objeto Estático)',  
    olisão Com Objeto Estático': 'Colisão (Objeto Estático)',  
    'Colisão Com Bicicleta': 'Colisão (Objeto Móvel)',  
    'Colisão Com Objeto Móvel': 'Colisão (Objeto Móvel)',  
    'Colisão Com Objeto Em Movimento': 'Colisão (Objeto Móvel)',  
    'Colisão Com Objeto': 'Colisão (Objeto)',  
}
```

```
if df_clean['ano_base'].iloc[0] >= 2017:
```

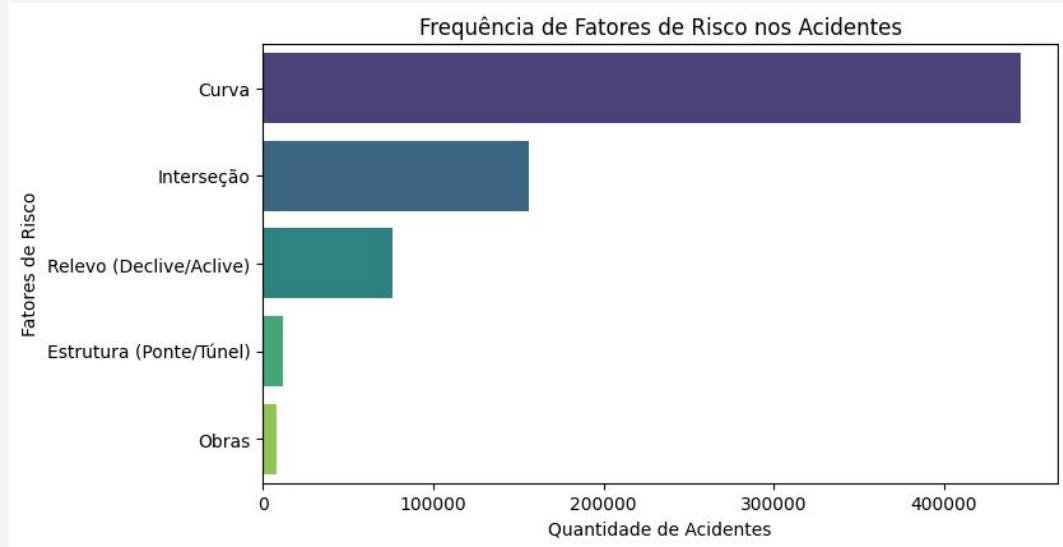
```
    mapping = { 'Sim': 'Urbano', 'sim': 'Urbano', 'Não': 'Rural', 'não': 'Rural' } # Mapping for new schema  
  
    df_clean['uso_solo'] = df_clean['uso_solo'].map( mapping ).fillna( df_clean['uso_solo'] ) # Apply mapping  
    df_clean['uso_solo'] = df_clean['uso_solo'].astype( 'category' ) # Set type to category
```



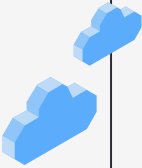


# Primeiras visualizações

Quais as primeiras impressões e conclusões que podemos tirar?

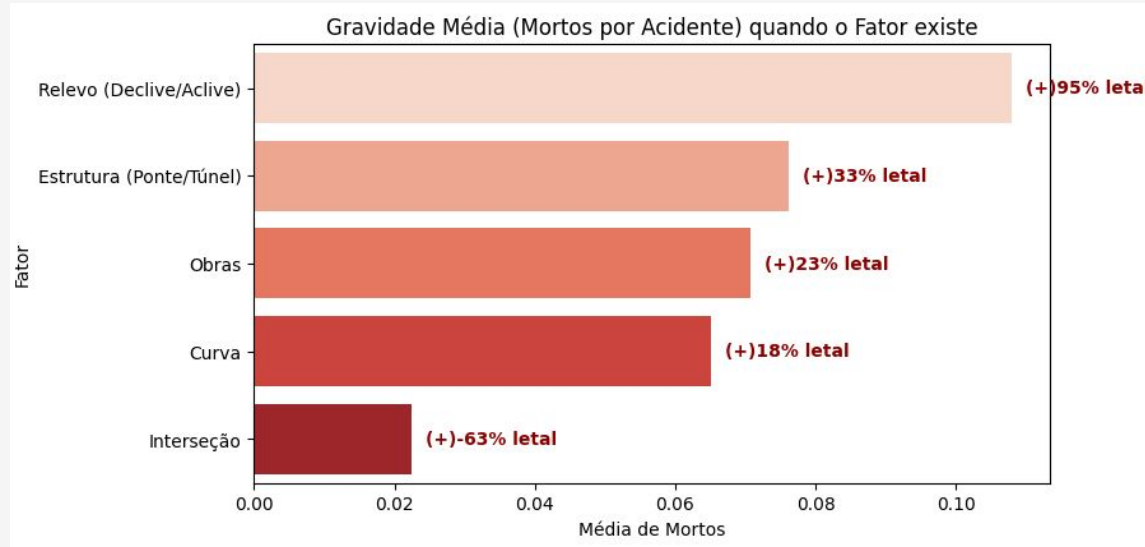


Acidentes totais: 2.122.328

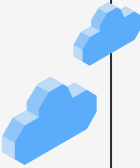


# Primeiras visualizações

Quais as primeiras impressões e conclusões que podemos tirar?

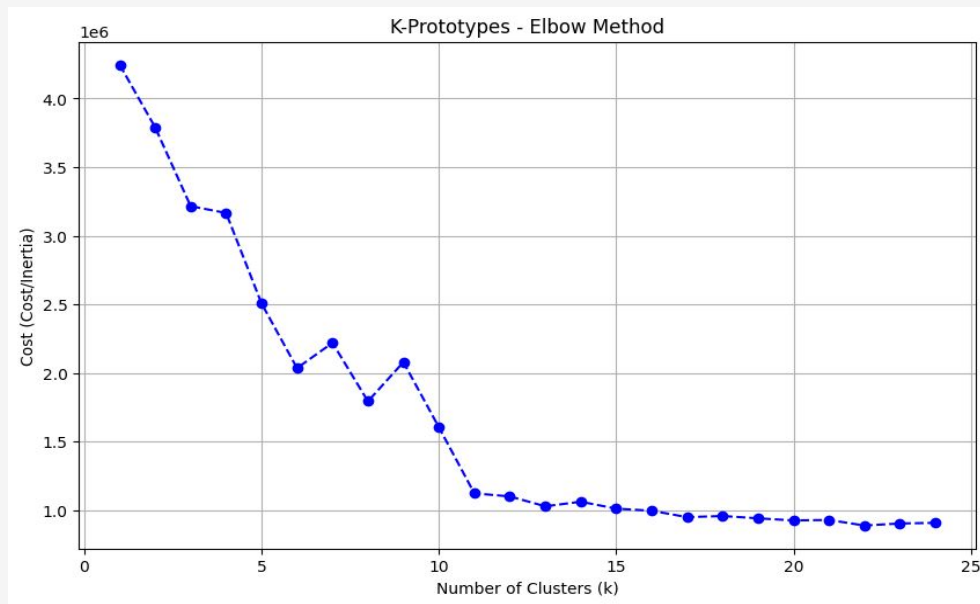


Em locais com relevo acentuado, a média de mortes por acidente é 95% maior do que em locais sem esse fator.

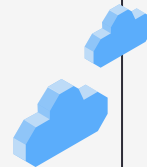


# Cluster **A** - K-Prototypes

Clusterização Eficiente para Dados Numéricos e Categóricos



# Cluster **A** - K-Prototypes



	% total	% morto	% feridos graves	Média veicul	Média pessoa	Curva	Relev	Estruc	Intersec	Obras	Quantida de	Fase dia	Dia semana	Tipo acidente	Causa acidente	Tipo pista	Condiç metere
0	3,40	0,1	0,27	2,03	2,59	0,25	1	0	0,04	0	72.001	Pleno Dia	Domingo	Saída De Leito Carro	Falta de Atenção/ Reação	Simp les	Céu Claro
1	0,90	0,07	0,26	2,04	2,54	0,11	0,16	0,59	0,09	0,43	20.012		Sábado	Colisão Traseira			
2	19,30	0,05	0,16	1,42	1,81	1	0	0	0	0	408.881		Domingo	Saída De Pista	Outras		
3	63,90	0,05	0,17	1,73	2,05	0	0	0	0	0	1.355.147		Sexta-Fei	Colisão Traseira	Falta de Atenção/ Reação		
4	5,40	0,21	0,53	3,65	5,95	0,12	0,01	0	0	0	114.841		Domingo				
5	7,10	0,02	0,17	1,94	2,3	0,01	0	0	1,0	0	151.446		Sexta-Feir	Colisão Transversal			

Planilhas de exploração:

[https://docs.google.com/spreadsheets/d/19wEdod6AtodFinwVzPH\\_JOCJTYuaHdDFw-9aTiHZiow/edit?gid=1685826567#gid=1685826567](https://docs.google.com/spreadsheets/d/19wEdod6AtodFinwVzPH_JOCJTYuaHdDFw-9aTiHZiow/edit?gid=1685826567#gid=1685826567)



# Cluster **B** - DBSCAN

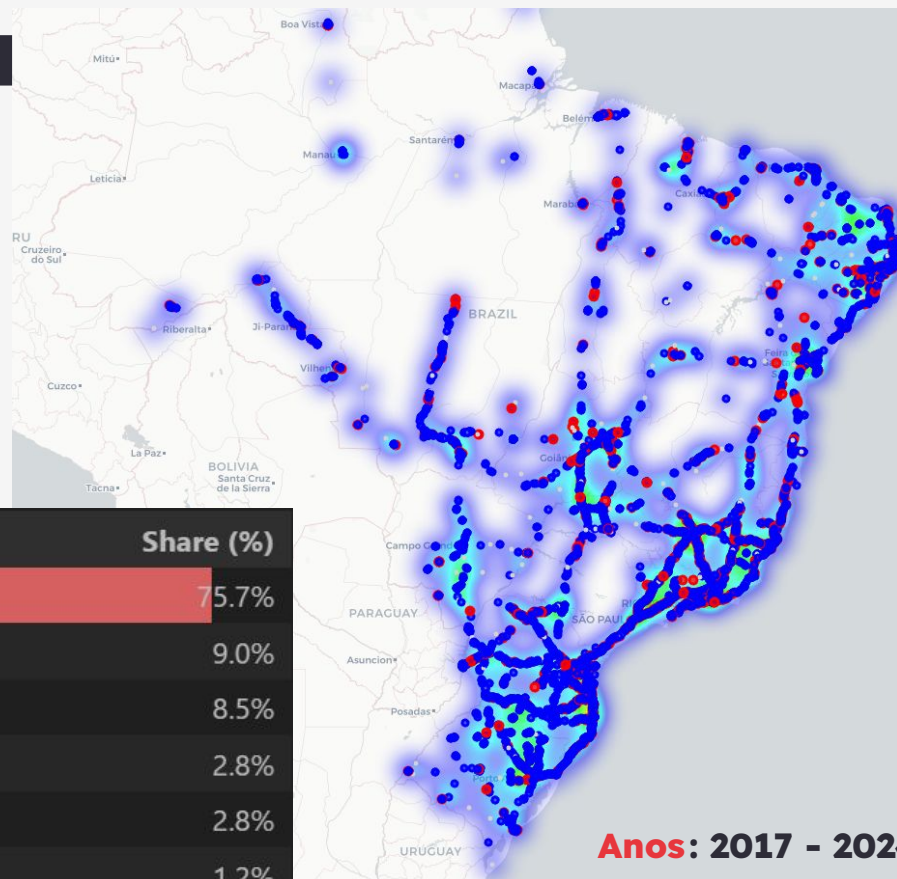
Agrupamento Espacial Baseado em Densidade.

**Micro** - BR-116 (SP), Km 207

Acidentes: Raio de 0,2 Km

Quantidade: 7.241

Mortos: 278 (3.8%)



	Profile	Share (%)
0	Standard Flow / Straight (Cluster A - 3)	75.7%
1	Mass Casualty Event (Cluster A - 4)	9.0%
2	Hazardous Terrain / Relief (Cluster A - 0)	8.5%
3	Curve Tangent (Cluster A - 2)	2.8%
4	Infrastructure Bottleneck (Cluster A - 1)	2.8%
5	Urban Conflict / Crossing (Cluster A - 5)	1.2%



# Cluster **B** - DBSCAN

Agrupamento Espacial Baseado em Densidade.

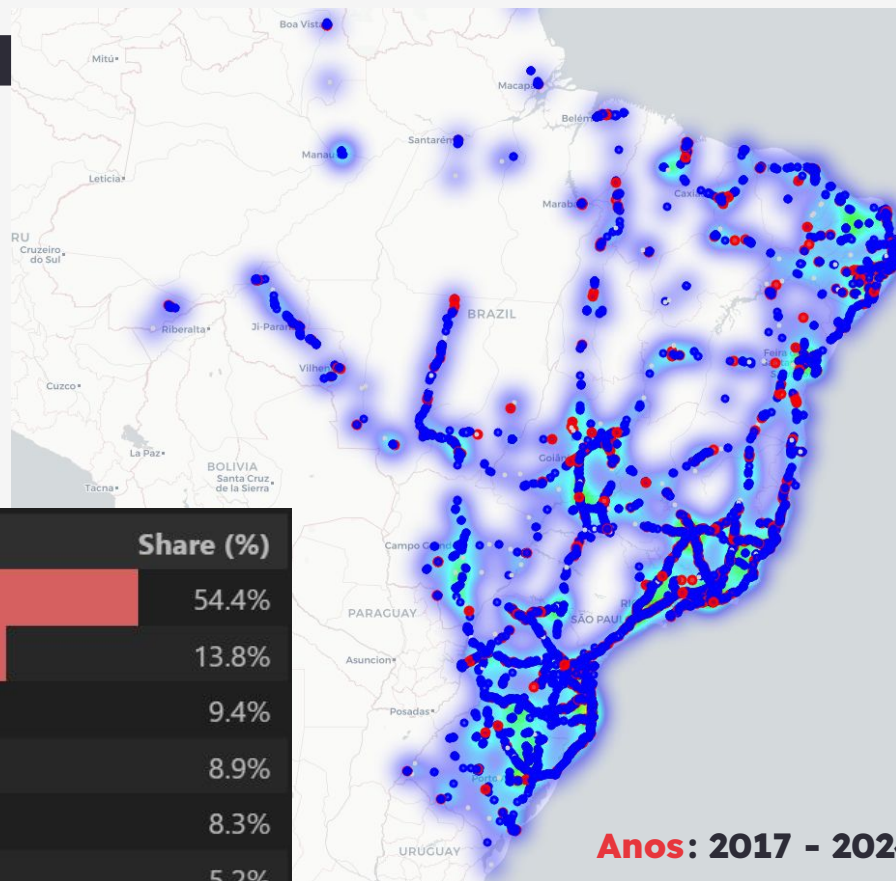
**Macro** - BR-101 (SC), Km

**169**

**Acidentes:** Raio de 5 Km

**Quantidade:** 72.943

**Mortos:** 2.904 (4.0%)



	Profile	Share (%)
0	Standard Flow / Straight (Cluster A - 3)	54.4%
1	Curve Tangent (Cluster A - 2)	13.8%
2	Hazardous Terrain / Relief (Cluster A - 0)	9.4%
3	Urban Conflict / Crossing (Cluster A - 5)	8.9%
4	Mass Casualty Event (Cluster A - 4)	8.3%
5	Infrastructure Bottleneck (Cluster A - 1)	5.2%

**Anos: 2017 - 2024**

