Perfil de pessoas desaparecidas no Estado de São Paulo

Fernando Poliano, Rafael Stern, Julio Trecenti, Eliana Vendramini

2016-03-16

Resumo

Há mais de dois anos o Ministério Público do Estado de São Paulo se dedica ao conhecimento do fenômeno do desaparecimento de pessoas. Através do Programa de Localização e Identificação de Desaparecidos – PLID, demonstrou-se a necessidade de adoção de uma política pública na temática. O programa permitiu a investigação do desaparecimento de vulneráveis e a criação de um banco de dados no nível Estadual. Neste estudo jurimétrico, analisamos um total de 24.261 queixas de desaparecimento realizadas nos anos de 2013 e 2014, de acordo com o sistema Prodesp do Governo de São Paulo. Utilizamos técnicas de mineração de texto para trabalhar com os Boletins de Ocorrência, classificando automaticamente possíveis causas dos desaparecimentos. Como resultados, observamos que os indivídios apresentam diferentes padrões nas idades de desaparecimento em relação ao sexo e causa do desaparecimento. Além disso, observamos que cerca de 70Em pesquisas futuras, exploraremos a dimensão geográfica, realizando cruzamentos de idade, sexo e causa com localidades, a fim de identificar focos de crimes organizados. Os resultados das pesquisas serão aproveitados na definição de estratégias de atuação efetivas no combate ao crime organizado e localização de desaparecidos.

1 Introdução

Há mais de dois anos o Ministério Público do Estado de São Paulo se dedica ao conhecimento do fenômeno do desaparecimento de pessoas. Através do Programa de Localização e Identificação de Desaparecidos – PLID,[1] demonstrou-se a **necessidade de adoção de uma política pública na temática**. O programa permitiu a investigação do desaparecimento de pessoas vulneráveis e a criação de um banco de dados no nível Estadual. Trabalhos que permitam direcionar políticas públicas de segurança e prevenção ainda estão em construção.[2]

O banco de dados do MPSP/PLID foi construído usando os números oficiais advindos da Secretaria de Segurança Pública em parceria com a Delegacia Geral de Polícia. A partir dessa base de dados, o MPSP/PLID partiu para uma análise que permitisse visualizar alguns fenômenos básicos, como, por exemplo, os locais de maior desaparecimento de pessoas, a população alvo e quais suas causas (dentre informadas e eventualmente

[1] O PLID nasceu como um banco de dados, no seio do Ministério Público do Rio de Janeiro, com o desafio de congregar informações sobre pessoas desaparecidas, controlar seu fluxo, fazer buscas guiadas e permitir o olhar global do fenômeno, posto que seu programa oferece, em tempo real, faixa de idade, sexo, naturalidade, nacionalidade, local de ocorrência, localização, circunstâncias da localização, motivação e tipo de identificação de cada fato. Hoje, RJ, SP, BA, AM, PI, CE, PA, PE e Distrito Federal e Territórios estão juntos nessa tarefa, em programa unificado. Esse olhar inédito, especialmente em meio aos sistemas de justiça, **foi premiado com o VII Innovare - Menção Honrosa - 2010**.

[2] Já existe procedimento junto à Comissão Permanente de Direitos Fundamentais do CNMP – GT5 para que o PLID se transforme em SINALID – Sistema Nacional de Localização e Identificação de Desaparecidos, com a atuação dos MPs de todos os Estados, ora reunindo RJ, SP, BA, AM, PI, CE, PA, PE e Distrito Federal e Territórios.

descobertas). Dentre essas causas, destacam-se o tráfico de pessoas (para os mais variados fins), a violência urbana (especialmente policial) e o tráfico de drogas (ocupando os espaços públicos); bem como do trato dos doentes, sejam mentais, sejam por drogadição ou por alcoolismo. Tais causas revelam que os próprios registros de desaparecimento de pessoas nos levam a temas de relevo para elaboração de políticas públicas.

1.1 Informações sobre a pesquisa

Nossa pesquisa partiu de base de dados referente às queixas de desaparecimento de pessoas realizadas nos anos de 2013 e 2014, de acordo com o sistema Prodesp do Governo de São Paulo. A obtenção dos dados foi realizada no mês de março de 2015, considerando somente desaparecidos não encontrados até o momento da extração. O estudo partiu de um universo de 25.682 queixas e, após retirar informações faltantes de sexo e faixa etária, restaram na base 24.261 queixas.

A Tabela 1 mostra a quantidade de queixas envolvendo indivíduos do sexo masculino e feminino[3]. Observamos que os desaparecidos são em sua maioria do sexo masculino.

Sexo	n	%
Masculino	14625	60.3%
Feminino	9636	39.7%
Total	24261	100%

Tabela 1: quantidade e proporção de queixas envolvendo indivíduos de cada sexo.

A Tabela 2 mostra a quantidade de queixas envolvendo indivídos em algumas faixas etárias[4]. É possível observar que existe um padrão diferenciado no desaparecimento de jovens, já que existe uma concentração de 40% das vítimas com idades entre 12 e 20 anos.

Faixa etária	n	%
[0,11]	1023	4.2%
(11,15]	5436	22.4%
(15,20]	4666	19.2%
(20,30]	4699	19.4%
(30,60]	7318	30.2%
(60,80]	983	4.1%
(80,99]	136	0.6%
Total	24261	100%

Tabela 2: quantidade e proporção de queixas em relação à faixa etária, em anos.

A Figura 1 mostra um histograma da idade em anos dos indivíduos envolvidos nas queixas. As cores representam faixas etárias de interesse. Observamos dois grupos, com picos nos 15 e 26 anos, sendo a primeira concentração significativamente mais acentuada.

[3] A informação sobre o sexo não é observada em 196 queixas (0.763%).

[4] A informação sobre o idade não é observada em 1089 queixas (4.24%).

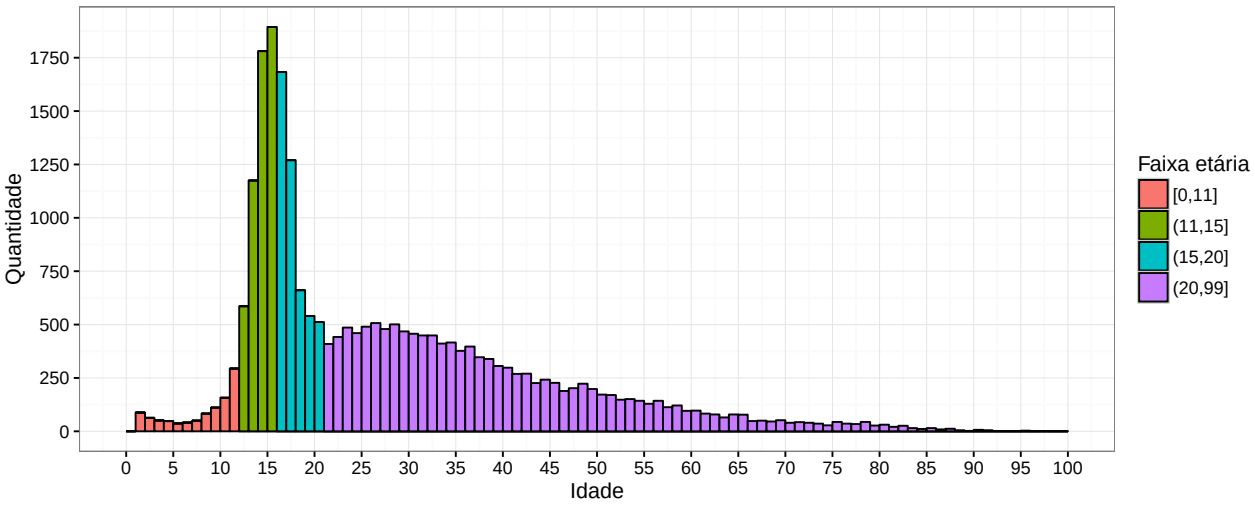

Figura 1: distribuição das idades dos indivídos envolvidos nas queixas, em anos.

1.1.1 Sexo e faixa etária

A Figura 2 mostra o mesmo histograma da Figura 1, mas separado por sexo. Observamos uma diferença na divisão de grupos entre homens e mulheres. Enquanto apenas há um grupo de mulheres, com pico aos 15 anos, os homens desaparecidos dividem-se em 2 grupos, com picos aos 15 e 26 anos. Na próxima seção tentamos caracterizar os principais motivos para o desaparecimento em cada um dos grupos.

2 Análise dos boletins de ocorrência

A base de dados contém, além de informações sobre idade e sexo, os relatos dos casos, na forma de Boletins de Ocorrência (BOs). Os BOs, quando bem documentados, trazem evidências de quais seriam as causas do desaparecimento de pessoas como, por exemplo, drogadição, uso de bebidas alcoólicas ou outras doenças, como mal de Alzheimer e esquizofrenia.

A análise das possíveis causas dos desaparecimentos de pessoas é um objeto de pesquisa de interesse. Tais informações podem trazer mais ingredientes para a compreensão do perfil dos desaparecidos e, com investigações mais aprofundadas, é possível identificar fenômenos para elaboração de ações estratégicas. Por exemplo, ao cruzar causas e localizações das pessoas desaparecidas, podemos encontrar focos de tráfico de drogas ou tráfico de pessoas.

Contudo, os BOs precisam ser adequadamente trabalhados antes de serem analisados. Como os textos foram escritos em linguagem natural, uma mesma informação (e.g. a vítima era usuária de drogas) pode aparecer de diversas formas. Para resolver esse problema, utilizamos técnicas de mineração de texto[5].

[5]Manning, Christopher D., and Hinrich Schütze. Foundations of statistical natural language processing. Vol. 999. Cambridge: MIT press, 1999.

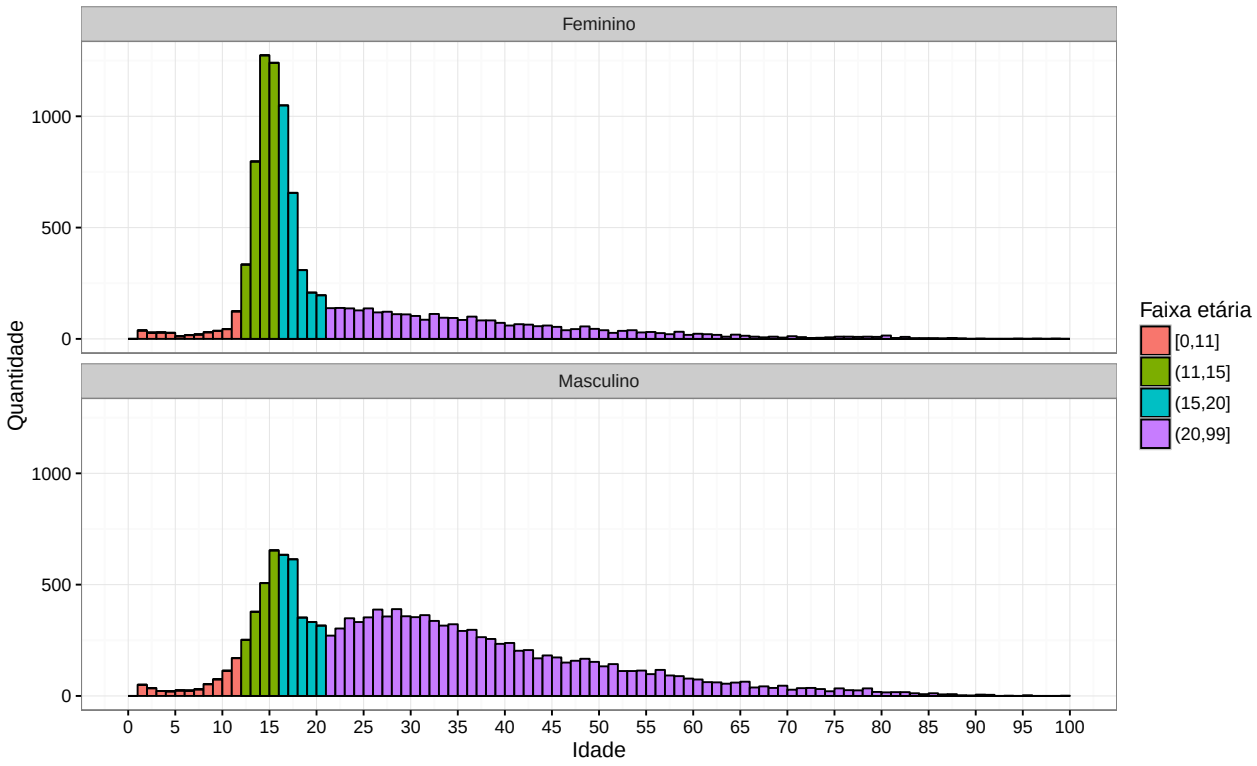

Figura 2: distribuição das idades dos indivídos envolvidos nas queixas, em anos, desagregado por sexo.

2.1 Classificação dos boletins de ocorrência

A classificação das possíveis causas a partir dos BOs foi realizada em três etapas.

No primeiro passo, fizemos uma amostra piloto de 30 casos e analisamos os textos. Com a leitura, buscamos identificar quais seriam as possíveis causas dos desaparecimentos de pessoas. O resultado desse estudo foi uma lista de causas relevantes: i) drogadição/alcoolismo; ii) doença mental/depressão; iii) desaparecimento voluntário; iv) desavenças familiares; v) desaparecimento sem causa aparente; vi) ausência de BO; vii) BO não relacionado a desaparecimentos; viii) provável vítima de crime; e ix) retorno da vítima.

No segundo passo, montamos uma amostra com 201 casos e os classificamos de acordo com as causas supracitadas. Como resultado, temos o que chamamos de "base de treino", uma base de dados utilizada para a construção de um modelo estatístico que associa automaticamente os textos dos BOs às classificações. Dada a discordância entre os autores em relação à identificação de desaparecimentos voluntários, desavenças familiares e prováveis vítimas de crimes, esses casos foram classificados como desaparecimentos sem causa aparente. Também retiramos os casos sem BO ou com erro e BOs de retorno.[6] No final, ficamos somente com as categorias (i), (ii) e (v).

No terceiro passo, propomos diferentes modelos e testamos seus poderes preditivos utilizando validação cruzada.[7] O modelo com melhor desempenho utilizou, além dos textos, um dicionário construído manualmente

[6] Os casos sem BO, com erro e retorno foram retirados tanto da base de treino quanto da base completa utilizando palavras-chave aplicadas aos textos.

[7] Validação cruzada é uma técnica que consiste em treinar um modelo preditivo em um subconjunto da base de dados, para então

contendo algumas expressões regulares associadas à drogas e doenças.[8] Outro modelo com desempenho equivalente foi definido manualmente e considera um dicionário de palavras e uma série de regras lógicas usadas para decidir entre as possíveis causas.

O modelo final escolhido resultou numa taxa acerto de aproximadamente 84% na validação cruzada. A Figura 3 mostra o funcionamento do modelo a partir de uma árvore binária de classificação. Cada variável indica a existência de uma ou mais palavras específicas no texto do BO. Descendo pela árvore e seguindo a direção apropriada para cada variável, é possível obter uma classificação da causa. Por exemplo, se temos palavras relacionadas a drogas, não temos palavras relacionadas a doenças, não aparece acolhimento, não temos desaparecimento sem motivo, não evadiu, não aparece a expressão "não faz uso", concluímos que a causa é Drogas.

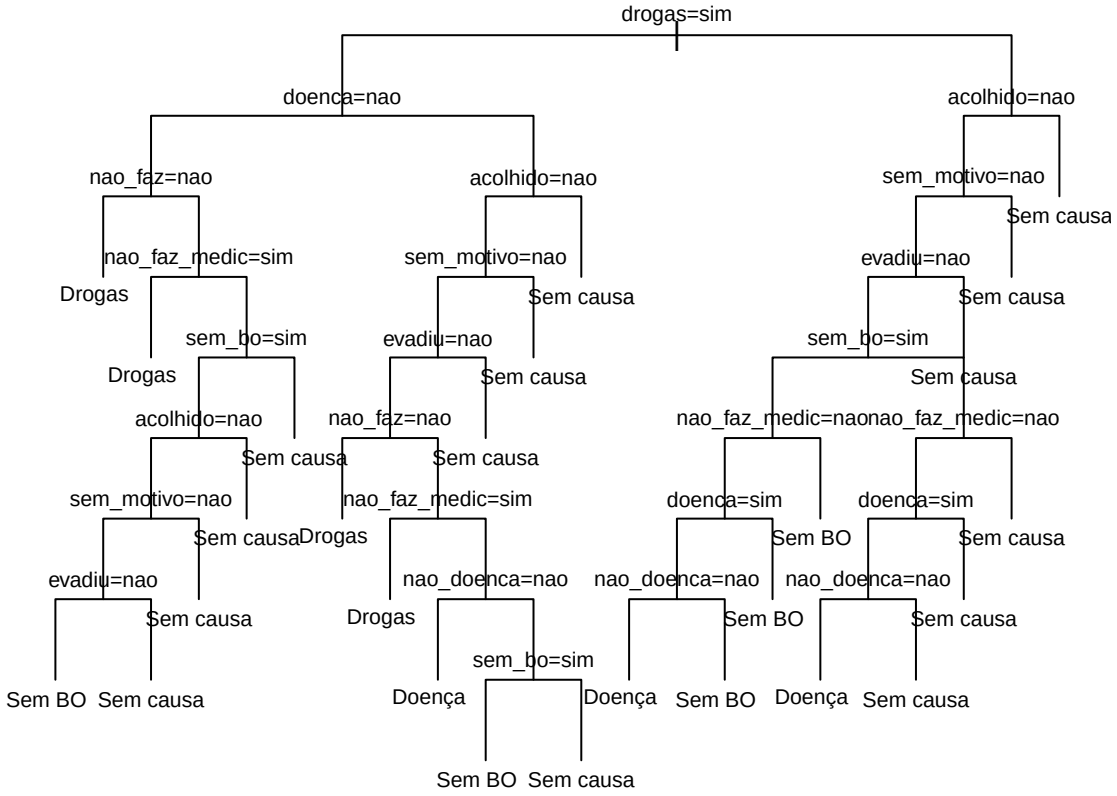

Figura 3: regras lógicas aplicadas pelo modelo.

O modelom construído com base na leitura de 201 casos foi utilizado para gerar predições a partir dos textos dos boletins de ocorrência. A classificação final é descrita na Tabela 3. Note que após a retirada dos casos sem BO restaram 22.888 casos. É possível observar que grande parte dos casos continuam sem causa aparente e que aproximadamente um em cada cinco casos de desaparecimento de pessoas tem relação com

testá-lo no subconjunto complementar. Ver Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.

[8] O modelo utilizado denomina-se Naive Bayes. Nesse modelo, cada palavra do dicionário recebeu um *score* definido empiricamente, que pode ser interpretado como um peso para a associação entre a presença da palavra no texto e uma causa de desaparecimento específica.

drogadição ou alcoolismo. A proporção de casos relacionados a outras doenças é menor do que um em vinte casos, mas ainda é relevante.

Causa	n	%
Sem causa	17031	74.4%
Drogas	4924	21.5%
Doença	933	4.1%
Total	22888	100%

Tabela 3: frequências das causas de desaparecimento de pessoas obtidas pelo modelo.

2.2 Relação entre causa de desaparecimento e perfil da vítima

A Figura 4 mostra a distribuição das causas em relação ao sexo. É possível observar que os homens apresentam maior proporção de casos relacionados a drogadição/alcoolismo que as mulheres.

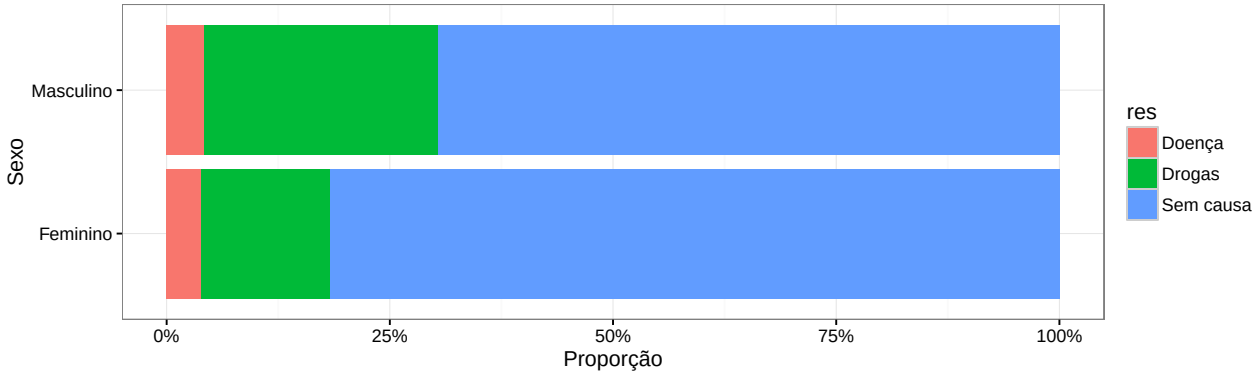

Figura 4: distribuição das causas por sexo.

A Figura 5 mostra histogramas das idades das vítimas em relação ao sexo e causa. A partir dos gráficos é possível observar vários fenômenos:

- Problemas relacionados a doenças aparecem após os dez anos de idade. Não observamos, no entanto, uma concentração de casos entre onze e vinte anos de idade nos homens. A concentração de casos na adolescência é relevante para mulheres, mas consideravelmente menor do que nos casos sem causa aparente;
- Mulheres apresentam uma concentração maior entre onze e vinte anos de idade quando a causa é drogas ou quando a causa não é identificada. Note que a proporção de desaparecidas acima de vinte anos com causa não identificada é menor que a mesma proporção quando a causa é relacionada a drogas.
- Em relação aos homens, a distribuição das idades das vítimas se concentra na faixa entre vinte e quarenta anos quando a causa é drogadição ou alcoolismo, diferente do que acontece na mesma faixa etária para vítimas sem causa de desaparecimento aparente.

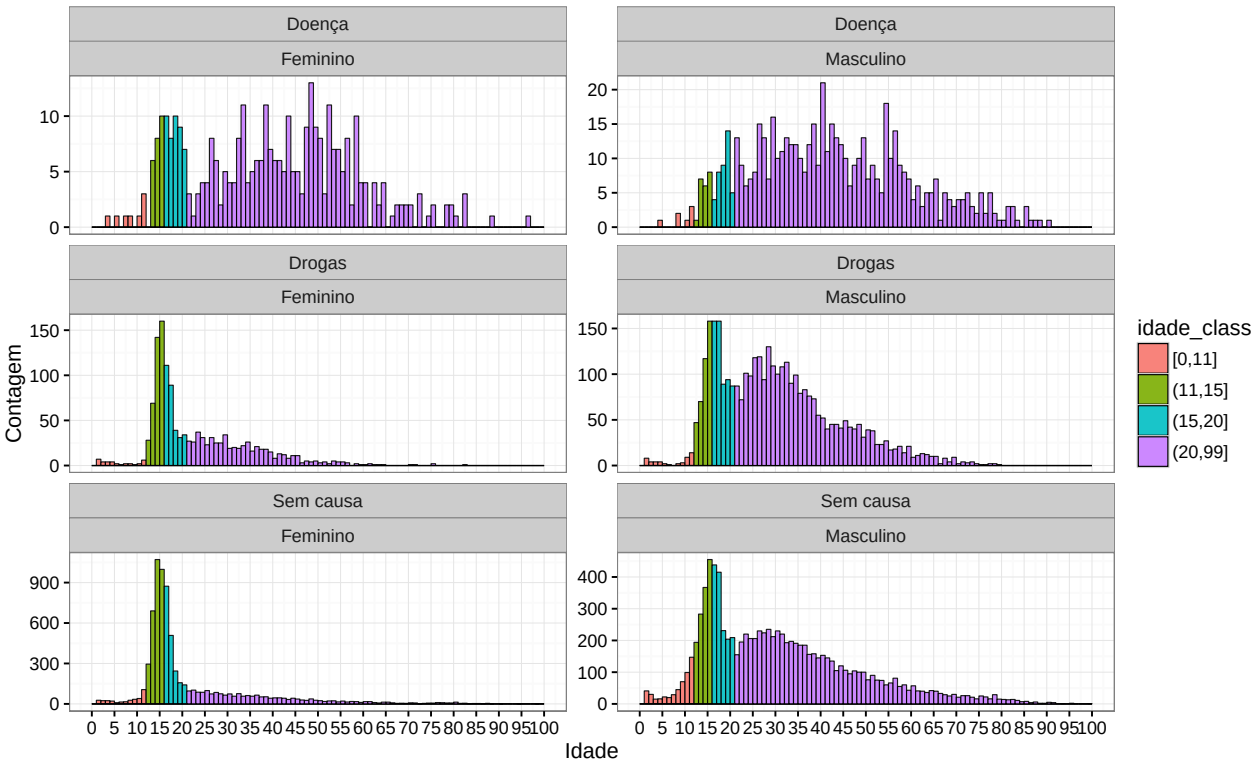

Figura 5: distribuição das idades dos indivídos envolvidos nas queixas, em anos, desagregado por sexo e causa.

2.3 Conclusões

Neste trabalho realizamos um estudo de caráter exploratório sobre uma base de dados concebida para fins gerenciais. Mesmo com ressalvas acerca da qualidade dos dados e a falta de padrão nos boletins de ocorrência, foi possível identificar padrões sobre sexo, idade e causas dos desaparecimentos de pessoas no estado de São Paulo. Dentre eles:

- 46% dos casos de desaparecimento de pessoas acontecem até a adolescência e, quando a causa é drogadição ou alcoolismo, essa proporção cai para 36%.
- Casos que envolvem doença apresentam menor proporção de vítimas adolescentes.
- Há diferenças entre a distribuição de idades de desaparecimento de homens e mulheres, a primeira com dois grupos e a segunda com apenas um grupo.
- 70% dos casos de desaparecimento sem causa aparente de mulheres acontecem até a adolescência.

O último ponto desperta atenção por acreditarmos que embora existam diversas explicações para o desaparecimento de mulheres na adolescência, nossa análise foi capaz de isolar parcialmente o rol de possíveis causas. Em particular, chamamos a atenção para crimes relacionados a tráfico de pessoas ou sequestros em geral.

Em futuras análises, a ABJ, juntamente com o MPSP e o PLID utilizará bases de dados atualizadas até 2015. Além da ampliação da abrangência temporal, iremos aprofundar mais a questão das causas, aproveitando a inteligência dos especialistas para produzir modelos capazes de identificar com maior acurácia as possíveis

explicações para os desaparecimentos de pessoas.

Além disso, exploraremos a dimensão geográfica, realizando cruzamentos de idade, sexo e causa com localidades, a fim de identificar focos de crimes organizados. Os resultados da pesquisa serão aproveitados na definição de estratégias de atuação efetivas no combate ao crime organizado e localização de desaparecidos.

Nessa pesquisa mostramos como a exploração cuidadosa de bases de dados acumuladas é útil na elucidação de fenômenos sociais complexos. Mais do que isso, mostramos que a jurimetria é indispensável quando o objetivo é definir políticas públicas de qualidade.