TPBCI | Tendências da Pesquisa Brasileira em Ciência da Informação

FORA DE SÉRIE: uso de outliers na predição criminal

OUTSTANDIG: use of outliers in criminal prediction

Manoel Camilo de Sousa Netto [1]
Adilson Luiz Pinto [2]

Resumo: O trabalho ora apresentado propõe um modelo de predição criminal baseado na Ciência da Informação, Estatística, Ciências Policiais e Ciências da Computação. Inicialmente o estudo destaca o papel da interdisciplinaridade na construção de modelos reutilizáveis que guiem atuações preventivas futuras. Apresenta o problema da pesquisa como sendo o impedimento da materialização delitiva por meio de um modelo que possa fortalecer a prevenção e, como objetivo principal propor um modelo interdisciplinar de predição criminal que aponte a um estágio de conhecimento dos cenários de maior risco de ocorrência criminal. Aponta, como fundamentação, conceitos acerca de predição criminal, outliers (discrepantes), medidas de centralidade, medidas de dispersão. Apresenta o modelo proposto na forma de um fluxograma de tarefas bem definidas, cada uma delas permeada pelas ciências que contribuíram em sua concepção. Explica que esse fluxograma pode ser aplicado para obter a predição de vários tipos de crimes. Explana cada tarefa do modelo: seleção do tipo de crime e variáveis, obtenção e tratamento dos dados, análise exploratória, cálculo de métricas, predição de amostra a ser monitorada e entrega da amostra às equipes de policiamento preventivo. Conclui que é possível propor um padrão de predição criminal de natureza interdisciplinar, apontando os outliers com acusações pré-existentes como o conjunto que compõe a predição que a aplicação da pesquisa buscou.

Palavras-Chave: Outliers. Discrepantes. Crime. Predição.

Abstract: *The paper presented here proposes a criminal prediction model based on Information Science, Statistics, Police Sciences and Computer Science. Initially the study highlights the role of interdisciplinarity in the construction of reusable models that guide future preventive actions. It presents the research problem as the prevention of criminal materialization through a model that can strengthen prevention and, as its main objective, to propose an interdisciplinary model of criminal prediction that points to a stage of knowledge of scenarios of higher risk of criminal occurrence. It points out, as grounds, concepts about criminal prediction, outliers, measures of centrality, measures of dispersion. It presents the proposed model in the form of a well-defined task flowchart, each one permeated by the*

[1] Doutorando em Ciência da Informação na Universidade Federal de Santa Catarina, UFSC, Brasil. Tutor credenciado pela Secretaria Nacional de Segurança Pública.

[2] Doutor em Documentação pela Universidade Carlos III de Madrid (2007). Coordenador do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Santa Catarina - UFSC (gestão 2017-2019).

sciences that contributed to its conception. Explains that this flowchart can be applied to predict various types of crime. Explains each task of the model: selection of crime type and variables, data collection and processing, exploratory analysis, metric calculation, sample prediction to be monitored, and sample delivery to preventive policing teams. It concludes that it is possible to propose an interdisciplinary criminal prediction pattern, pointing outliers with pre-existing accusations as the set that makes up the prediction that the research application sought.

Keywords: *Outliers Discrepant. Crime. Prediction.*

1 INTRODUÇÃO

A Ciência da informação (CI), com suas características multidisciplinares, tem estabelecido fronteiras com diversas ciências. Fronteiras que operam tal qual membranas semipermeáveis. O cenário resultante permite troca de ferramentas, dados e métodos entre as ciências envolvidas, tal qual uma osmose invertida[3]. O conhecimento da ciência mais densa atravessa a membrana-fronteira da ciência periférica vizinha, contribuindo ao aumento da menor densidade cognitiva; ulteriormente, o equilíbrio se reestabelece – mas não definitivamente.

Os cientistas, dentro ou fora da CI, atuam para mudar o quadro de equilíbrio quando, sistematicamente, atingem novos degraus de conhecimento nos diversos campos das suas ciências. Novamente a densidade do conhecimento científico vizinho se altera e, naturalmente, o ciclo se reinicia. A habitual recorrência gera vantagens científicas óbvias: recursos de uma ciência reforçam outra.

A Ciência Policial não pode desprezar tal potencial. E, na sua missão de gerar ciência anticrime, pode associar-se com a CI, a Estatística e a Ciência da Computação para realizar suas osmoses cognitivas inversas com vistas a construir modelos que guiem atuações policiais futuras.

Nenhuma ciência atuando *per si*, atingiria a sinergia do conjunto delas trocando conhecimentos. O trabalho ora abordado fundamenta-se nesse princípio pois, em seu cerne, há o uso da interdisciplinaridade para obtenção de um modelo básico de predição de crimes que usufrui de dados como matéria prima.

A Polícia dispõe de muitos e diversificados dados, mas a mera posse desses elementos – estacionários em uma mídia de armazenamento computacional – não garante um estado de cognição minimamente proveitoso. Os padrões criminais camuflam-se em um amontoado de bits infrutíferos e os indicativos dos crimes futuros, mesmo de existência prévia à consumação, continuam velados. É possível realizar previsão criminal se conhecermos esses indicativos? Quiçá a Ciência Policial e as ciências preditas, atuando em parceria osmótica invertida, transformem o estado anômalo de conhecimento em um estado que preveja os crimes futuros. Certamente não preverá a totalidade deles, mas, ao menos, uma parcela significativa dessas mazelas.

[3] Na osmose real o fluxo de solvente atravessa a membrana semipermeável no sentido (meio menos denso)→(meio mais denso), contrário da metáfora ora empregada.

A fuga do estado anômalo do conhecimento é questão de estudo desde sempre, mas foi formalizada por Brookes (1980, p. 131) quando elaborou a expressão denominada "Equação Fundamental da Ciência da Informação" cujo formato é ilustrado por meio da seguinte expressão algébrica:

$$\boldsymbol{K(S) + \Delta I = K(S + \Delta S)}$$

A Equação exibe um estado geral do conhecimento anômalo denominado ***K(S)***, o qual é modificado para um novo estado do conhecimento ***K(S+ΔS)*** em decorrência do incremento da informação ***ΔI***. No caso da predição de crimes, o estado final ***K(S+ΔS)*** é aquele em que, baseado em métodos científicos, revelam-se cenários onde a probabilidade de crimes é maior. Noutras palavras, é o ato de tomar conhecimento da amostra resultante da própria predição. Apesar de ser uma área nova das Ciências Policiais, a predição tem gerado interesse em polícias de todo o mundo.

O problema abordado na pesquisa é o impedimento da materialização delitiva por meio de um modelo que possa fortalecer a prevenção.

Seguindo os fundamentos da Equação de Brookes, o objetivo do presente artigo é propor um modelo interdisciplinar de predição a que aponte a um estágio de conhecimento dos cenários de maior risco de ocorrência criminal.

A pesquisa também objetiva aplicar o modelo em um cenário real: a predição de crimes de desvio de recursos públicos.

2 DESENVOLVIMENTO

Segundo o Programa das Nações Unidas para o Desenvolvimento (PNUD), em 2015, 3,8% da população brasileira - aproximadamente 7,7 milhões de pessoas - encontrava-se em condições de pobreza multidimensional. O termo multidimensional refere-se ao fato de que houve uma quebra de paradigmas quando o PNUD adotou outras variáveis além da baixa renda como influenciadoras da medição do grau de pobreza. Dentre as novas variáveis estão a saúde, a educação e a segurança.

Ao considerar que esses fatores devem ser priorizados pelo Estado como promotor primeiro do bem-estar social, releva-se a importância da escrupulosa aplicação dos recursos

públicos, especialmente em cenários tais quais o da crise econômica ora vigente em nosso país. Entretanto, apesar desse cenário, a atuação policial continua se pautando na repressão.

Primeiro os prejuízos ocorrem, depois a ação estatal surge, quase sempre tarde demais. A reatividade é prenúncio de danos permanentes, pois o crime consumado pode realizar perdas irreparáveis em bens e direitos. Nesse caso, quando muito, espera-se a punição penal contra o criminoso.

Isto posto, faz-se necessário a adoção de estratégias de prevenção ao crime. Uma delas é a predição criminal. Baseada no que já se sabe historicamente, a predição pode utilizar técnicas estatísticas que atuem antes que o crime ocorra. O cenário deixa de ser reativo e posterior para tornar-se proativo e prévio.

A predição também promove a economicidade, ao evitar que o erário público seja poupado do custo da repressão estatal posterior ao crime e dos prejuízos decorrentes da materialização do crime evitável.

2.1 Fundamentação teórica

A fundamentação teórica da pesquisa se utiliza conceitos como predição criminal, outliers, medidas de centralidade e de dispersão.

2.1.1 Predição criminal

As atividades policiais repressivas são aquelas que se manifestam apenas após a ocorrência de um crime específico, tal como um homicídio: ato contínuo ao delito, a ação estatal busca esclarecer as circunstâncias e individualizar seus autores. As atividades preventivas, de outro modo, operam antes que os crimes ocorram, pois são prévias à materialização delitiva.

Embora a atuação da polícia judiciária seja, via de regra, repressiva, a prevenção tem sido gradativamente valorizada na medida em que é capaz de minimizar, por meio do agir proativo, a atuação reativa e póstera. É uma tentativa da mitigação dos esforços e prejuízos provavelmente porvindouros.

O policiamento preventivo pode usar técnicas de predição criminal. Segundo Tayebi e Glässer (2016, p.7) O policiamento preditivo baseia-se na ideia de que, embora alguns crimes

sejam aleatórios, a maioria deles não é. A predição, entretanto, não obterá a exatidão em apontá-los, mas indicará os cenários de maior risco de que eles ocorram.

Sabendo que não se pode monitorar todos os cenários sociais prévios aos crimes, ao menos é possível encontrar as observações estatísticas que se destacam e que podem representar maior risco delitivo. Por meio de variáveis categóricas e quantitativas, essas observações são segregadas das demais, e passam a ser nominadas como outliers ou discrepantes.

2.1.2 Outliers ou discrepantes

Segundo Rousseeuw e Van Zomeren (1990, p. 633) outliers são observações que não seguem o padrão da maioria dos dados. Segundo Gladwell (2008, p. 13) um outlier é uma observação estatística cujo valor na amostragem é marcadamente diferente dos demais.

Segundo Bruce e Bruce (2016, p. 22), em contraste com a análise de dados típica, onde os outliers são às vezes informativos e incômodos, na detecção de anomalias os pontos de interesse são os próprios outliers, e a maior massa de dados serve principalmente para definir o "normal" contra o qual as anomalias são confrontadas.

Os outliers podem ser dados a serem desprezados na pesquisa por possuírem potencial de gerar erros estatísticos, entretanto eles também podem ser os dados que se busca encontrar.

A discrepância, portanto, pode ser uma característica que gera desprezo ou relevância, a depender do que almeja o pesquisador. Hair Junior et al. (2009) apresentam o seguinte posicionamento sobre a manutenção ou eliminação dos discrepantes, os quais chama de observações atípicas:

> Depois que as observações atípicas foram identificadas, descritas e classificadas, o pesquisador deve decidir sobre a retenção ou eliminação de cada uma. Entre os pesquisadores há muitas filosofias sobre como lidar com as observações atípicas. Nossa visão é de que elas devem ser mantidas, a menos que exista prova demonstrável de que estão verdadeiramente fora do normal e que não são representativas de quaisquer observações na população. No entanto, se elas representam um elemento ou segmento da população, devem ser mantidas para garantir generalidade à população como um todo.

2.1.3 Medidas de centralidade, quartis e outliers

A média é uma medida de cômputo simples e a mais conhecida dentre as medidas de centralidade. Seu cálculo se resume a razão entre a soma das observações pela sua quantidade, conforme a seguir:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

A mediana é a realização que ocupa a posição central da série de observações quando estão elas ordenadas em ascendência.

A mediana é uma medida mais resistente (ou robusta), no sentido que que ela não é muito afetada pelos valores discrepantes. O mesmo não ocorre com a média, que sofre distorções consideráveis causadas por eventuais outliers.

Segundo Pinheiro et al. (2012, p. 247) A mediana pode ser descrita matematicamente como:

$$md(X) = \begin{cases} X_{\left(\frac{n+1}{2}\right)} \, se \; n \text{ é } \acute{i}mpar, \\ \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} se \; n \text{ é } par \end{cases}$$

Devore (2006, p. 28) relata que os quartis dividem o conjunto de dados em quatro partes iguais e que aproximadamente 1/4 dos dados recai sobre ou abaixo do primeiro quartil q_1, metade dos dados sobre ou abaixo do segundo quartil q_2 (a mediana) e aproximadamente 3/4 dos dados sobre ou abaixo do terceiro quartil q_3.

Larson e Farber (2016, p. 81) afirmam que a distância (ou amplitude) interquartil (d_q) de um conjunto de dados é uma medida de variação que fornece a amplitude da porção central (aproximadamente metade) dos dados. A d_q, portanto, é a diferença entre o terceiro e o primeiro quartis, conforme abaixo:

$$d_q = q_3 - q_1$$

Segundo Bussab e Morettin (2010, p. 103), outliers são valores maiores do que o limite superior (***Ls***) e menores que o limite inferior (***Li***) os quais, por sua vez, são determinados respectivamente por:

$$\begin{cases} Ls = q_3 + 1{,}5d_q \\ Li = q_1 - 1{,}5d_q \end{cases}$$

2.1.4 Medidas de Dispersão

A interpretação de um conjunto de observações estatísticas por uma única medida representativa de posição central pode ser distorcida porque não considera toda a informação sobre a variabilidade do conjunto.

Outros autores destacam os perigos do uso isolado das medidas de centralidade, vide o relato a seguir:

> Informar apenas a medida de tendência central fornece apenas informações parciais sobre um conjunto de dados ou uma distribuição. Diferentes amostras ou populações podem ter medidas de tendência central idênticas e apresentar diferenças entre si em outros aspectos importantes (Devore, 2006, p. 31).

Essa dificuldade pode ser superada pelo uso das medidas de dispersão. Segundo Pinheiro (2012, p. 37) uma medida de dispersão para uma variável quantitativa é um indicador do grau de espalhamento dos valores da amostra em torno da medida de centralidade. Maiores dispersões indicam menos representatividade dos valores centrais.

Uma das medidas de dispersão, a variância populacional de um conjunto de dados com ***n*** elementos, segundo Larson e Farber (2016, p. 81) pode ser calculada da seguinte forma:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Uma das desvantagens da variância como medida de dispersão é que, por ser elevada ao quadrado, sua unidade é diversa daquela usada pelo conjunto de dados.

A solução é o uso do desvio padrão, outra medida de dispersão. Segundo Haslwanter (2016, p.93) o desvio padrão é calculado pela raiz quadrada da variância.

De uma forma mais detalhada, o desvio padrão pode ser obtido por:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Um desvio-padrão perto de zero aponta que os pontos tendem a estar próximos da média e, quanto maior o desvio padrão, mais longe da média eles tendem a estar.

O desvio padrão analisado por si, entretanto, não carrega consigo ideia da variação em relação à média dos dados. Segundo Pinheiro (2009, p. 32) a interpretação da magnitude do desvio padrão deve ser feita pelo coeficiente de variação, uma grandeza adimensional.

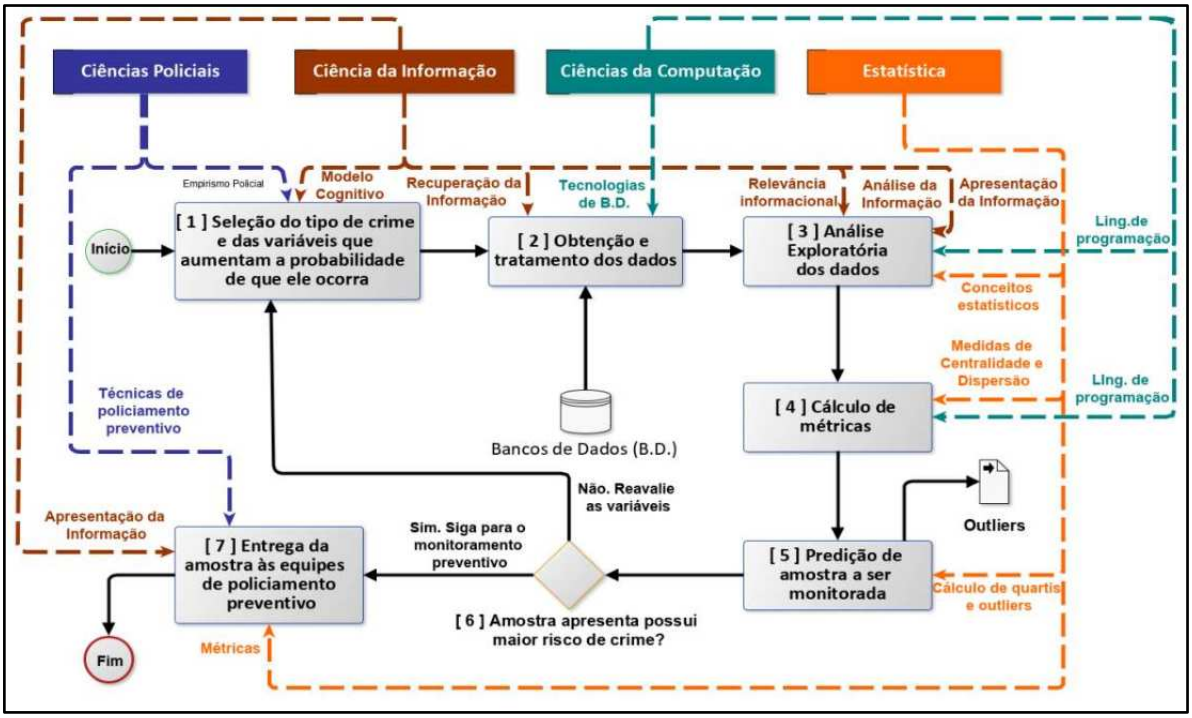
Uma solução para esse óbice é a utilização do coeficiente de variação, definidor por Larson e Farber (2016, p. 81) como sendo a razão entre o desvio padrão e a média dos dados, conforme abaixo:

$$Cv = \frac{\sigma}{\bar{x}}$$

3 MÉTODO

De natureza quali-quantitativa, o tipo de pesquisa ora conduzido é descritiva-exploratória. O método da pesquisa segue um processo: uma sequência de tarefas, onde cada uma delas é uma ação executada por uma pessoa ou por algum sistema. Em cada uma delas há algum nível de interdisciplinaridade característica das pesquisas em Ciência da Informação, conforme a Figura 1:

Figura 1 – Fluxo de tarefas da metodologia proposta



Fonte: Elaborado pelos autores

O modelo ilustrado pode ser adequado a qualquer tipo de crime. Mais adiante aplicar-se-á o fluxograma aos crimes relacionados ao desvio de recursos públicos, entretanto também poderia ser utilizado, por exemplo, em crimes de tráfico internacional de drogas. Normalmente o perfil das pessoas que executam o transporte desses entorpecentes é incomum: passagens compradas às vésperas das viagens, rotas repetidas várias vezes, preço das passagens acima da capacidade financeira presumida do passageiro, viagens solitárias (quase nunca acompanhando amigos ou família), etc.

Um estudo detalhado dos casos de tráfico certamente será capaz de determinar variáveis categóricas e quantitativas, expressas por dados armazenados em sistemas governamentais, que expressam o perfil do traficante com boa margem de acerto. Isso indica que o modelo, baseado em variáveis que aumentam os riscos de crime, pode ser adequado ao crime de tráfico de drogas, bem como a outros tantos tipos penais existentes.

É necessário, entretanto, que para cada tipo de crime as variáveis sejam novamente determinadas.

Cada uma das tarefas da Figura 1 é explanada, na ordem em que aparecem no processo proposto.

3.1 Seleção do tipo de crime e de variáveis

O Brasil tem sido alvo de verdadeira pilhagem aos cofres públicos em todas as esferas de poder: o Ministério Público Federal estima em 200 bilhões em prejuízos causados por esse tipo de crime aos cofres estatais brasileiros anualmente[4].

Parte considerável desses delitos são operacionalizados através das contratações estatais de particulares agindo sob manto de legalidade teatral, todavia acompanhada da obscura utilização de artifícios criminosos.

Pelas razões expostas, os crimes de desvio de recursos públicos foram escolhidos para testar o modelo, mas quais seriam algumas das variáveis que concorrem para que esse tipo de crime ocorra? Antes de responder essa questão, é relevante pensar como obtê-las com supedâneo conceitual dos modelos da Ciência da Informação (CI).

Dentre os paradigmas da CI, o modelo cognitivo é um dos mais adequados ao meio policial. Tentar prever crimes é uma atividade semelhante ao modelo descrito por Capurro (2003, p. 8) quando afirma que, no paradigma cognitivo da CI, os processos informativos transformam ou não o usuário, entendido em primeiro lugar como sujeito cognoscente possuidor de "modelos mentais" do "mundo exterior".

Pressupondo esse paradigma, é possível pensar que empirismo policial permeia o meio investigativo, transformando o homem da polícia em um repositório de modelos mentais do mundo laboral que o cerca. O tamanho desse repositório é diretamente proporcional a experiência acumulada pelo policial. Trata-se da manifestação real do modelo cognitivo de Capurro.

Contemporaneamente é comum, no meio policial, a obtenção de demonstrações científicas daquilo que outrora já se conhecia empiricamente. Por meio do conhecimento resultante da experiência, mais referências empíricas o policial acumula, o que o torna uma fonte importante de proposições inaugurais da Ciência Policial.

O processo de escolha de variáveis que implicam em risco criminal é, portanto, um exercício baseado no empirismo policial, onde os modelos mentais são decorrentes da prática desde muito vivenciada.

Por óbvio, a escolha de cada uma das variáveis deve ser fundamentada, caso contrário sua concepção será criação de aleatoriedade *nonsense*.

4 Segundo **https://istoe.com.br/brasil-perde-cerca-de-r-200-bilhoes-por-ano-com-corrupcao-diz-mpf/**

Seguindo os pressupostos, um perfil bivariado foi adotado na aplicação da pesquisa. A primeira variável é categórica e segunda, quantitativa. Ambas são explicitadas a seguir, com fundamentações baseadas, principalmente, no empirismo policial adotado como um método de evidência inaugural.

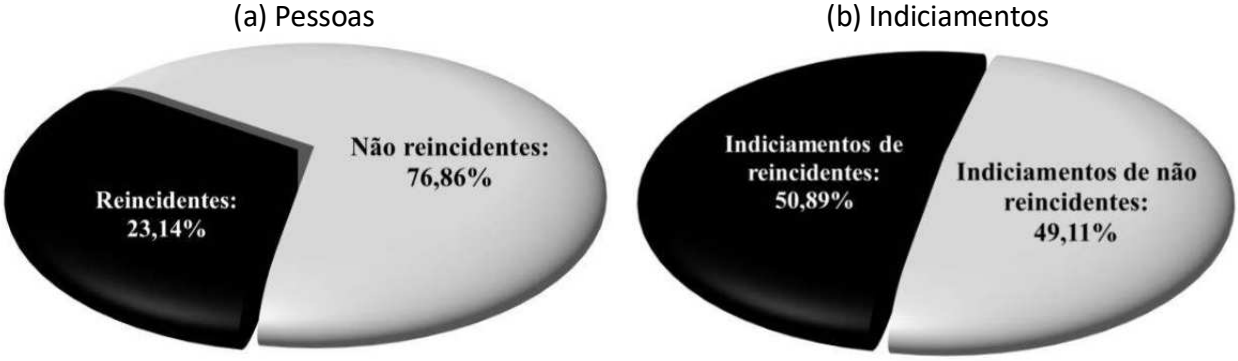
A primeira variável se relaciona com a reincidência criminal, fenômeno que parece ser um fator relevante na detecção de crimes futuros. Isso porque há um grande percentual de novos crimes praticados por pessoas outrora já acusadas (reincidentes).

Nos dados avaliados, das 1.327 (mil trezentas e vinte e sete) pessoas indiciadas em pelo menos um inquérito policial, 307 delas (trezentas e sete), ou seja, 23,14% do total, possuíam mais de um indiciamento, ou seja, são reincidentes.

Os reincidentes respondem a um total de 1.056 (mil e cinquenta e seis) indiciamentos, ou seja, 50,89% de um total de 2.075 (dois mil e setenta e cinco).

Assim, na população estatística utilizada na pesquisa, os reincidentes, os quais representam 23,14% das pessoas respondem por mais de 50% de todos os indiciamentos. Esse cenário é condensado na figura 2, abaixo:

Figura 2 – 23,14% das pessoas (os reincidentes) respondem por 50,89% de todos os indiciamentos



Fonte: Elaborado pelos autores

Por vezes as pessoas indiciadas se tornam sócias de empresas que, por sua vez, tornam-se fornecedores de órgãos públicos. Doravante, e para fins de simplificação textual, essas empresas dotadas de sócios indiciados por crimes serão chamadas de fornecedores acusados. A 1ª variável, portanto, é o indicativo booleano que indica se o fornecedor possui (ou não) sócios acusados de crimes.

Sabe-se que o princípio da presunção da inocência tutela, inclusive, essas pessoas acusadas. Isso não invalida a aplicação da presente pesquisa como método de predição criminal,

pois não há objetivo de gerar consequências negativas – sejam penais ou administrativas – que prejudiquem os fornecedores assim classificados.

A segunda variável escolhida é aquela que expressa a busca por altos ganhos financeiros: a quantidade de recursos públicos recebidos por um fornecedor. A escolha dessa variável se justifica porque cifras vultosas são perseguidas pelos delinquentes como forma de compensação aos perigos penais e sociais decorrentes do crime (reprovação social, prisão e condenação). Essa variável, portanto, é impactada pelo comportamento do criminoso. Trata-se de uma variável numérica e contínua.

Apesar do exposto, não se pode afirmar que o recebimento de vultosos recursos públicos é um indicador suficiente para predizer crimes, pois mesmo as pessoas não previamente acusadas buscam auferir ganhos mais altos.

Considere-se, entretanto, o grupo 1 aquele formado por fornecedores acusados e o grupo 2 formado por fornecedores destituídos de acusações; isto posto, propõe-se a seguinte questão: os maiores recebedores de recursos públicos possuem participações proporcionais de componentes de ambos os grupos?

A resposta a essa pergunta poderá indicar que uma característica pessoal – a acusação criminal (indiciamento) – pode alterar a rede social de contratações estatais em favor dos acusados (embora os motivos não sejam o alvo da corrente pesquisa).

A finalização dessa etapa corresponde a conclusão da tarefa número [1] do fluxograma ilustrado pela Figura 1.

3.2 Obtenção e tratamento dos dados

Antes da realização do estudo ora proposto foi necessário idealizar, sob uma ótica da CI, como seriam obtidos os insumos: as informações que dão embasamento à pesquisa. Coadic (1994, p. 27) relata que sem informação a ciência não pode se desenvolver e que a atividade de pesquisa constitui a aplicação do raciocínio ao corpo de conhecimentos acumulados ao longo do tempo e armazenados nas bibliotecas e centros de documentação. Ora, considerando que os dados disponíveis em sistemas de informática são novos tipos de centros de documentação, a ideia de Coadic foi aplicada a presente pesquisa. Isso deriva do fato de que a informação foi coletada em sistemas informacionais advindos de diversas instituições. A CI, portanto, fornece corpo à pesquisa, conforme será esmiuçado a seguir.

Os dados foram utilizados na pesquisa estão quantificados e qualificados quanto à origem na tabela 1, abaixo:

Tabela 1: Dados utilizados na pesquisa.

UNIDADE DE INFORMAÇÃO	QUANTIDADE	FONTE DA INFORMAÇÃO
Empresas	11.409	Secretaria da Fazenda da UF[5]
Inquéritos Policiais	380	Sistemas Cartorários da Polícia Federal [6]
Indiciamentos	4.139	Sistemas Cartorários da Polícia Federal
Sócios de Empresas	309.653	Secretaria de Fazenda da UF
Pagamentos de Recursos Públicos	583.072	Tribunal de Contas da UF
Órgãos públicos	224	Tribunal de Contas da UF

Fonte: Elaborado pelo(a) autor(a).

O tratamento dos dados, o qual envolveu remoção de inconsistências e valores nulos, resultou numa redução de menos de 1% do total inicial. O abatimento desse percentual já está contemplado na tabela 1 e não afetou significativamente os resultados.

A interdisciplinaridade característica da pesquisa permitiu o uso da Ciências da Computação pelo uso de plataformas de banco de dados baseados em grafos. A Ciência da informação contribuiu com técnicas de recuperação da informação tais como filtragem e agrupamento.

A finalização dessa etapa corresponde a conclusão da tarefa número [2] do fluxograma ilustrado pela Figura 1.

3.3 Análise exploratória dos dados: conhecendo o terreno

Segundo Pinheiro (2009, p. 12), a análise exploratória trata-se de um conjunto de técnicas de tratamento de dados que, sem implicar em uma fundamentação matemática mais rigorosa, nos ajuda a fazer uma sondagem do terreno, ou seja, a obter um primeiro contato com a informação disponível. Para execução dessa tarefa, a Ciência da Informação contribuiu via

[5] UF: Unidade Federativa.
[6] Todos os dados foram tornados anônimos.

análise da informação com vistas a obter a relevância informacional e depois apresenta-la com uso de técnicas de apresentação da informação.

A Ciência da Computação contribuiu pelo uso de uma linguagem de programação com ampla disponibilidade de bibliotecas gráficas, estatísticas e relacionadas a *data science*.

A estatística contribuiu com conceitos estatísticos básicos.

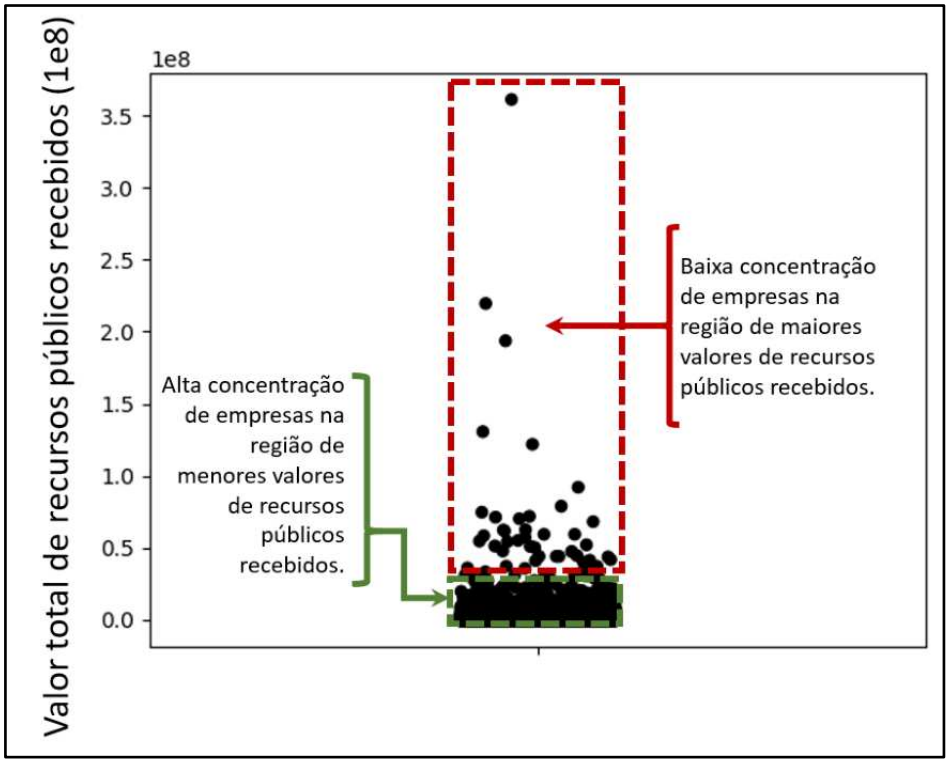
Ao lidar com um conjunto de dados, muitas vezes, a primeira atividade a ser realizada é obter uma ideia de como as variáveis estão distribuídas. Por isso, foi traçado um diagrama dos fornecedores e o valor recebido por cada um deles foi marcado na área de plotagem. Percebe-se, por meio de uma análise visual preliminar, a alta concentração de fornecedores na área de plotagem que corresponde aos menores valores (na base do gráfico).

Alguns poucos fornecedores estão dispersos no alto do gráfico, o que significa que receberam os maiores valores.

Os fornecedores marcados na parte superior da área de plotagem contrastam com a grande maioria das demais observações, cuja regra é a concentração basal da distribuição.

Há, portanto, um grau de dispersão da população que pode ser consideravelmente alto, conforme a figura 3 abaixo:

Figura 3: Gráfico de dispersão dos recursos públicos recebidos por empresas.

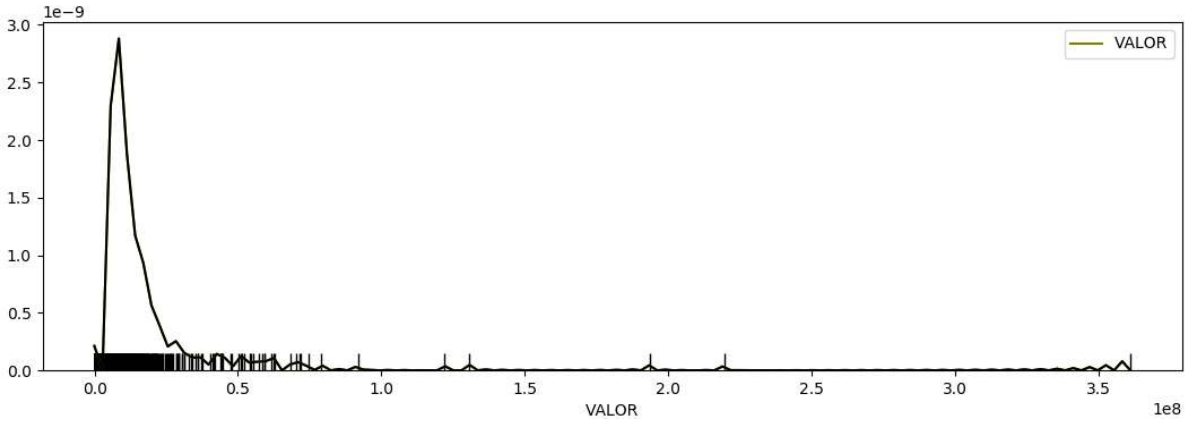


Fonte: Elaborado pelos autores

Outro gráfico que pode indicar a dispersão e variabilidade dos valores é o denominado *Kernel Density*, cuja finalidade é indicar a densidade de ocorrências de uma variável. No caso dos valores recebidos pelos fornecedores, percebe-se uma alta distribuição deles no gráfico.

Cada linha curta e perpendicular ao eixo das abscissas é a marca da ocorrência de um fornecedor. A alta concentração dessas linhas significa alta concentração dos fornecedores naquele intervalo de valor considerado. A maioria esmagadora dos fornecedores se situa nas faixas dos menores valores recebidos, conforme ilustra a Figura 4.

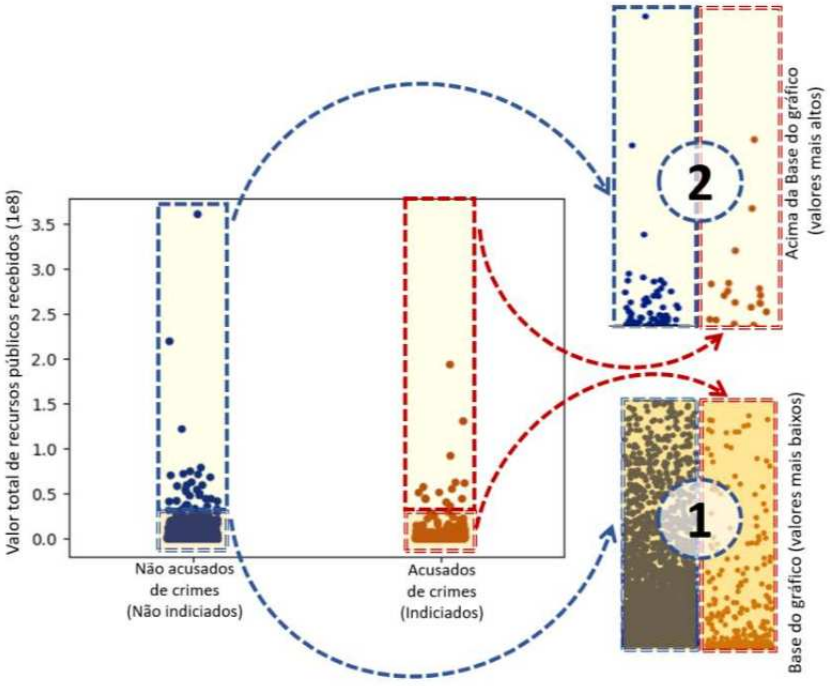
Figura 4: Gráfico de densidade por valor.



Fonte: Elaborado pelos autores

Também é possível separar as plotagens em duas áreas pela variável categórica que indica a presença (ou ausência) de acusações criminais prévias em um gráfico de dispersão, conforme a Figura 5:

Figura 5: Gráfico de dispersão dos recursos públicos recebidos por empresas, categorizadas por possuírem acusações criminais (ou não) contra os sócios.



Fonte: Elaborado pelos autores

A análise da base do gráfico (destacada no item ① da figura), onde os valores recebidos individualmente são menores, indica haver uma distinção muito latente onde os fornecedores que não possuem sócios acusados de crimes são uma classe muito predominante. Entretanto, quando se analisa a região superior do gráfico (destacadas no item ② da figura) ocorre um

equilíbrio entre as duas classes, o que induz a concluir que a medida que os valores recebidos aumentam, também aumenta a presença relativa de empresas com sócios acusados de crimes.

Ao que parece, os acusados estão mais presentes quando os valores recebidos são maiores, mas resta necessário comprovar essa afirmação com medidas estatísticas mais concretas.

A finalização dessa etapa corresponde a conclusão da tarefa número [3] do fluxograma ilustrado pela Figura 1.

3.4 Cálculo de métricas de centralidade e de dispersão

Nessa etapa também houve uso da interdisciplinaridade.

A Ciência da Computação contribui com o uso de uma linguagem de programação com ampla disponibilidade de bibliotecas gráficas, estatísticas e relacionadas a *data science*.

A Estatística contribuiu com métricas de dispersão e centralidade. Para entender melhor a dispersão das observações estatísticas é necessário compreensão da variabilidade dos dados populacionais. Para tanto, foram calculadas a média, o desvio padrão e o coeficiente de variação, cujos resultados são apresentados na tabela 2:

Tabela 2: Quadro de medidas da população.

Medida	Valor
Média (R$)	919.215,80
Variância $(R\$)_2$	37.160.004.557.207,50
Desvio Padrão (R$)	6.095.900,64
Coeficiente de variação (adimensional)	663,16%

Fonte: Elaborada pelos autores.

Percebe-se alta dispersão dos dados indicada pelo alto coeficiente de variação: 663,16%. A média nesse caso, portanto, é uma medida de centralidade pouco robusta como uma referência ao conjunto dos dados. A pouca representatividade da média em relação a população é um indicativo de que observações estatísticas extremas – os outliers – a estão distorcendo.

A análise do limite inferior, dos quartis e do limite superior pode ajudar a esclarecer a existência desses outliers. Abaixo, a tabela contendo os valores dos quartis (Q_1, Q_2 e Q_3), do limite superior (L_s) e do limite inferior (L_i):

Tabela 3: Quartis, valor máximo e valor mínimo.

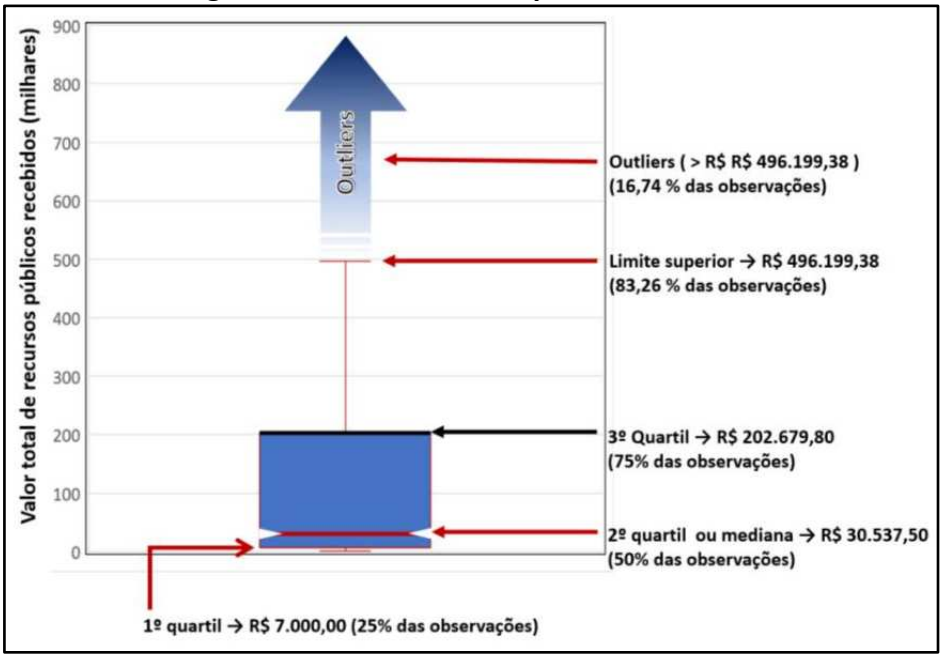
Medida	Valor (em R$)

Q_1 (1º Quartil, concentra 25% das empresas)	7.000,00
Q_2 (2º Quartil ou mediana, concentra 50% das empresas)	30.537,50
Q_3 (3º Quartil, concentra 75% das empresas)	202.679,80
Distância Interquartil ($Q_3 - Q_1$)	195.679,75
Limite Superior (L_s)	496.199,38
Limite Inferior (L_i)	Não relevante

Fonte: Elaborado pelo(a) autor(a).

O Limite Inferior foi ignorado porque não há observações estatísticas abaixo dele, logo não há outliers inferiores. Ao lançar os valores em um gráfico do tipo *boxplot,* obtemos o resultado da Figura 6:

Figura 6: Quartis, limite superior e Outliers.

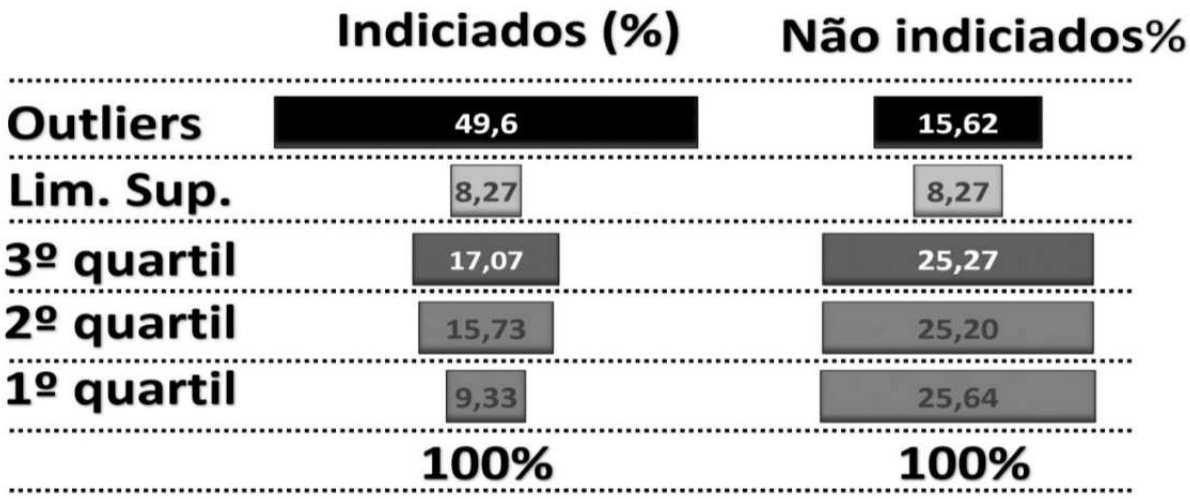


Fonte: Elaborado pelos autores

O *boxplot* evidenciou ainda mais a dispersão dos dados em relação à primeira variável utilizada: os valores de recursos públicos recebidos.

A partir de então a segunda variável, agora não mais numérica, mas categórica, foi alvo de análise: a presença de fornecedores acusados nos quartis e nos outliers. Os percentuais de acusados em cada um dos quartis foi expresso em um gráfico tipo funil, conforme a da figura 5:

Figura 5: Concentração de acusados de crimes dentre os outliers e nos principais percentis.



Fonte: Elaborado pelos autores

Percebe-se, pelo gráfico acima, que a presença de acusados dentre os outliers é muito alta. A finalização dessa etapa corresponde a conclusão da tarefa número [4] do fluxograma ilustrado pela Figura 1.

3.5 Predição da amostra a ser monitorada e entrega às equipes de monitoramento

A predição, produto da pesquisa, é a própria amostra de dados onde se sabe ser o risco criminal maior. Tendo em vista que reincidentes criminais possuem participação maior nos indiciamentos, o grupo que recebe maiores valores e que possui acusação criminal prévia é a amostra e produto da predição (resultado concreto da aplicação do modelo).

Uma consequência óbvia é que se esse grupo possui maior risco criminal que os demais, não é necessário reavaliar as variáveis, já que elas influenciam no risco de crime. Isso responde com um "sim" ao *gateway* exclusivo (tarefa que significa uma decisão que deve ser tomada), do passo [6] do fluxograma.

Caso as variáveis não influíssem no aumento do risco de crime, a resposta ao gateway exclusivo seria "não" e, por isso, as variáveis não seriam adequadas para a predição. Deveriam então ser reavaliadas. Com isso o fluxograma do modelo retornaria ao passo [1].

Entretanto, como as variáveis escolhidas influenciaram o risco, o próximo passo seria a entrega da amostra às equipes de policiamento preventivo para pôr em prática o monitoramento.

O profissional que fornece a predição poderá, quando entregar a amostra às equipes de policiamento preventivo, indicar contribuições das Ciências Policiais, tal como técnicas de policiamento preventivo e softwares que auxiliem esse monitoramento.

A depender do tipo de crime que se deseja prevenir, as equipes nominadas de policiamento preventivo podem ser, inclusive, compostas por órgãos de controle administrativo e auditoria, pois não há motivações para executar repressão policial a crimes ainda não existentes. A palavra "policiamento", nesse caso, refere-se ao sentido *lato senso* do termo, não se referindo especificamente a equipes de policiais.

A finalização dessa etapa corresponde a conclusão das tarefas número [5], [6] e [7] do fluxograma ilustrado pela Figura 1.

4 CONSIDERAÇÕES FINAIS

É possível propor um padrão de predição criminal de natureza interdisciplinar, com contribuições de ciências como a CI, a Computação, a Estatística e as Ciências Policiais. O modelo ora aventado pode ser aplicado a qualquer tipo de crime, desde que as variáveis que determinam o risco criminal sejam concebidas com um fundamento conceitual sólido. Esse fundamento dever ser baseado no tipo de crime e nas ciências interdisciplinares propostas, processo que se configura como uma aplicação concreta da equação fundamental da Ciência da Informação. Nesse contexto, se busca um novo estado do conhecimento mediante o incremento de informações fornecidas pelo empirismo policial.

Na aplicação do modelo em crimes de desvio de recursos públicos, os dados estatísticos demonstraram conclusões peculiares. Uma característica individual – a acusação criminal pré-existente – pode mudar o comportamento da rede social de contratações: nela, há concentração de acusados na região dos outliers. Os dados demonstraram que há 49,6% de fornecedores acusados foram classificados como outliers contra 15,62% de fornecedores sem acusações. Ser um outlier, nesse caso, significa estar na região de maiores ganhos financeiros.

E, como a reincidência criminal dos previamente acusados é maior, a região que detém a maioria deles estará mais predisposta a novas ocorrências de crimes.

Isto posto, o resultado da aplicação da pesquisa é a seguinte predição criminal: monitorar preventivamente o conjunto formado por outliers com acusações pré-existentes, com vistas a diminuir a incidência de crimes relacionados ao desvio de recursos públicos.

A medição da eficácia da aplicação da pesquisa não foi objeto do estudo porque seria alvo de um novo trabalho até mais complexo do que a proposição do modelo. Isso porque após vários anos seria possível avaliar se houve melhora nos índices criminais decorrentes da prevenção baseada na predição.

Dito isso, como proposta de trabalhos futuros, há a possibilidade de aplicar o modelo e medir, no decorrer do tempo, se houve a diminuição das estatísticas criminais para atestar a eficácia do modelo ora proposto.

REFERÊNCIAS

BROOKES, B. C. The foundations of information science: philosophical aspects. **Journal of Information Science**, v.2, p.125-133, 1980.

BRUCE, Peter C.; BRUCE, Andrew G.. **Practical Statistics for Data Scientists:** 50 Essential Concepts. Sebastopol: O'reilly, 2016.

BUSSAB, Wilton de O.; MORETTIN, Pedro A.. **Estatística Básica.** 6. ed. São Paulo: Saraiva, 2010.

CAPURRO, R. **Epistemologia e Ciência da Informação**. Belo Horizonte: V Encontro Nacional de Pesquisa em Ciência da Informação, 2003.

COADIC, Yves-françois Le. **A Ciência da Informação.** Brasília: Lemos Informação e Comunicação, 1994.

DEVORE, Jay L.. **Probabilidade e Estatística para Engenharia e Ciências.** 6. ed. São Paulo: Cengage Learning, 2006.

GLADWELL, Malcom. **Fora de Série**: Outliers. Rio de Janeiro: Sextante, 2008.

HAIR JUNIOR, Joseph F. et al. **Análise Multivariada de Dados.** 6. ed. Porto Alegre: Bookman, 2009.

HASLWANTER, Thomas. **An Introduction to Statistics with Python:** With Applications in the Life Sciences. Berlin: Springer, 2016.

LARSON, Ron; FARBER, Betsy. **Estatística Aplicada**. 6. ed. São Paulo: Pearson, 2016.

PINHEIRO, João Ismael D. et al. **Estatística Básica:** A Arte de Trabalhar com Dados. Rio de Janeiro: Elsevier, 2009.

PINHEIRO, João Ismael D. et al. **Probabilidade e Estatística:** Quantificando a Incerteza. Rio de Janeiro: Elsevier, 2012.

ROUSSEEUW, Peter J.; VAN ZOMEREN, Bert C. Unmasking multivariate outliers and leverage points. **Journal Of The American Statistical Association**. Londres, p. 633-651. set. 1990.

TAYEBI, Mohammad; GLÄSSER, Uwe. **Social Network Analysis in Predictive Policing**: Concepts, Models and Methods. Vancouver: Springer, 2016.