Versão do registro: https://www.sciencedirect.com/science/article/pii/S0167404820301619
Manuscrito_24661802c33db776d96dbe5f87465677

Previsão do número de intervenções de bombeiros por região com dados baseados em privacidade local diferencial

H´eber H. Arcolezia,ÿ , Jean-François Couchota , Selene Cernaa , Christophe ´echara Al Bouna , Guillaume Royerb , Guyeuxa , B , Xiaokui Xiaod

aFemto-ST Institute, Univ. Bourgogne Franche-Comt´e, UBFC, CNRS, Belfort, France bSDIS 25 - Service D ´epartemental d'Incendie et de Secours du Doubs, France cLab., Antonine University, Hadath-Baabda, Lebanon dSchool of Computing, National University of Singapore , Singapura

Resumo

Estudos estatísticos sobre o número e tipos de intervenções dos bombeiros por região são essenciais para melhorar o atendimento à população. É também uma etapa preliminar se quisermos prever essas intervenções para otimizar a colocação de recursos humanos e materiais dos corpos de bombeiros, por exemplo. No entanto, isso tipo de dado é sensível e deve ser tratado com o máximo cuidado. Em ordem para evitar qualquer vazamento de informações, pode-se pensar em anonimizá-los usando Privacidade Diferencial (DP), um método seguro por construção. Este trabalho foca em prever o número de intervenções dos bombeiros em determinadas localidades enquanto respeitando o forte conceito de DP. Uma abordagem de Privacidade Diferencial local foi usado pela primeira vez para anonimizar dados de localização. Estimadores estatísticos foram então aplicados para reconstruir um conjunto de dados sintéticos não correlacionados com os usuários. Finalmente, um abordagem de aprendizado supervisionado usando aumento de gradiente extremo foi usado para fazer as previsões. Experimentos mostraram que a previsão de anonimização método é muito preciso: a introdução de ruído para higienizar os dados não afetam a qualidade das previsões, e as previsões refletem fielmente o que

ÿEndereços de e-mail do autor para correspondência: heber.hwang_arcolezi@univ-fcomte.fr (H´eber H. Arcolezi), jean-francois.couchot@univ-fcomte.fr (Jean-Fran¸cois Couchot), selene_leya.cerna_nahuis@univ-fcomte.fr (Selene Cerna), christophe.guyeux@univ-fcomte.fr (Christophe Guyeux), guillaume. ROYER@sdis25.fr (Guillaume Royer), bechara.albouna@UA.EDU.LB (B ´echara Al Bouna), xkxiao@nus.edu.sg (Xiaokui Xiao)

Pré-impressão enviada para Computadores e Segurança 26 de dezembro de 2020

aconteceu na realidade.

Palavras-chave: privacidade diferencial local, mecanismo RAPPOR, bombeiros
localização da intervenção, previsão multialvo, XGBoost.

---

1. Introdução

O transporte médico de emergência inclui os vários serviços úteis para trans
transporte de pessoas feridas de suas casas ou do local do acidente para o
hospital mais capacitado para cuidar do paciente. Como esta emergência médica
transporte é implementado depende do país que está sendo considerado, sua história,
e as escolhas saudáveis que foram feitas no passado. Geralmente inclui o
serviços de transporte próprios dos hospitais e, muitas vezes, também de operadores privados especializados
(condutores de ambulância privados licenciados). Pode também incluir outros serviços públicos,
como brigadas de incêndio. Na França, por exemplo, este último não é apenas responsável
para extinguir incêndios, mas também está escrito em seu status que eles devem
encarregar-se de parte destes transportes médicos de emergência, e este encargo
representa mais de 80% da sua actividade.

Esta estruturação funcionou bem no passado, no entanto, tanto na França como na
vários outros países, temos enfrentado uma grande crise na emergência médica
transporte já há algum tempo, por várias razões. O envelhecimento da população
nos países ocidentais e o fato de que os idosos precisam de assistência com mais frequência
leva a uma maior demanda por transporte. O endividamento dos países e
a grande crise econômica da última década levaram seus governos
racionalizar ainda mais os gastos sociais, tomando medidas como o fechamento de pequenas
centros ou mudança para atendimento ambulatorial (os pacientes devem ser encaminhados para casa o
mais rápido possível para reduzir o número de leitos a serem administrados). No entanto, o fechamento
de pequenos centros não só leva à saturação de grandes centros de emergência, mas
também aumenta as distâncias a serem percorridas pelos transportadores de saúde. De forma similar,
atendimento ambulatorial aumenta o risco de re-hospitalização e, portanto, o retorno
percurso entre o hospital e a casa do paciente. O modelo econômico de
motoristas de ambulâncias particulares só é viável se a parte de “guarda” for fraca diante do

transporte médico planejado (excluindo emergências). Esses e outros elementos estão, portanto, levando a uma crise de emergência no transporte de saúde em várias partes do o mundo.

Uma das soluções pensadas para aliviar a pressão sobre estes transportadores é otimizar o uso de seus recursos, de forma a fortalecer as equipes durante horários de pico, reduzindo-os durante os fora de pico. Tais otimizações são geralmente implementado por pessoal assimétrico diurno e noturno e, às vezes, por a distinção entre dias úteis e fim de semana. Mas a situação de crise é de tal forma que agora é preciso ir muito além nessas otimizações, que requer uma visão relativamente clara das necessidades de curto, médio e longo prazo. Esta a previsão é possível até certo ponto, uma vez que este transporte médico de emergência atividade está diretamente relacionada à atividade humana: esta é reduzida à noite à medida que as pessoas dormem, há menos acidentes e a necessidade de transporte é consequentemente mais baixo à noite (daí os turnos reduzidos). No entanto, poderíamos ir muito mais longe, considerando que a atividade muda de acordo com as estações do ano (queda de peões no gelo no inverno, afogamento em piscinas no verão...), feriados, dias da semana, ocorrência de eventos programados como como festivais ou eventos, etc.

Não apenas o fluxo de intervenção pode, portanto, provavelmente ser previsto, mas também seu tipo e localização. De fato, prever o número de intervenções por local ção pode reduzir o tempo necessário para chegar ao local do acidente. Por exemplo, em megacidades e durante ondas de calor, áreas altamente poluídas (como rodoviárias) e com alta densidade de pessoas com risco de problemas respiratórios (os idosos, cuja distribuição geográfica é conhecida pelas estatísticas nacionais e institutos demográficos) são claramente sensíveis e pré-posicionam um ambu lança nesses locais permite uma ação mais rápida em caso de emergência de do tipo desconforto respiratório. Redução do tempo de chegada ao local da o acidente tem benefícios materiais, humanos e econômicos. Materiais, antes de tudo, porque é possível redistribuir recursos quando uma sobrecarga de intervenções em uma determinada área é esperada (por exemplo, devido à inundação de certos rios): com visibilidade, é possível otimizar o uso dos recursos atuais. Além disso, chegando

no local o mais rápido possível é crucial no caso de incêndios, e chegar a
o início do incêndio permite limitar os danos e economizar prop
erty e edifícios. Essa previsão também permite otimizar o uso de
humanos, mas também para salvar vidas em situações como parada cardíaca e
afogamento, para o qual cada segundo conta. E essas otimizações se traduzem em
benefícios econômicos, tanto por causa da salvaguarda da propriedade, como porque
a morte prematura tem um custo social significativo.

O aumento dos níveis de água após fortes chuvas leva a eventos de inundação
perto de rios, envolvendo resgate pessoal. Estradas de alta altitude têm um risco maior
de cobertura de neve no inverno do que as de baixa altitude, aumentando o risco de acidentes
amassados, etc. É por isso que alguns autores recentemente procuraram explorar
técnicas de inteligência [1, 2, 3], baseadas em características que condicionam a atividade humana
(variáveis meteorológicas, informações de tráfego rodoviário, monitoramento de epidemias, etc.),
a fim de prever a demanda futura no transporte médico de emergência. No entanto,
para ser supervisionado, o aprendizado automático requer a capacidade de colocar o número de
intervenções ao longo do período (hora, dia...) para o qual temos estas explicações
variáveis históricas. Em outras palavras, é necessário ter acesso à intervenção
fluxo dos operadores cuja carga estamos tentando prever (ambulância particular
motoristas, bombeiros, etc.). Os últimos geralmente não têm nem o humano e o companheiro
recursos riais nem a competência para implantar soluções baseadas em inteligência artificial
e são, portanto, obrigados a transmitir esses dados a um terceiro confiável
com esta capacidade ou liberar seus dados para que o mundo acadêmico ou privado
operadores podem propor soluções de aprendizado de máquina ad hoc.

A divulgação desses dados é, portanto, de inegável interesse e pode ajudar
para fornecer soluções para a crise de transporte de saúde de emergência. Mas este lançamento
dos fluxos de intervenção é, por sua vez, problemático. Em primeiro lugar, são dados pessoais,
e várias estruturas legais naturalmente bloqueiam sua divulgação. Então é sensível
dados, ligados a acidentes, ao resgate de pessoas, possíveis óbitos. como saúde
as transportadoras trabalham de forma just-in-time e urgente, erros humanos ou organizacionais
são sempre possíveis, o que pode trazer sérias consequências, levar a ações judiciais, etc.
É por isso que esses dados, que foram divulgados recentemente para ver

ferramentas aparecem, foram liberados após anonimização. Na França, por exemplo, nós
teve recentemente duas publicações de tais fluxos em data.gouv.fr, um site do governo
dedicado a tais iniciativas, em uma abordagem de dados abertos. A primeira diz respeito ao
Intervenções 2007-2017 do Service D´epartemental d'Incendies et de Secours
de Saˆone-et-Loire (SDIS 71), contendo o número de intervenções por tipo e
por município [4], enquanto o segundo diz respeito aos mesmos tipos de dados para SDIS 91
(departamento de Essonne) para o período 2010-2018 [5]. Em cada caso, anonimização
foi feito por agregação: mensalmente para o primeiro conjunto de dados e semanalmente para o
segundo.

Embora a intenção desses SDIS seja louvável, a forma como eles lançaram
esses dados apresentam dois problemas: a anonimização alcançada é muito forte
e muito fraco. Muito forte, em primeiro lugar, porque realizar uma agregação
por mês resulta na perda de todas as informações úteis e resume o
intervenções em uma nuvem de 120 pontos (12 por ano), para os quais apenas um simples
a regressão linear permanece possível: impossível imaginar o aprendizado de máquina com
tal conjunto de dados - isso é verdade, em menor grau, para dados agregados semanalmente.
Então, muito baixo, porque essa agregação por mês, ou por semana, era feita
de forma cega e generalizada: se algumas comunas têm uma população suficientemente
número de intervenções, o que permite uma agregação temporal simples para alcançar
anonimização dos dados, outros, inversamente, não têm o suficiente. Dentro do estojo
de agregação mensal, por exemplo, são mais de 600 situações em que
houve apenas uma intervenção em uma comuna em um determinado mês: neste
nível, o simples 2-anonimato [6] não é mais satisfeito, e as informações
vazamento é óbvio. Tais vazamentos de informações também são numerosos no caso de
dados semanais e anonimização falhou para ambos os conjuntos de dados. Ao analisar
Neste arquivo, aprendemos, por exemplo, que na comuna de Ballore (FR-71220), ocorreu uma
intervenção do corpo de bombeiros em agosto de 2014. Considerando que
o município tem 86 habitantes, não seria muito difícil encontrar o
pessoa que recebeu ajuda este mês.

O objetivo deste artigo é, portanto, mostrar que é possível processar
tais fluxos de forma que 1) o anonimato seja garantido, e 2) previsões corretas

pode ser feito por aprendizado automático nesses dados. Isso é verdade mesmo que os dados considerados têm densidades espaciais muito variáveis.

A anonimização de dados é de fato um campo de pesquisa muito ativo e grandes avanços como a Privacidade Diferencial (DP) [7] permitem encontrar um meio justo compromisso entre privacidade e informações contidas. E agora é possível capaz de preservar tanto a segurança quanto a utilidade dos dados liberados [8]. O primeiro de enfim, a anonimização visa proteger as informações sobre cada indivíduo, quando o machine learning busca entender tendências gerais, grupais (periodicidade, sazonalidade, etc.): esses dois objetivos, portanto, não têm, a priori, razão de ser oposto. E várias semelhanças podem ser destacadas nessas duas abordagens. Por exemplo, indivíduos que se destacam na multidão são obviamente problemas atic e não pode ser preservado se o objetivo for produzir dados anonimizados; esses indivíduos também representam um problema durante a aprendizagem e são freqüentemente disfuncionais perdidos como outliers. Da mesma forma, os dados de aprendizado geralmente são ruidosos, e esse ruído é geralmente não uniforme. Essa assimetria na parte não informativa do sig nal torna o aprendizado mais complexo. Por outro lado, a adição de ruído uniforme é um método clássico de anonimização de dados, e essa adição pode, de certa forma, suavizar a parte do sinal de aprendizado que é polarizada por ruído não uniforme.

Com esses elementos em mente, neste artigo, nosso objetivo é aplicar um versão padronizada de Privacidade Diferencial, para transformar dados reais para que ambos sejam adequadamente anonimizados e úteis para aprendizado automático. Mais especificamente, em esta configuração local (nomeadamente Privacidade Diferencial Local - LDP), cada utilizador perturba seus dados antes de enviá-los para o servidor não confiável. O LDP tem sido amplamente aplicado e aceitos no processo de coleta de dados. Google aplica o RAPPOR mecanismo [9] para coletar comportamento de navegação na web e configurações do usuário no Chrome. Apple [10] aplica o LDP para coletar estatísticas populacionais com o objetivo de encontrar emojis e novas palavras comumente usadas. Para outros domínios de aplicação, [11] aplicou o LDP para coletar dados de posição interna; os autores em [12] propuseram uma variante do LDP adequada para espaços métricos (por exemplo, dados de localização); e [13] propuseram um protocolo para encontrar itens frequentes na configuração LDP de valor definido.

Depois de receber os dados ruidosos do LDP, o servidor pode calcular a população

estatísticas sobre o conjunto de dados sanitizado. Esses dados processados são então usados para propósitos de aprendizagem e predição: uma tarefa de predição múltipla do número de intervenções por região, com dados brutos e anonimizados, é então proposta.

Abordagens consideradas abrangem o uso de memória de longo prazo para o número total de intervenções [2]; um perceptron multicamadas para o número total de intervenções novamente [3]; e, finalmente, o uso de XGBoost em intervalo de tempo de 3h, um modelo por duas regiões importantes e modelos por motivo [8].

Na Figura 1, um fluxograma resume a abordagem proposta e implementada nesse papel. Primeiramente, o algoritmo toma como entrada a base de dados bruta (apresentada na Seção 2) e o parâmetro de DP (onde seu embasamento teórico é explicado na Seção 3). Em segundo lugar, o mecanismo baseado em LDP é aplicado para anonimizar cada ponto de dados (localização de uma intervenção), que é apresentado como uma metodologia para uma coleta de dados que preserva a privacidade na Seção 4. Em terceiro lugar, uma abordagem intuitiva baseada em estatísticas é usada para estimar estatísticas e construir um conjunto de dados sintético (abordagem não interativa em DP), que é detalhado na Seção ção 6. Tanto a segunda quanto a terceira etapas são baseadas no mecanismo RAPPOR apresentado em [9]. Finalmente, usando uma versão anônima do conjunto de dados (sintético), a técnica XGBoost é treinada e testada para a tarefa específica de prever o número de intervenções por região (apresentado na Seção 7). Este artigo termina com uma seção de conclusão, na qual a contribuição é resumida e o trabalho futuro pretendido é delineado.

2. Apresentação de dados

A base de dados à nossa disposição foi fornecida pelo corpo de bombeiros e salvamento mento, SDIS 25, na região de Doubs-França. Este arquivo contém informações sobre 382.046 intervenções atendidas pelo corpo de bombeiros de 2006 a 2018 dentro de suas departamento. Cada intervenção é registrada em um arquivo como uma linha e a principal em tributos deste arquivo são mostrados na Tabela 1 com informações artificiais e descrito a seguir:

- ID é o identificador da intervenção, que é utilizado em arquivos complementares;

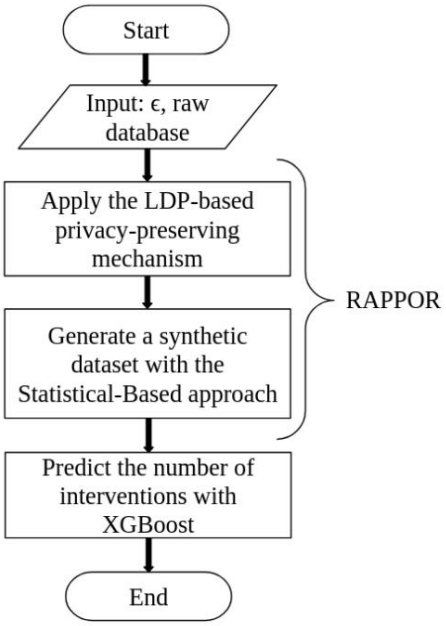

Figura 1: Fluxograma da abordagem proposta e implementada.

EU IRIA	SDate	Estação	Cidade	Localização
8	2008/08/08 08:08	Besan¸con	East Besan¸con	(47.2380, 6.0243)

Tabela 1: Principais atributos dos dados de operações do corpo de bombeiros

180 • SDate é a data de início da intervenção;

• Estação é o nome do corpo de bombeiros que atendeu a intervenção;

• Vila é o nome do município onde ocorreu a operação;

• Localização dá a localização precisa (latitude, longitude) da intervenção.

Além disso, a Tabela 2 apresenta a análise dos dados das intervenções agrupadas por dia
185 em cada ano. As métricas são o número total de intervenções (Total Interv.),
a média (Average), o desvio padrão (Std. Dev.), o máximo e
o número mínimo de intervenções (Max. e Min. Interv.). Como se pode
veja na Tabela 2 que há um grande incremento no número de intervenções ao longo

os anos. Ou seja, em 10 anos o número de intervenções duplicou de
17.333 em 2006 para 34.436 em 2016 e continuou aumentando até 40.510 em
2018. Este incremento representa mais trabalho para os próximos anos, onde uma melhor
otimização de recursos deve ser considerada para continuar melhorando a resposta
tempos às ocorrências e para melhor atender a população.

Ano Total Interv. Padrão médio Dev. máx. Interv. mín. Interv.					
2006	17.333	47	20	131	17
2007	19.277	53	16	116	23
2008	18.021	49	14	117	26
2009	28.669	79	38	257	22
2010	29.604	81	26	328	42
2011	33.645	92	39	403	48
2012	29.079	79	16	143	52
2013	29.760	82	14	145	47
2014	30.641	84	14	164	54
2015	33.518	92	17	154	57
2016	34.436	94	28	556	60
2017	37.553	102	16	165	61
2018	40.510	111	21	265	73

Tabela 2: Análise dos dados das intervenções durante 2006-2018.

3. Fundamento teórico sobre privacidade diferencial (local)

Seja A um algoritmo usado para publicar informações agregadas de um
base de dados. Privacidade diferencial (DP)[14] é como uma restrição (propriedade) em A que
limita a divulgação de informações privadas de registros cujas informações estão em
o banco de dados. Grosso modo, A é diferencialmente privado se um observador vendo sua saída
não pode dizer se as informações de um determinado indivíduo foram usadas no cálculo.

Seja um número real positivo que intuitivamente corresponde ao vazamento

nível. Quanto maior o valor dessa variável, mais importante é a informação
vazamento de informação. Seja im (A) denota a imagem de A, ou seja, o conjunto de todos os possíveis
resultados por A. Diz-se que o algoritmo A fornece -privacidade diferencial se, por
todos os conjuntos de dados D1 e D2 que diferem nos dados de uma pessoa e para todos os subconjuntos
205 R de im (A), temos

$$\Pr[A(D1) \ddot{y} R] \ddot{y} e \times \Pr[A(D2) \ddot{y} R]. \quad (1)$$

Intuitivamente, dado Pr[A(D2) ÿ R] (a probabilidade de que um conjunto de dados D2 possa ser
anonimizado em um elemento de R) e dada a quantidade de vazamento. Esta
equação dá um limite superior da probabilidade de que um conjunto de dados D1 pode ser
anonimizado em um elemento de R, que é, portanto, um vazamento de informações.

210 A privacidade diferencial permite a composição (de mecanismos independentes que
são 1, . . . , n DP. . .), robustez ao pós-processamento (F(A) é n DP para qualquer
função F).

No entanto, esta abordagem requer que todo o conjunto de dados seja completo, armazenado em
de forma segura e ainda mais anonimizada. A anonimização não é feita antes. Isso é
215 o objetivo da privacidade diferencial local introduzido em [15]. Nesta abordagem,
os dados são higienizados pelo usuário de forma probabilística antes de enviá-los para
o coletor. Um exemplo simples é pedir a uma pessoa que responda à pergunta “Faça
você mora em Belfort?”, conforme procedimento a seguir:

Jogue uma moeda.

220 • Se sair coroa, jogue a moeda novamente (ignorando o resultado) e responda à pergunta
pergunta honestamente.

• Se for cara, jogue a moeda novamente e responda “Sim” se for cara, “Não” se for coroa.

Este método estocástico básico é resumido na Figura 2. Seja ty o
proporção de respostas “sim” verdadeiras e cy é a proporção de “sim” observado
respostas. A equação a seguir fornece uma relação estimada entre esses dois
variáveis

$$\frac{1}{2} - cy. \frac{1}{4} \quad \frac{ty + \ddot{y}}{}$$

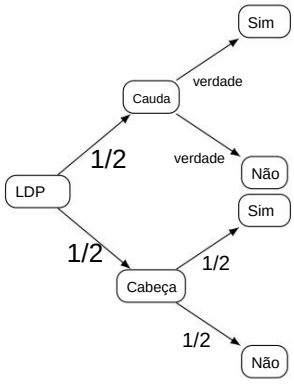

Figura 2: Resumo do que é enviado ao coletor pelo LDP básico

Quanto maior o número de experimentos, mais próxima a proporção de resultados aleatórios As respostas “Sim” serão até 1/4 e quanto mais próximo for o número de vezes que a verdade é informado, mais precisa será a estimativa. Neste caso, ty pode ser estimado de

$$1 \text{ ty ÿ } 2.\text{cy ÿ } \frac{}{2}.$$

Diz-se que o algoritmo A fornece privacidade diferencial -local se, para todos os pares dos possíveis dados privados do usuário v1 e v2 e todos os subconjuntos R de imA:

$$\Pr[A(v1) \text{ ÿ } R] \text{ ÿ } e \times \Pr[A(v2) \text{ ÿ } R]. \quad (2)$$

4. Coleta de Dados de Localização de Intervenções de Bombeiros para Preservação da Privacidade ção (lado do usuário)

A primeira pergunta que se pode fazer é se uma intervenção é um atributo sensível. A resposta é certamente sim porque o corpo de bombeiros não teria sido chamado se a situação não tivesse sido grave o suficiente. Por exemplo, considere o cenário onde uma pessoa que mora em uma cidade pequena adquiriu uma doença muito particular facilidade. Se é sabido que neste período ocorreu uma intervenção nesta localidade onde normalmente raramente acontece, há uma grande probabilidade de que o corpo de bombeiros interveio por esta pessoa.

Portanto, o objetivo desta tarefa é implementar um sistema de preservação da privacidade mecanismo de localização de intervenção dos bombeiros utilizando o conceito de local dif privacidade diferencial descrita anteriormente. Em seguida, dado um período específico, o chal O objetivo é estimar o número de intervenções dos bombeiros dentro do locais usando os dados anônimos para construir um conjunto de dados sintético. Para resumir, mais do que determinar com precisão as coordenadas de cada intervenção, o objetivo deste trabalho está ocultando a informação do local da intervenção de forma que estatisticamente tiques sobre o número de intervenções por local podem ser adquiridos com Utilitário. A Figura 3 ilustra um esboço da abordagem e é resumida em a seguir.

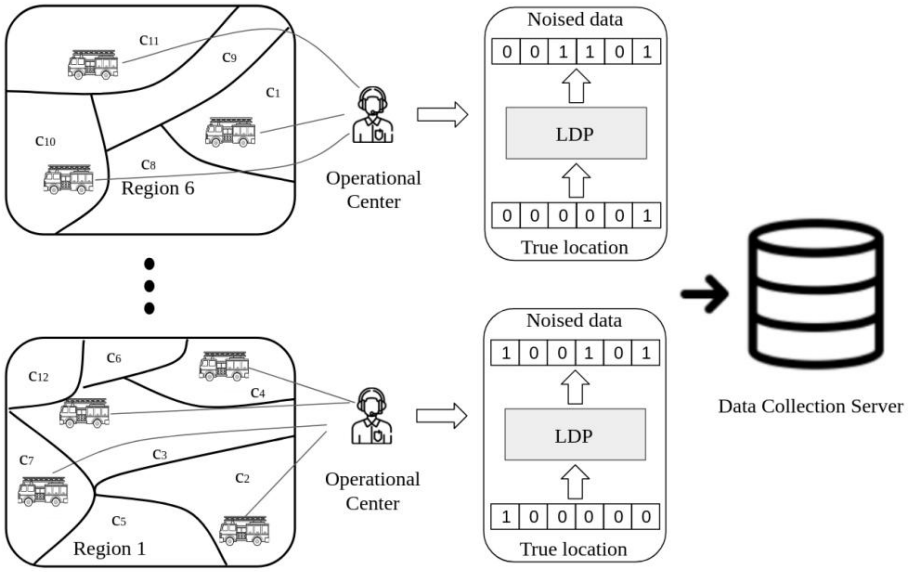

Figura 3: Um esboço da abordagem aplicada para coletar os dados de localização da intervenção dos bombeiros preservando a privacidade.

Na abordagem proposta, o primeiro passo para garantir a privacidade de cada a localização das intervenções é o agrupamento dos municípios onde aconteceram cada intervenção nível de uma cidade maior (região) para obter eventos suficientemente representativos representativo em número. Por exemplo, pode-se notar na Figura 3 que um conjunto de C = {c1, c2, ..., c12, ..., cm} pequenas cidades são agrupadas em n = 6 regiões.

Neste contexto, com base nos dados de que dispomos, 608 localidades onde foram realizadas intervenções ocorridos no departamento de Doubs foram generalizados para n = 17 regiões usando o conjunto de dados público disponível em [16]. As 17 regiões são: (1) CA du Grand Besan¸con, (2) CA Pays de Montbéliard Agglomeration, (3) CC Altitude 800, (4) CC de Montbenoit, (5) CC des Deux Vall´ees Vertes, (6) CC des Lacs et Montagnes du Haut-Doubs, (7) CC des Portes du Haut-Doubs, (8) CC du Doubs Baumois, (9) CC du Grand Pontarlier, (10) CC du Pays d'H´ericourt, (11) CC du Pays de Maˆÿche, (12) CC du Pays de Sancey-Belleherbe, (13) CC du Plateau de Frasne et du Val Rasne et du Val de Drugeon (CFD), (14) CC du Plateau de Russey, (15) CC du Val de Morteau, (16) CC du Val Marnaysien, (17) CC Loue-Lison.

A Figura 4 ilustra o departamento de Doubs com as respectivas cidades e seus aglomeração de regiões.

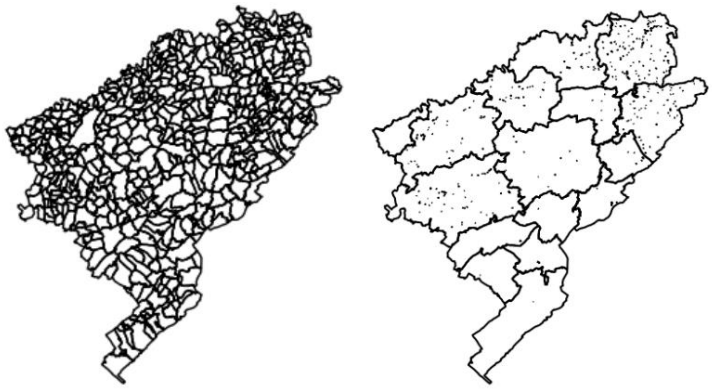

Figura 4: Municípios do departamento de Doubs agrupados por regiões.

Em segundo lugar, para melhorar o nível de privacidade de cada intervenção, o LDP Mecanismo baseado em “Basic One-time RAPPOR” introduzido por [9] é aplicado. Esta algoritmo é uma simplificação do mecanismo RAPPOR, que usa bloom filtros e funções hash para mapear os relatórios enviados pelos usuários e possui dois níveis de respostas aleatórias, ou seja, permanentes e instantâneas.

No entanto, no “Basic One-time RAPPOR”, é aplicado apenas um passo de resposta aleatória usando um mapeamento determinístico das n = 17 regiões em

vetores one-hot-codificados. A motivação para usar este algoritmo direto
270 é baseado em duas suposições:

- As aglomerações de cidades (regiões) são conhecidas a priori permitindo a mapeamento determinístico em vez de usar funções hash e filtros bloom;
- A localização exata de cada intervenção tem coordenadas únicas (x, y), que permite enviar um único relatório por intervenção com base em seu grande
275 aglomeração. Observe na Figura 3 que todas as intervenções que aconteceram na área de “Região 1” reportará um local com ruído baseado no mesmo valor real.

Uma aplicação técnica deste algoritmo em nosso estudo de caso é descrita abaixo:

1. Sinal de localização real. Seja R = {r1, r2, ..., rn} um conjunto de n regiões em
280 consideração, onde cada subscrito representa um ID de região exclusivo. Conseqüentemente, uma matriz de n bits, B (que denota o local de intervenção atual) é de multado como

$$Bk = \begin{cases} 1, \text{ se } k = i \\ 0, \text{ caso contrário} \end{cases} \quad (3)$$

onde neste caso, Bk representa o valor do k-ésimo bit em B com k ÿ [1, n]. Ou seja, o bit correspondente ao ID da região é definido como um,
285 enquanto os outros são definidos como zero (como o local verdadeiro na Figura 3).

2. Resposta aleatória permanente. Em seguida, cada bit em B (do pré passo anterior) é perturbado pela aplicação do conceito de resposta aleatória do seguinte modo:

$$\text{Reino Unido} = \begin{cases} 1, \text{ com probabilidade } \frac{f}{2} \\ 0, \text{ com probabilidade } \frac{f}{2} \\ Bk, \text{ com probabilidade } 1 - f \end{cases} \quad (4)$$

onde f é um valor de probabilidade entre 0 e 1, que controla o nível de garantia de privacidade diferencial (ver [9] para provas matemáticas). 1

pode perceber a relação direta entre privacidade e utilidade variando f
onde aumentá-lo garante mais privacidade com o custo de adicionar mais
290 ruído de B para U.

3. Relatório final. A resposta aleatória permanente B é transmitida para
o servidor coletor de dados.

O nível de privacidade diferencial -local foi mostrado em [9] como sendo o pior ý
definido como

$$ý = 2 \ln \frac{1 - \frac{1}{2}f}{\frac{1}{2}f}. \quad (5)$$

A prova original contém etapas que não são fáceis de seguir. O Apêndice A
apresenta outra prova para este valor.

295 4.1. Exemplo

Suponha que ocorreu uma intervenção na área de r3, que representa o
3ª região de n = 8 unidades. Portanto, seu verdadeiro sinal de localização B é descrito
do seguinte modo:

$$B = [0, 0, 1, 0, 0, 0, 0, 0] \quad (6)$$

considerando que pode-se ver que o 3º bit de B é definido como um. Nesta etapa, o
300 garantia de privacidade do local de intervenção é assegurada pela aglomeração
área, no entanto, em vários cenários, um invasor pode fazer uso de fundo
conhecimento e fontes externas para inferir a localização exata (neste caso, o
Cidade). Portanto, aplicando a Equação (4) com, por exemplo, f = 0,3, uma possível
resposta permanente U é a seguinte:

$$U = [1, 0, 1, 0, 0, 1, 0, 0] \quad (7)$$

305 onde dadas propriedades aleatórias dependendo de f, tanto o 1º bit quanto o 6º são
também definido como um.

Portanto, como se pode ver, as informações de localização não são mais fáceis de serem
descoberto, pois o conceito de LDP garante que qualquer região verdadeira (entrada) possa
geraram a saída com ruído U com uma razão de probabilidade limitada de e ÿ.

5. Gerando um conjunto de dados sintético (lado do servidor)

Considerando um período específico de estudo, o objetivo é estimar o número de intervenções por local associado à i-ésima região, ri . Nesse contexto, um conjunto de dados sintético pode ser construído com essa estimativa, que é considerada como um caso não interativo de DP. Mais especificamente, este conjunto de dados é gerado por estatísticas usando apenas dados de localização anônimos e são liberadas apenas uma vez para todas as outras tarefas pretendidas.

Portanto, dentro de um tempo específico, seja set(U) um conjunto de respostas e set(B) seja o conjunto correspondente de matrizes de bits de localização originais. Além disso, suponha que |set(U)| e |set(B)| denotar o número de elementos em cada respectivo conjunto. Naturalmente, |set(U)| = |conjunto(B)|.

Portanto, o número estimado de intervenções NBintest por região O cátion ri para i ÿ [1, n] é adquirido por uma abordagem baseada em estatística (SB) como segue [9]:

$$\text{NBint}(ri) = \frac{1}{\ddot{y}\ f\ 2} \cdot \text{Ni}\ \ddot{y}\ 1 \quad \frac{f \cdot \text{Ntotal}}{} \tag{8}$$

onde Ntotal é o número de respostas aleatórias permanentes |conjunto(U)| e Ni é o número total de respostas aleatórias permanentes cujo i-ésimo bit é definido como 1. Vale ressaltar que a Equação (8) pode estimar números negativos, portanto, a função max(0, NBintest) é usada.

Para avaliar o resultado do SB, a estimativa de densidade de um local da i-ésima região associado com ri é calculado da seguinte forma [11]:

$$\text{Densidade máxima}(ri) = \frac{\text{NBint}(ri)}{\sum_{y=1}^{n} \text{NBint}(ry)} \tag{9}$$

onde n é o número da região e, portanto, a métrica da taxa de erro (ER) é definido como:

$$\text{ER} = \frac{1}{n} \sum_{i=1}^{n} |\text{Densidadereal}(ri)\ \ddot{y}\ \text{Densidadeest}(ri)| \tag{10}$$

onde Densityactual(ri) e Densityest(ri) correspondem ao real e es
densidade estimada, respectivamente, da região associada ao i-ésimo local.
Em vez de calcular a raiz do erro quadrático médio sobre o estimado e ac
número real de intervenções, a taxa de erro é calculada sobre o valor da densidade
motivados para valores normalizados entre 0 e 1.

6. Experiências de Anonimização

Avaliar a abordagem proposta de anonimizar a intervenção dos bombeiros
localização, várias simulações são realizadas com diferentes valores de f, que
determina o nível de privacidade ÿ-diferencial. Nos experimentos, f será
variam em [0,1, 0,2, ..., 0,8, 0,9], o que garante privacidade ÿ-diferencial entre
[5,89, 4,39, ..., 0,81, 0,4].

Portanto, usando a abordagem baseada em estatística (Equação (8)), o objetivo é
estimar o número de intervenções por região considerando diferentes cenários
de tempo. Os cenários de tempo são descritos a seguir. O primeiro a
analisar é com dados de um ano (13 pontos de dados), o que permite no início
de um ano o corpo de bombeiros para melhor distribuir seu orçamento em seus centros
de acordo com o número de intervenções por região. Em seguida, um cenário de um mês
(156 pontos de dados) é considerado. E, como antes, o corpo de bombeiros pode
têm estatísticas de alta utilidade de uma empresa terceirizada para reorganizar orçamentos
e pessoal a cada mês. Por último, um cenário de um dia (4748 pontos de dados) é
levados em consideração para que as tarefas de aprendizado de máquina possam ser aplicadas em
esta quantidade de dados.

Esses experimentos permitirão avaliar a relação entre ER versus
tamanho dos dados (período de análise) de acordo com $_{y}$ para encontrar a melhor privacidade
compensação de utilidade para diferentes aplicações. Cada cenário permite ao corpo de bombeiros
ter um banco de dados anônimo de locais de intervenção onde terceiros
empresas ou o próprio departamento de recursos humanos poderiam adquirir
Estatisticas. Mais especificamente, conjuntos de dados sintéticos serão construídos com base no SB
Abordagem 360º , que conterá o número de intervenções por região para cada

cenário do tempo.

6.1. Resultados

Por uma questão de brevidade, considerando apenas três valores para $\ddot{y}$ = [5,89, 2,19, 0,40] (resp. f = [0,1, 0,5, 0,9]), a Tabela 3 apresenta os seguintes resultados
365 rics: o ER médio (ER Av.), o desvio padrão ER (ER. Std.), os erros mínimo (Min. ER) e máximo (Max. ER), para cada cenário de
Tempo. Ou seja, conforme as estatísticas são adquiridas, por exemplo, para cada ano, o erro serão resumidos de uma só vez (considerando todos os anos) na Tabela 3.

Para melhor ilustrar os resultados da Tabela 3, a Figura 5 mostra a relação
370 do ER e $\ddot{y}$ para: todos os anos (2006-2018), com zoom para os últimos 8 meses de 2018 e com zoom para os últimos 8 dias de dezembro de 2018, respectivamente. Mais Além disso, a Figura 6 ilustra as estatísticas obtidas sobre o número de intervenções para o ano de 2013, o primeiro mês de 2017 e um dia preciso em janeiro de 2016, com os três valores para $\ddot{y}$ = [5,89, 2,19, 0,40] (f = [0,1, 0,5, 0,9] um baixo, um
375 médio e uma alta garantia de privacidade). Todas as três datas específicas foram escolhidas em aleatório para fins de ilustração.

$\ddot{y}$	Cenário ER Av.	ER Std.	mín. ER	máx. emergência	
5.89	Um ano 0,001209	0,000271		0,000894	0,001750
	Um mês 0,004045	0,001081		0,002054	0,007426
	Um dia 0,017675	0,005265		0,005115	0,048845
2.20	Um ano 0,003992	0,000922		0,002116	0,005644
	Um mês 0,012813	0,002920		0,006475	0,021311
	Um dia 0,042584	0,010536		0,014006	0,092509
0,40	Um ano 0,018785	0,003726		0,012430	0,024008
	Um mês 0,043107	0,010174	0,022024		0,070537
	Um dia 0,077103	0,015029		0,029918	0,117647

Tabela 3: Resultados das métricas para comparação do ER em diferentes cenários de tempo e ÿ-diferencial privacidade.

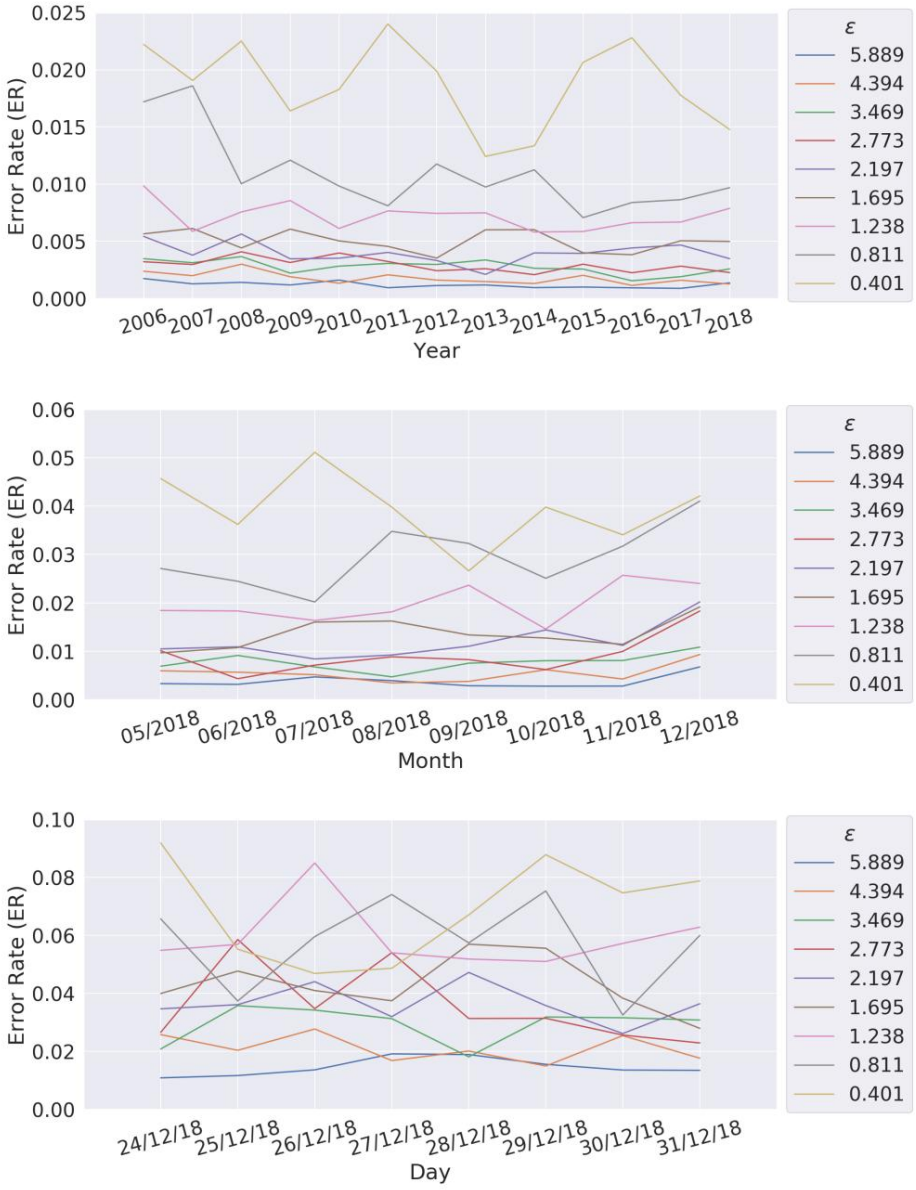

Figura 5: Comparação entre taxa de erro e período de análise (tamanho dos dados) variando ÿ.

6.2. Discussões

Como se pode notar na Tabela 3 e nas Figuras 5 e 6, o mecanismo baseado em LDP pode ser bem aplicado para a coleta de localização das intervenções dos bombeiros para o

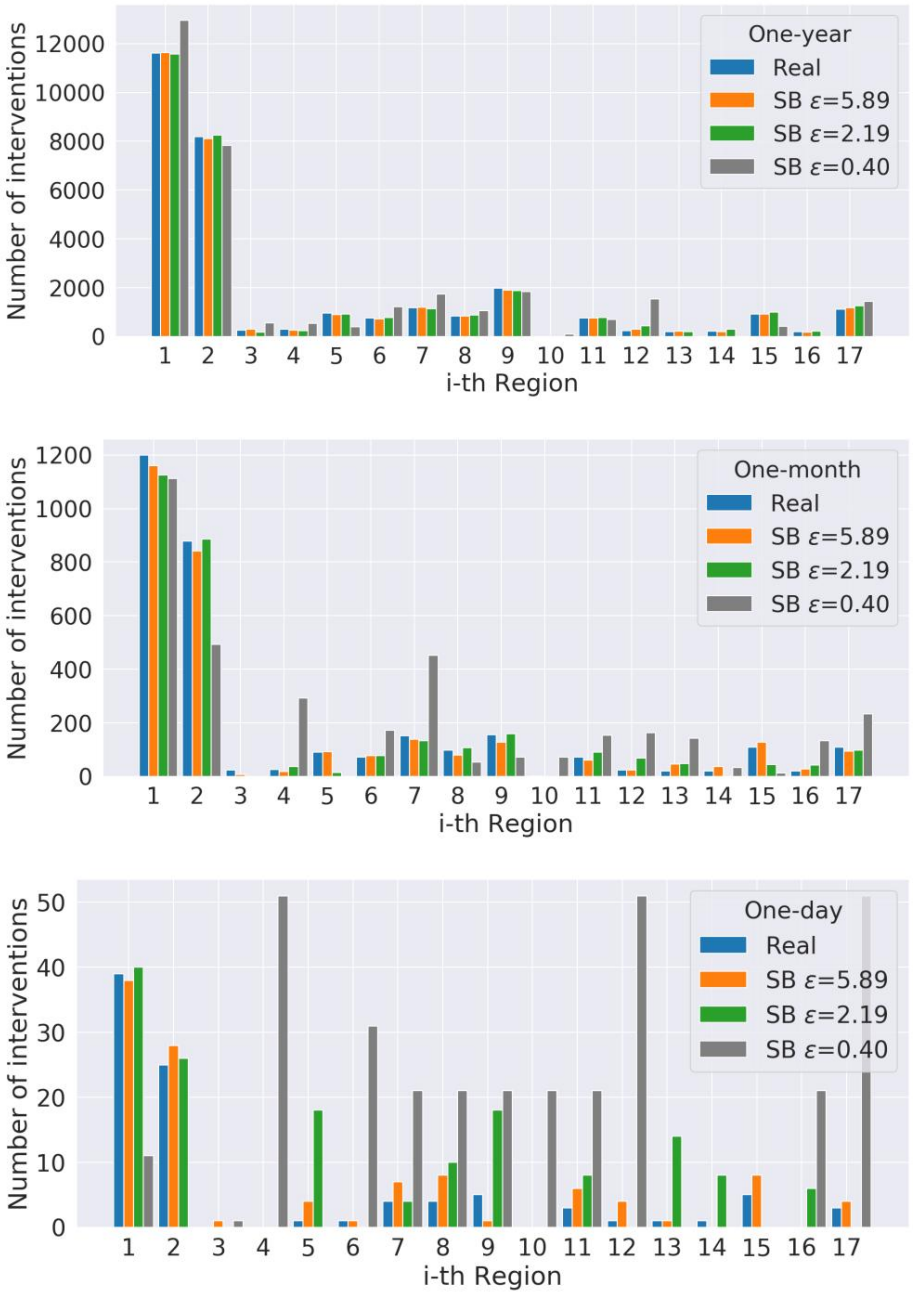

Figura 6: Análise entre o número real e estimado de intervenções por região.

objetivo de inferir o número de intervenções por região. Como o LDP garante a privacidade dos indivíduos, perturbando os dados antes de enviá-los para os dados cobrador, neste caso, o bombeiro responsável por relatar as intervenções aplicará a perturbação para o local das intervenções antes de enviá-lo para os dados servidor (como a Figura 3 ilustra).

Vale ressaltar que o ER diminui à medida que o tamanho dos dados aumenta. Isso é devido à configuração LDP, que requer uma grande quantidade de dados para garantir uma bom equilíbrio de ruído. Por exemplo, para uma análise de um ano, o número de intervenções é de pelo menos 17.333 em 2006, enquanto a média por dia é de apenas 47 para o mesmo ano. Por esta razão, a utilidade dos dados diminui para pequenas cenários como casos de um mês e um dia apresentados neste artigo. Conseqüentemente, é preciso equilibrar a aplicação dos dados anonimizados. Por exemplo, se um pretende adquirir estatísticas por ano, os resultados são muito precisos com boa privacidade garantias. No entanto, se alguém pretende aplicar tarefas de aprendizado de máquina a esses dados (conforme apresentado na próxima seção), um cenário de um dia é mais apropriado, mas com erro maior.

Além disso, a relação entre ER e garantias de privacidade é natural, Considerando que a garantia de privacidade ÿ-diferencial é aprimorada, mais ruído é adicionado aos dados e sua utilidade diminui. No entanto, como já mencionado, os dados o tamanho influencia muito nessa etapa. Pode-se ver na Tabela 3 que enquanto o 'ER. Av.' para a análise de um ano é de cerca de 1eÿ3 para os dois primeiros níveis de ÿ-Garantia DP, não é o caso do cenário de um dia com a mesma métrica entre 1e ÿ 2 e 4e ÿ 2. O cenário de um mês é o cenário médio-baixo com pontos de dados razoáveis, mas não suficientes como o caso de um ano para fornecer bons resultados. Neste caso, métricas razoáveis são adquiridas em comparação com o esquema de um dia.

A Figura 5 resume ambas as relações de ER com tamanho de dados conforme o nível de garantia de privacidade ÿ-diferencial aumenta (menor fornece fortes garantias de privacidade). Enquanto o cenário de um ano com o nível máximo de privacidade garantias tem ER em torno de 0,02, o cenário de um mês atinge esse ER para ambos os últimos valores de garantias de privacidade e o cenário de um dia atinge este ER

já com a segunda menor garantia de privacidade. Além disso, na Figura 5, um
pode ver as estatísticas adquiridas sobre o número de intervenções para cada período,
onde pequenos erros são adquiridos para o caso de um ano e consideráveis para
esquemas de um mês e de um dia.

Portanto, como também destacado na literatura, a escolha de depende
vários fatores (tamanho dos dados, o domínio do aplicativo) e é preciso
equilibrá-lo considerando a privacidade dos usuários e a utilidade dos dados. No nosso caso, como 608
cidades foram generalizadas para n = 17 regiões, a privacidade pode ser ligeiramente diminuída para
adquirir boa utilidade para gerar estatísticas (por exemplo, com $_{ÿ}$ = 5,89 conforme apresentado
na Figura 6). Na literatura, valores comuns para estão na faixa de 0,01 ÿ 10 [17]. No artigo original
do RAPPOR [9], os autores experimentaram (f, q, p) = (0, 0,75, 0,5) para dados não longitudinais
(enviados apenas uma vez), o que garante 1 1,09; (f, q, p) = (0,75, 0,75, 0,5), o que garante ÿ =
2,05 e coleta de páginas iniciais do Google Chrome (com aproximadamente 14 milhões de $_{1}$ = 0,53 para
e (f, q, p) = (0,5, 0,75, 0,5), o que garante ÿ = 4,39 e $_{1}$ = 1,07.
Em [11], os autores usaram (f, q, p) = (0,2, 0,75, 0,25), o que fornece ÿ = 4,39 e
$_{1}$ = 1,69 para coletar posições internas usando dados reais.

7. Previsão das Intervenções dos Bombeiros por Região

O objetivo desta tarefa é implementar um aprendizado de máquina de última geração
, ou seja, o aumento de gradiente extremo (XGBoost), para prever o
número de intervenções por dia das n = 17 regiões em Doubs-França. Enquanto o
objetivo principal, arquivos anônimos serão usados para construir modelos de interesse
de avaliar a utilidade dos dados com diferentes níveis de privacidade ÿ-diferencial
em comparação com o original.

7.1. Preparação de dados

Três fontes iniciais foram consideradas:

- Uma lista de localizações geométricas com projeção de mapa epsg:2154 para cada cidade
  pertencente ao departamento de Doubs, obtido no SDIS 25.

- Uma lista de cidades agrupadas em 17 regiões para o departamento de Doubs. O arquivo foi extraído do conjunto de dados público disponível em [16].

- Uma lista de intervenções de 2006 a 2018, compartilhada pelo SDIS 25. Foi organizado em um conjunto de dados, onde cada linha, representando um dia, compreende o número de intervenções por região. Como mostrado na seção anterior, estatísticas sobre o número de intervenções por dia podem ser obtidas com um margem de erro aceitável, que tem pontos de dados suficientes (4748).

Da primeira fonte extraíram-se os polígonos que descrevem cada localidade. Em seguida, foram agrupados por região considerando a segunda fonte. Assim, é obteve uma lista final com os novos polígonos para cada região conforme ilustrado em Figura 4.

A terceira fonte tem 10 versões: os dados reais e as outras 9 os anônimos de seguir o LDP conforme descrito na Seção 6 (onde f ÿ [0.1, 0.2, ..., 0.8, 0.9], ou seja, que garante privacidade ÿ-diferencial entre [5.89, 4.39, ..., 0.81, 0.4]) . Para ambos os tipos de conjuntos de dados, foram adicionadas informações temporais como ano, mês, dia, dia da semana, dia do ano, valores (1 para 'sim', 0 455 para 'não') para indicar anos bissextos, primeiro ou último dia do mês e primeiro ou último dia do ano como atributos.

Devido à função max(0, NBintest) , na maioria dos casos, os dados anônimos descrevem um número maior de intervenções do que o real. A fim de manter a integridade dos dados, um filtro é aplicado a cada conjunto anônimo. Como exemplo, , um conjunto de dados anônimo específico é obtido; para cada cidade nela contida, uma razão é obtido. A razão é o resultado da divisão das médias do número de incidentes ocorridos no ano anterior (2017) a partir do conjunto de dados real e do um anonimizado, de acordo com a cidade. Assim, o novo número de anonimizados intervenções em cada ponto de dados de uma cidade é o resultado da divisão novamente número de intervenções anônimas por sua respectiva proporção calculada.

Os dados são considerados sequenciais em cada conjunto de dados. O alvo é um vetor, onde cada posição e valor representam a região e o número de suas inter venções respectivamente, para a próxima hora (t+ 1) de uma amostra presente (t). Um presente

amostra é composta pelo número atual de intervenções em cada região e
as variáveis temporais naquele momento. Como o banco de dados fornecido pelo SDIS25 possui
informações sobre intervenções atendidas de 2006 a 2018, modelos são treinados
usando os anos 2006-2017 e testado em 2018.

7.2. Modelagem

Para fazer uma multi-previsão do número de intervenções por região,
a regressão multi-alvo é usada para resolver esta tarefa. Assim, o “MultiOut
putRegressor” da biblioteca scikit-learn [18] é aplicado. Nesse sentido, um regressor por alvo
(região) é ajustado usando o regressor XGBoost com o
parâmetro objetivo = 'count : poisson e o resto como padrão.

Seis modelos foram construídos. Dois modelos treinados com os dados reais: um como
linha de base que descreve o número médio de intervenções em cada dia do
semana por região; e um segundo construído com XGBoost que prevê o número
de intervenções por região durante um dia inteiro. Além disso, foram construídos quatro modelos
com dados anônimos considerando diferentes níveis de garantias de privacidade usando
XGBoost também.

A suposição feita aqui é: o corpo de bombeiros divulga o anonimato
dados e as informações de relação (“filtro”) do último ano para empresas de terceiros
nies e instituições acadêmicas para construir modelos apropriados para o sistema real.
Assim, para avaliar a eficácia dos modelos, todos eles são testados usando o
Dados reais de 2018.

7.3. Resultados

Os modelos são avaliados com as métricas Root Mean Square Error (RMSE) e Mean
Absolute Error (MAE). Além disso, como é uma saída múltipla
cenário, as pontuações para cada meta são calculadas com uma ponderação uniforme
média sobre saídas ('média uniforme') [18].

Por uma questão de brevidade, considerando apenas quatro valores para $_{y}$ =
[4,39, 2,77, 1,69, 0,81] (resp. f = [0,2, 0,4, 0,6, 0,8]), a Tabela 4 apresenta métricas
resultados para uma previsão de linha de base, para modelos treinados com os dados originais e

para modelos treinados com dados anônimos. Para conjuntos de dados anônimos, os resultados são apresentados para ambos os casos em que a 'razão' é usada para normalizar o número de 500 intervenções por região de acordo com o ano de 2017 ou não.

Modelo	Razão normalizada		Razão não normalizada	
	MAE	RMSE	MAE	RMSE
Linha de base (média)	-	-	2.5556	3.3237
Original	-	-	1,8552	2.5821
f = 0,20 ÿ = 4,39	1,8666	2.5963	2.1748	2.8822
f = 0,40 ÿ = 2,77	1,9271	2.7194	2.7436	3.6736
f = 0,60 ÿ = 1,69	1,9151	2,6848	4,2475	4.9567
f = 0,80 ÿ = 0,81	1,9403	2.7002	7.8542	8.4985

Tabela 4: Resultados da métrica para prever o número de intervenções por região em cada dia de 2018 usando dados originais e anonimizados normalizados e não normalizados.

Além disso, as Figuras 7 e 8 ilustram melhor os resultados da Tabela 4 em em relação às métricas RMSE e MAE com o parâmetro f variando de f = 0,1 a f = 0,9. Na Figura 9, os melhores resultados de previsão são ilustrados para cada região comparando o número original de intervenções com modelos treinados 505 com os dados brutos e anonimizados (f = 0,60; ÿ = 1,69) para um único dia de março de 2018.

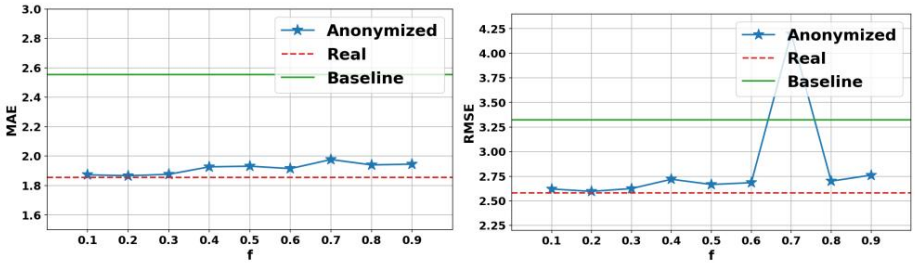

Figura 7: Métricas MAE e RMSE para os modelos de predição normalizados.

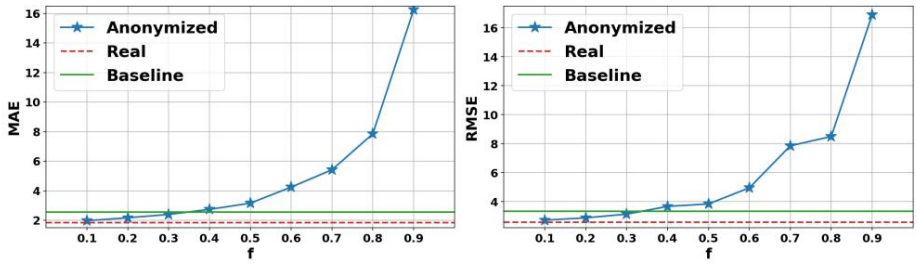

Figura 8: Métricas MAE e RMSE para os modelos de previsão não normalizados.

7.4. Discussão

Com o objetivo de avaliar o trade-off privacidade-utilidade dada a implementação de um mecanismo de privacidade diferencial local para coletar dados de intervenções , esta pesquisa implementa um algoritmo de aprendizado de máquina para prever o número de intervenções por região. Em comparação com a literatura, este trabalho introduz um modelo de previsão para várias regiões, em vez de apenas o total número de intervenções por período, o que é uma tarefa mais difícil. Além disso, é notável a melhora da pontuação com os modelos treinados para tal tarefa complexa em vez de desenvolver um modelo de previsão simples como linha de base (média) assumida neste artigo.

Como se pode notar na Tabela 4 e nas Figuras 7 e 8, os modelos treinados com dados anonimizados e normalizados também podem garantir uma boa utilidade dos dados para fins de previsão. Vale ressaltar o uso de um 'filtro' para normalizar a número de intervenções por região e dia, sendo neste caso a previsão o desempenho não diminuiu muito em comparação com o modelo treinado com os dados brutos. Em contraste, para conjuntos de dados não normalizados, os resultados diminuem muito rápido à medida que a garantia de privacidade é aplicada e, após f = 0,4, MAE e As métricas RMSE são piores do que o modelo de linha de base (média).

Os números em negrito na Tabela 4 representam os resultados das métricas usando o conjunto de dados anônimo que tem a melhor relação privacidade-utilidade. Ainda que melhores resultados foram encontrados com f = [0,1, 0,2, 0,3], ÿ = [5,89, 4,39, 3,46] (como se pode ver na Figura 7), suas garantias de privacidade são muito baixas considerando um real

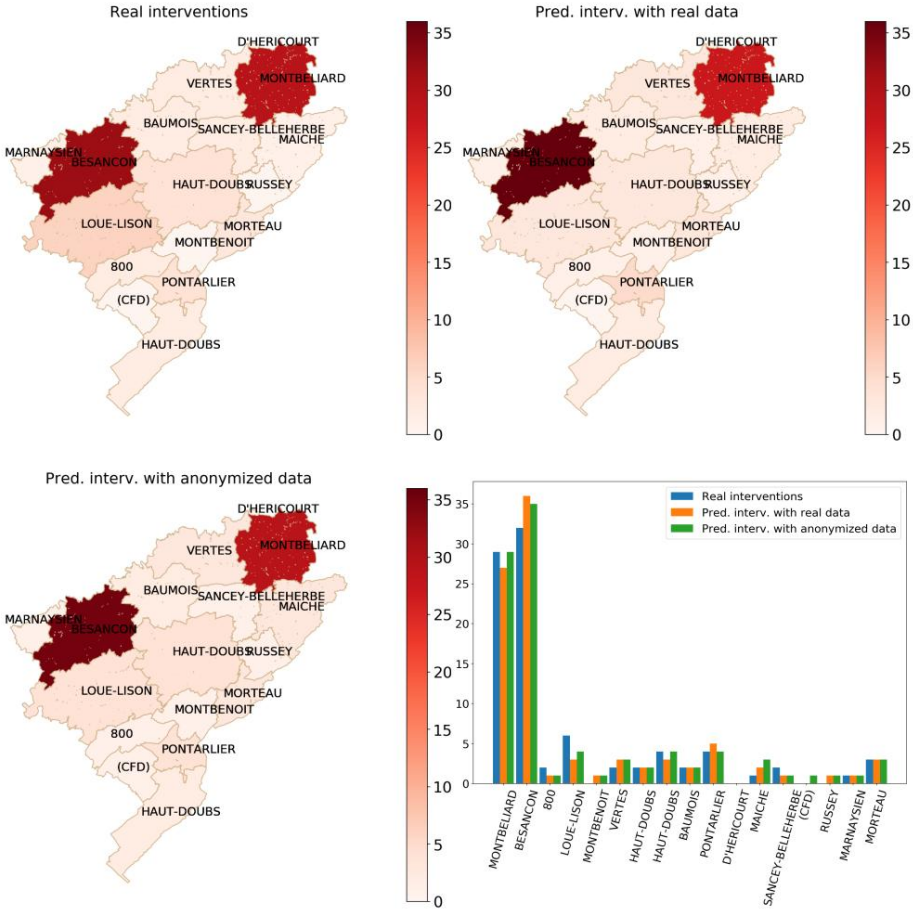

Figura 9: Comparação do número real e previsto de intervenções por região para um único dia.

aplicação mundial. Além disso, em nossa análise, resultados ainda melhores foram encontrados com f = 0,05 ef = 0,15; no entanto, ambos têm garantia de privacidade ainda menor camisetas com $_{ý}$ = 7,33 e $_{ý}$ = 5,02 respectivamente (quanto maior ÿ representa mais vazamento de informações na teoria de DP).

Assim, na Figura 9, é mostrado para um determinado dia de março de 2018 a comparação do número real e previsto de intervenções por região usando os dados brutos e melhor versão de dados anônimos (f = 0,60, ÿ = 1,69). Com tal resultado, a previsão do número de intervenções por região para o dia seguinte, o

brigada de incêndio pode se preparar com eficiência para curto, médio e longo prazo
cenários. Em particular, sabendo que certas regiões são mais propensas a
aconteçam incidentes, o corpo de bombeiros pode alocar melhor as pessoas e máquinas
recursos, bem como o planejamento da construção de novos quartéis.

8. Conclusão

A privacidade diferencial local é uma abordagem de última geração usada para proteger um
privacidade do indivíduo no processo de coleta de dados. Ao invés de confiar em um
curador de dados para ter os dados brutos e anonimizá-los para consultas de saída (como a abordagem
geral de privacidade diferencial 545), o LDP permite que os usuários anonimizem seus próprios
dados antes de enviá-los para o servidor coletor de dados.

Neste artigo, a aplicação de um mecanismo LDP para preservar a privacidade
é introduzida a recolha de dados para fins de localização das intervenções dos bombeiros. Como
mostrado nos resultados da Seção 6, o mecanismo 'Basic One-Time RAPPOR'
pode adquirir estatísticas adequadamente com um bom nível de garantias de privacidade. No
Nesse caso, um invasor não consegue distinguir entre os valores v1 ou v2 (denominados B como os
locais reais das intervenções), porque ambos têm aproximadamente o mesmo
probabilidade de gerar a saída com ruído (U).

Além disso, conforme mostrado na Seção 7, é possível prever o número futuro
ber de intervenções por região com dados anônimos, bem como com o bruto
dados. Mais especificamente, o trabalho deste artigo mostra que fluxos de dados como
transporte de saúde de emergência, que é sensível no início, mas pode ser muito
úteis, podem ser adequadamente anonimizados para evitar vazamento de informações, enquanto
permanecendo útil para fins de otimização. Eles podem ser usados para desenvolver previsões
ferramentas úteis, e essas ferramentas podem ser usadas para muitas coisas. Previsões de curto prazo
permitiria otimizar os turnos para a próxima semana, antecipar por
fornecer reforço de emergência durante picos e veículos de pré-posição. No
a médio prazo, essas previsões permitiriam redistribuir temporada
aliar os recursos materiais e humanos aos quartéis existentes, bem como auxiliar
no planejamento de férias, dada a carga de trabalho esperada nos próximos meses. Fi

finalmente, a longo prazo, tais previsões, possibilitadas por tal aprendizado de dados anonimizados, permitiria antecipar as necessidades futuras (humanas e materiais) necessárias para manter uma certa qualidade de serviço, ajudando ao mesmo tempo escolher a localização geográfica dos futuros quartéis.

Para trabalhos futuros, são planejadas melhorias no modelo multi-previsão. Por exemplo, serão adicionadas ao conjunto de dados mais variáveis explicativas, como dados meteorológicos e de tráfego, onde serão utilizadas técnicas de seleção de características para melhorar o desempenho dos modelos. Além disso, técnicas para afinar o hiperparâmetros dos modelos serão implementados.

Reconhecimento

Este trabalho foi apoiado pela Região de Bourgogne Franche-Comt´e CADRAN Projeto, pela EIPHI-BFC Graduate School (contrato “ANR-17-EURE-0002”), pelo projeto Interreg RESponSE e pela brigada de bombeiros SDIS25.

Referências

Referências

[1] TT Dang, Y. Cheng, J. Mann, K. Hawick, Q. Li, previsão de risco de incêndio usando dados de várias fontes: um estudo de caso na área de humberside, em: 2019 25th Conferência Internacional de Automação e Computação (ICAC), 2019, pp. 1–6. doi:10.23919/IConAC.2019.8894971.

[2] S. Cerna, C. Guyeux, HH Arcolezi, ADP Lotufo, R. Couturier, G. Royer, Memória de longo prazo para prever intervenções de bombeiros, em: 6ª Conferência Internacional sobre Controle, Decisão e Tecnologia da Informação nologies (CoDIT 2019), Paris, França, 2019. URL: https://doi.org/10.1109/codit.2019.8820671. doi:10.1109/codit.2019.8820671.

[3] C. Guyeux, J.-M. Nicod, C. Varnier, ZA Masry, N. Zerhouny, N. Omri, G. Royer, previsão de bombeiros usando redes neurais: um estudo de caso real,

em: Avanços em Sistemas Inteligentes e Computação, Springer Interna
Publicação Nacional, 2019, pp. 541–552. URL: https://doi.org/10.1007/
978-3-030-29516-5_42. doi:10.1007/978-3-030-29516-5_42.

[4] Satistiques mensuelles fournies par le service d´epartemental d'incendies et de
secours (sdis 71), https://www.data.gouv.fr/fr/datasets/
intervenções-des-pompiers-od71/, 2013. Acesso: 2019-12-13.

[5] Donn´ees hebdomadaires sur les intervenciones des sapeurs
Pompiers de l'essonne, https://www.data.gouv.fr/fr/datasets/
intervenções-des-pompiers/, 2018. Acesso: 2019-12-13.

[6] L. Sweeney, k-anonimato: Um modelo para proteger a privacidade, International
Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10 (2002)
557–570.

[7] C. Dwork, A. Roth, et al., Os fundamentos algorítmicos do diferencial
privacidade, Fundamentos e TendênciasR em Ciência da Computação Teórica 9
(2014) 211–407.

[8] J.-F. Couchot, C. Guyeux, G. Royer, anonimamente prevendo o número
e natureza das operações de combate a incêndio, in: Anais da 23ª Interna
Simpósio Nacional de Aplicações de Banco de Dados e Engenharia - IDEAS19,
ACM Press, 2019. URL: https://doi.org/10.1145/3331076.3331085.
doi:10.1145/3331076.3331085.

[9] U. Erlingsson, V. Pihur, A. Korolova, Rappor: Agregação aleatória
resposta ordinal preservando a privacidade, em: Proceedings of the 2014 ACM
Conferência SIGSAC sobre Segurança de Computadores e Comunicações, CCS '14,
ACM, Nova York, NY, EUA, 2014, pp. 1054–1067. URL: http://doi.acm.
org/10.1145/2660267.2660348. doi:10.1145/2660267.2660348.

[10] A. Equipe de Privacidade Diferenciada da Apple, Learning with privacy at scale, 2017.

[11] JW Kim, D.-H. Kim, B. Jang, Aplicação de privacidade diferencial local para coleta de
dados de posicionamento interno, IEEE Access 6 (2018) 4276–4286.

[12] MS Alvim, K. Chatzikokolakis, C. Palamidessi, A. Pazii, Privacidade diferencial local baseada em métricas para aplicações estatísticas, CoRR abs/1805.01456 (2018).

[13] T. Wang, N. Li, S. Jha, Mineração frequente de conjuntos de itens privados localmente diferenciados, em: 2018 IEEE Symposium on Security and Privacy (SP), IEEE, 2018. URL: https://doi.org/10.1109/sp.2018.00035. doi:10.1109/sp.2018.00035.

[14] C. Dwork, F. McSherry, K. Nissim, AD Smith, Calibrando ruído para sensibilidade na análise de dados privados, J. Priv. Confidencialidade 7 (2016) 17–51.

[15] SP Kasiviswanathan, HK Lee, K. Nissim, S. Raskhodnikova, A. Smith, O que podemos aprender em particular?, em: 2008 49º Simpósio Anual do IEEE sobre Fundamentos da Ciência da Computação, IEEE, 2008. URL: https://doi.org/10.1109/focs.2008.27. doi:10.1109/focs.2008.27.

[16] Lista e composição 2018, https://www.collectivites-locales.gouv.fr/liste-et-composition-2018/, 2018. Acesso: 2019-12-01.

[17] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, BC Pierce, A. Roth, Privacidade diferencial: um método econômico para escolher ep silon, em: Proceedings of the 2014 IEEE 27th Computer Security Foun Dations Symposium, CSF '14, IEEE Computer Society, Washington, DC, EUA, 2014, pp. 398–410. URL: https://doi.org/10.1109/CSF.2014.35. doi:10.1109/CSF.2014.35.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Pas sos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Aprendizado de máquina em Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

Apêndice A. Valor do nível de privacidade diferencial ÿ-local.

Vamos provar que o algoritmo A de RAPPOR único verifica a diferença -local privacidade encial com igual a $_{ý}$ conforme definido em (5).

Vamos então encontrar um limite $\frac{\Pr[A(v1) ÿ R]}{\text{de } \Pr[A(v2) ÿ R]}$, para todos os possíveis pares de usuários 650 dados privados v1 e v2 e todos os subconjuntos R de im(A):

$$\frac{\Pr[A(v1) ÿ R]}{\Pr[A(v2) ÿ R]} \quad \max_{UÿR} \frac{\Pr[A(v1) = U] ÿ}{\Pr[A(v2) = U]}$$

$$= \text{máximo}_{UÿR} \frac{\Pr[B1 = U]}{\Pr[B2 = U]}$$

$$= \text{máx.}_{UÿR} \frac{\prod_{k=1}^{n} P[B_k^1 = \text{Reino Unido}]}{\prod_{k=1}^{n} P[B_k^2 = \text{Reino Unido}]}$$

Graças à Equação (4), é fácil estabelecer que P(Uk = Bk) = 1 ÿ $\frac{f}{2}$ e que P(Uk = Bk) = $\frac{f}{2}$ para qualquer k, 1 ÿ k ÿ n. Temos assim

$$\frac{\prod_{k=1}^{n} P[B_k^1 = \text{Reino Unido}]}{\prod_{k=1}^{n} P[B_k^2 = \text{Reino Unido}]} = \frac{\prod_{k=1}^{n} 1\text{-}\frac{f}{2}^{|\text{Reino Unido}ÿB1_k|}\frac{f}{2}^{1ÿ|UkÿB1|_k}}{\prod_{k=1}^{n} 1\text{-}\frac{f}{2}^{|\text{Reino Unido}ÿB2_k|}\frac{f}{2}^{1ÿ|UkÿB2|_k}}$$

$$= \prod_{k=1}^{n} \frac{ÿ 1}{\frac{2}{f} ÿ 1}^{|\text{Reino Unido}ÿB1_k|}$$

$$= \prod_{k=1}^{n} \frac{2}{ÿ} ÿ 1^{|\text{Reino Unido}ÿB1_k ÿ |\text{Reino Unido}ÿB2_k|}.$$

Para qualquer f, 0 ÿ f ÿ 1 o número $\frac{2}{f}$ ÿ 1 é maior ou igual a 1. Estamos então à esquerda para encontrar três vetores booleanos de comprimento n $B1, B2$, e U que maximizam 655 |U ÿ B1 | ÿ |U ÿ B2 |, ou seja, que maximiza |U ÿ B1 | minimizando |U ÿ B2 | Sem perda de generalidade, podemos considerar que B1 = (1, . . . , 1, 0, . . . , 0), ou seja cujos primeiros h bits são definidos com 1. O vetor U que maximiza |Uk ÿ B1 | é o $_k$

reverso de B1 , ou seja, U = (U1, . . . , Uh, Uh+1, . . . , Un) = (0, . . . , 0, 1, . . . , 1), que contém n ÿ h bits definidos com 1 .O vetor booleano de comprimento n B2 que minimiza |U ÿB2 | tem que definir seus h bits iguais a 1 nos mesmos índices que os de U. é possível se h ÿ nÿh, ou seja, h ÿ $\frac{n}{2}$, que é o caso na prática. Em outras palavras B2 = (B2 1 , . . . , B2 h , B2 B2h+1, . . . , 2h , B2 B2 2h+1, . . . , n) = (0, . . . , 0, 1, . . . , 1, 0, . . . , 0). Temos assim:

- para k, 1 ÿ k ÿ h, |Uk ÿ B1 k | ÿ |Reino Unidoÿ B2 | = 1 ÿ 0;
- para k, h + 1 ÿ k ÿ 2h, |Uk ÿ B1 k | ÿ |Reino Unidoÿ B2 | = 1 ÿ 0;
- para k, 2h ÿ k ÿ n, |Uk ÿ B1 k | ÿ |Reino Unidoÿ B2 | = 1 ÿ 1;

Portanto

$$\max_{U,B1,B2 \ddot{y} Bn} \sum_{k=1}^{n} \frac{2}{f} \ddot{y} 1 \; |\text{Reino Unido}\ddot{y}_{k}^{B1} \ddot{y} |\text{Reino Unido}\ddot{y}_{k}^{B2}| \; \ddot{y} \; 2^{2h} \ddot{y} 1 f$$

e a prova está estabelecida.