UnB - UNIVERSIDADE DE BRASÍLIA

FGA - FACULDADE GAMA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA BIOMÉDICA

USO DE BUSINNES INTELIGENCE NO MONITORAMENTO EPIDEMIOLÓGICO DA DENGUE NO DISTRITO FEDERAL.

JORGE LUIS DA SILVA LUSTOSA

ORIENTADORA: Dra. Lourdes Mattos Brasil

CO-ORIENTADOR: Dr. Marcos Takashi Obara

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA BIOMÉDICA

PUBLICAÇÃO: 103A/2018

BRASÍLIA/DF: DEZEMBRO – 2018

UNIVERSIDADE DE BRASÍLIA
FACULDADE DO GAMA
ENGENHARIA BIOMÉDICA

"USO DE BUSINNES INTELIGENCE NO MONITORAMENTO EPIDEMIOLÓGICO DA DENGUE NO DISTRITO FEDERAL"

JORGE LUIS DA SILVA LUSTOSA

DISSERTAÇÃO DE MESTRADO SUBMETIDA À FACULDADE UNB GAMA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO TÍTULO DE MESTRE EM ENGENHARIA BIOMÉDICA.

APROVADA POR:

Prof. Dr. Marcos Takashi Obara; FCE / UnB
(COORIENTADOR)

Profa. Dra. Marília Miranda Forte Gomes; FGA / UnB
(EXAMINADOR INTERNO)

Prof. Dr. Georges Daniel Amvame Nze; ENE / UnB
(EXAMINADOR EXTERNO)

Brasília, 10 de dezembro de 2018

BRASÍLIA/DF, 10 DE DEZEMBRO DE 2018.

FICHA CATALOGRÁFICA

Jorge Luis da Silva Lustosa
USO DE BUSINNES INTELIGENCE NO MONITORAMENTO EPIDEMIOLÓGICO DA DENGUE NO DISTRITO FEDERAL, [Distrito Federal] 2018.
81., 210 x 297 mm (FGA/UnB Gama, Mestre, Engenharia Biomédica, 2018). Dissertação de Mestrado - Universidade de Brasília. Faculdade Gama. Programa de Pós-Graduação em Engenharia Biomédica.

1. DENGUE	2. GEOCODIFICAÇÃO
3. BUSINESS INTELLIGENCE	4. EPIDEMIOLOGIA
I. FGA UnB Gama/ UnB.	II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

LUSTOSA, J. L. S. (2018). USO DE BUSINNES INTELIGENCE NO MONITORAMENTO EPIDEMIOLÓGICO DA DENGUE NO DISTRITO FEDERAL. Dissertação de Mestrado em Engenharia Biomédica, Publicação NO./2018, Programa de Pós-Graduação em Engenharia Biomédica, Faculdade Gama, Universidade de Brasília, Brasília, DF, 81.

CESSÃO DE DIREITOS

AUTOR: Jorge Luis da Silva Lustosa.
TÍTULO: USO DE BUSINNES INTELIGENCE NO MONITORAMENTO EPIDEMIOLÓGICO DA DENGUE NO DISTRITO FEDERAL.

GRAU: Mestre

ANO: 2018

2018
St. Leste Projeção A - Gama Leste, Brasília - DF.
CEP 72444-240 Brasília, DF – Brasil.

DEDICATÓRIA

Para minha querida Lili, com amor.

"Não estar ocupado e não existir representam a mesma coisa. Todas as pessoas são boas, exceto as ociosas. Cada um deve dar a si mesmo todo o trabalho que possa para tornar a vida suportável neste mundo. Quanto mais avanço em idade, mais sinto a necessidade do trabalho. Ele se torna pouco a pouco o maior dos prazeres e substitui as ilusões da vida".

Voltaire

AGRADECIMENTOS

Antes de mais nada agradeço a Deus, que é o verdadeiro mestre e criador de tudo. Pelo dom da vida e pela oportunidade poder colaborar e criar soluções que possam ajudar as pessoas. A minha querida esposa Liana Queiros Fontelles, pelo seu amor e compreensão ao decorrer desta jornada, sempre me apoiando e me motivando nos momentos mais difíceis desta pesquisa. Ela me fez acreditar em meu potencial e sem ela de fato nada disso seria possível.

Minha mais profunda gratidão a minha querida orientadora Professora Dra. Lourdes Mattos Brasil, por ter me acolhido e me direcionado no decorrer de todo meu mestrado, por toda sua atenção e apoio ao longo do tempo que nos conhecemos, por sua dedicação e gentileza que me inspiraram não apenas a ser um aluno melhor, mas um ser humano melhor. Aos meus queridos mestres Professor Dr. Marcos Takashi Obara e Professor Dr. Georges Daniel Amvame Nze, por suas valiosas contribuições a minha pesquisa, bem como a sua paciência e dedicação no decorrer destes anos de parceria.

Obrigado a minha família por acreditar em mim. Obrigado especialmente a minha Mãe Geni e meu Pai Augusto, por todo seu amor, que nunca mediram esforços para meu sucesso e felicidade. Tudo o que sou devo a vocês. Agradeço postumamente a minha querida avó Idália por sua doçura e iniciativa que me inspirou profundamente com sua grande abnegação no serviço ao próximo.

Obrigada a minha querida amiga Delmira Ferreira Lima e meu querido amigo Marcus Vinicius Tinoco Gonçalves Ribeiro por estarem sempre prontos a me ajudar e aconselhar, sempre escutando minhas idéias e me ajudando a enxergar que o caminho do equilíbrio e da ponderação trazem sempre os melhores resultados. Sei que sempre posso ir mais longe por contar com a valiosa amizade de vocês e tenho a convicção que fiz amigos para esta vida.

Obrigado ao mestre Márcio Lima de Medeiros, que me ensinou muito sobre *Business intelligence* e sobre gestão da melhor maneira que se pode ensinar alguém, na prática! Obrigado pela sua paciência e por ter acreditado em meu potencial. Foi muito gratificante ver o resultado dos vários projetos que criamos juntos.

Meu muito obrigado também aos meus queridos amigos Jorge Luiz Ferreira da Silva Junior e Douglas Médice Rocha por estarem ao meu lado nesta jornada do mestrado, compartilhando e criando conhecimento, pessoas com histórias de vida belíssimas. Pessoas das quais sempre me recordarei.

RESUMO

USO DE BUSINNES INTELIGENCE NO MONITORAMENTO EPIDEMIOLÓGICO DA DENGUE NO DISTRITO FEDERAL

Autor: JORGE LUIS DA SILVA LUSTOSA
Orientadora: Profa. Dra. Lourdes Mattos Brasil
Coorientador: Prof. Dr. Marcos Takashi Obara
Programa de Pós-Graduação em Engenharia Biomédica
Brasília, dezembro de 2018.

Os casos de dengue aumentam em todo o mundo e informação de qualidade é essencial para o controle e prevenção. Atualmente, é uma das doenças infecciosas que mais crescem no mundo no século XXI de acordo com a Organização Mundial de Saúde. No Brasil os casos de dengue são notificados ao ministério da saúde e registrados no Sistema Nacional de Notificação de Agravos (SINAN). Os endereços dos pacientes são registrados em texto livre e em muitos casos são registrados de maneira incorreta ou incompleta dificultando as análises espaciais. Analisar estes casos sob perspectiva espacial, permite criar estratégias de controle do vetor muito mais eficazes concentrando o esforço de combate em localidades com mais casos e maior necessidade. Para se realizar este tipo de analise é necessário transformar os endereços dos pacientes em coordenadas geográficas em um processo chamado de geocodificação. Utilizando técnicas de mineração de dados foi possível melhorar a qualidade dos registros de endereços, corrigindo abreviações e informações incorretas ou incompletas. Permitindo assim geocodificar mais de 70.000 casos de dengue, e criar uma série painéis de análise com *business intelligence*. Nesta pesquisa a taxa de geocodificação dos endereços obteve uma taxa de sucesso na geocodificação de 82%. Quando comparada a outros processos de geocodificação dos dados do SINAN esta taxa de sucesso geralmente é de 62%. Esta melhoria no resultado foi possível devido ao uso de um processo dividido em três etapas. Sendo a primeira, de melhoria da qualidade dos endereços, a segunda, de geocodificação utilizando a base nacional de CEP e a terceira, de agregação dos dados utilizando o CEP dos endereços. Após os dados georreferenciados, com o uso do software de *business intelligence* Tableau foi possível analisar os dados e compreender os casos de forma visual, identificando padrões geográficos e volumétricos que permitem decisões no combate ao vetor da dengue muito mais eficazes.

Palavras-chaves: Dengue, Geocodificação, Saúde Pública, Epidemiologia, Mineração de Dados, SINAN.

ABSTRACT

BUSINNES INTELIGENCE USE TO DENGUE EPIDEMIOLOGIC MONITORING OF BRASIL FEDERAL DISTRICT
Author: Jorge Luis da Silva Lustosa

Supervisor: Dra. Lourdes Mattos Brasil

Cosupervisor: Dr. Marcos Takashi Obara

Post-Graduation Program in Biomedical Engineering – Qualify of Master Degree

Brasília, December of 2018.

Dengue cases increase worldwide and quality information is essential for control and prevention. Currently, it is one of the fastest growing infectious diseases in the world in the 21st century according to the World Health Organization. In Brazil, cases of dengue are reported to the Ministry of Health and registered in the National System of Notification of Injuries (SINAN). Patient addresses are recorded in free text and in many cases are incorrectly or incompletely recorded, making it difficult to perform spatial analyzes. Analyzing these cases from a spatial perspective allows us to create much more effective vector control strategies by concentrating the combat effort in locations with more cases and greater need. In order to perform this type of analysis it is necessary to transform the patient's addresses into geographic coordinates in a process called geocoding. Using data mining techniques it was possible to improve the quality of address records by correcting abbreviations and incorrect or incomplete information. Thus allowing geocoding more than 70,000 cases of dengue, and create a series of analysis panels with business intelligence. In this research the geocoding rate of the addresses obtained a success rate in geocoding of 82%. When compared to other processes of geocoding the SINAN data, this success rate is generally 62%. This improvement in outcome was possible due to the use of a three step process. The first one, to improve the quality of the addresses, the second, geocoding using the national base of CEP and the third, of aggregation of the data using the CEP of the addresses. After the georeferenced data, using the business intelligence software Tableau, it was possible to analyze the data and to understand the cases in a visual way, identifying geographic and volumetric patterns that allow decisions in the fight against dengue vector much more effective.

Keywords: Dengue, Geocoding, Public health, Epidemiology, Data Mining, SINAN .

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE SÍMBOLOS, NOMENCLATURAS E ABREVIAÇÕES

AE. AEGYPTI – Aedes aegypti

ANVISA - Agência Nacional de Vigilância Sanitária

API - *Application Programming Interface*

BCE/UNB – Biblioteca Central da Universidade de Brasilia

BD – Banco de Dados

BI – *Bussiness Intelligence*

CAPES - Comissão de Aperfeiçoamento de Pessoal do Nível Superior

CEP – Código de Endereçamento Postal

CENEPI - Centro Nacional de Epidemiologia

CHIKV- Vírus da *Chikungunya*

CPqAM – Centro de Pesquisa Ageu Magalhães

DBF - *Data Base Format*

DENV – *Vírus* da Dengue

DEN 1 – Variação 1 do *Vírus* da Dengue

DEN 2 – Variação 2 do *Vírus* da Dengue

DEN 3 – Variação 3 do *Vírus* da Dengue

DEN 4 – Variação 4 do *Vírus* da Dengue

DIVAL – Diretoria de Vigilância Ambiental

DNE – Diretório Nacional de Endereços

DW – *Data Wharehouse*

DM – *Data Mart*

ETL – *Extract Transform Load*

FGV- Fundação Getúlio Vargas

FHD – Febre Hemorrágica da Dengue

FIN – Ficha Individual de Notificação

FIOCRUZ – Fundação Oswaldo Cruz

FUNASA - Fundação Nacional de Saúde

GCI - *Geocoding Certainty Indicator*

GPS – *Global Positioning System*

IBGE – Instituto Brasileiro de Geografia e Estatística.

ICEHTMC – Congresso Internacional de Engenharia Clínica e Gestão da Tecnologia da Saúde

IEEE – Instituto de Elétrica, Eletrônica e Engenharia

IOC – Instituto Oswaldo Cruz

LABOCLIMA/UFPR - Laboratório de Climatologia/Universidade Federal do Paraná

JSON – *Java Script Object Notation*

LIS – Laboratório de Informática em Saúde

MS – Ministério da Saúde

NCBI – *National Center for Biotechnology Information*

NML – *National Library of Medicine*

NMT/UNB – Núcleo de Medicina Tropical / Universidade de Brasília

OLAP – *On Line Analytical Processing*

OLTP – *On Line Transaction Processing*

OMS – Organização Mundial de Saúde

OPAS – Organização Pan-Americana de Saúde

PHP - *Personal Hypertext Preprocessor*

PLISA – Plataforma de Informação em Saúde das Américas

PNCD – Programa Nacional de Controle da Dengue

PUBMED – Publicações Médicas

RA – Região Administrativa

RL – Regressão Linear

SACdengue – Serviço de Alerta Climático de Dengue

SAPIO – Sistema de Aquisição e Processamento de Imagens de Ovitrampas

SCIELO – *Scientific Eletronic Library Online*

SE – Semana Epidemiológica

SESA/PR – Secretaria de Estado de Saúde

SIG - Sistema de Informação Georreferenciada

SIGESON-DENGUE - Sistema de Informação Geográfica e Sonoro para Ovitrampas da Dengue

SIGO-DENGUE - Sistema de Informação Geográfica para Ovitrampas da Dengue

SINAN- Sistema de Informação de Agravos de Notificação

SPDI – Sistema para o processamento Digital de Imagens

SUS-DF – Sistema Único de Saúde – Distrito Federal

STRING – Sequência de Caracteres

TIC – Tecnologia da Informação e Comunicação

UFPE – Universidade Federal de Pernambuco

UNB – Universidade de Brasília

UNB/FGA – Universidade de Brasília / Faculdade Gama

UPU – União Postal Universal

USP – Universidade de São Paulo

VizQL – *Visual Query Language*

ZIKV – Vírus da *Zika*

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO E FORMULAÇÃO DO PROBLEMA

A dengue é uma doença viral transmitida por fêmeas infectadas de mosquito da espécie *Aedes aegypti* e *Aedes albopictus*. Atualmente é uma das doenças infecciosas que mais crescem no mundo no século XXI de acordo com a Organização Mundial de Saúde (OMS) (WHO, 2012; WHO, 2009). O vírus da dengue se apresenta e quatro variedades de sorotipos diferentes, DEN-1, DEN-2, DEN-3 e DEN-4, que estão presentes na maioria dos países que reportaram casos. Dados epidemiológicos de pesquisas sobre dengue no Brasil, Caribe, Malásia, México, Filipinas e Tailândia, patrocinadas pela OMS, comprovaram o comportamento da expansão do vírus (HORSTICK; MORRISON, 2014) e os constantes surtos globais de dengue nas regiões endêmicas e não endêmicas chamam a atenção para um acompanhamento e gerenciamento mais efetivo para estes surtos.

De acordo com Runge et al. (2014) a expansão mundial dos sorotipos da dengue vem se tornando uma das doenças infecciosas que mais cresce e preocupa a saúde pública global. Sendo uma doença endêmica em 128 países, e atualmente exposta a aproximadamente 4 bilhões pessoas. A OMS estima que aproximadamente 390 milhões de pessoas sejam infectadas anualmente em todo mundo (WHO, 2009). No Brasil, a primeira evidência documentada clínica e laboratorialmente, aconteceu em 1981, em Boa Vista (RR), onde foram detectadas as variações DEN-1 e DEN-4 do vírus. Desde então, a dengue vem crescendo no país de forma sustentada e continuada (BRASIL, 2009).

Uma possível saída para este problema seria o desenvolvimento de uma vacina que pudesse imunizar todo este contingente populacional, combatendo de uma maneira mais eficaz os surtos de dengue. O laboratório *Sanofi Pasteur* vem pesquisando há alguns anos a vacina DENGVAXIA que imuniza a pessoa dos quatro sorotipos atuais da dengue (DEN-1, DEN-2, DEN-3, DEN-4). Segundo Sabchareon, em artigo publicado na revista *The Lancet*, em 2012, pesquisadores reportaram taxa de eficiência entre 60% e 90% de sucesso em vacina tetravalente testada em 4 mil crianças com idades entre 4 e 11 anos na Tailândia.

Mesmo a vacina apresentando bons resultados e sendo recomendada pela OMS, ela ainda não previne contra outros tipos de vírus que também são transmitidos por *Aedes aegypti* e *Aedes albopictus*, como os vírus da *Chikungunha* e o *Zika Vírus*. Sendo assim, a

opção mais eficaz, ainda é controlar o vetor destas doenças, que é o mosquito. Segundo Barrera (2015), este esforço deve estar basicamente concentrado em duas frentes: A redução dos criadouros do mosquito, usando-se larvicidas e, no caso dos mosquitos adultos, usando-se inseticida aplicados por meio de Ultra Baixo Volume.

O controle desta doença, tipicamente urbana, é muito complexo e envolve vários eixos, tais como: infraestrutura das cidades, meio ambiente, logística de pessoas e materiais, saúde, educação, entre outros. Em decorrência destas dificuldades e dos custos envolvidos para solucionar estes problemas, faz-se necessário e urgente, explorar novas alternativas para auxiliar o controle da dengue.

Algumas experiências têm obtido sucesso no controle do vetor. Uma delas aconteceu no Brasil, e foi conduzida pela Fundação Oswaldo Cruz do Rio Grande do Norte em parceria com a Universidade Federal de Pernambuco (UFPE). Neste caso, o uso de um Sistema Informações Georreferenciadas (SIG) e o apoio da população local e autoridades governamentais, permitiu a redução de aproximadamente 90% do índice de Ovitrampas positivas com os ovos do *Aedes aegypti* (REGIS et al., 2013).

Outro exemplo foi estudado na Tailândia, onde o envolvimento da comunidade local na eliminação de criadouros e o uso de um SIG para localização dos potenciais criadouros de dengue, que foram a chave do sucesso na supressão do *Aedes aegypti*. A população e os pesquisadores com acesso ao SIG e os dados de ovitrampas positivas conseguiram concentrar maior esforço em locais com mais casos de dengue, direcionando as pessoas e as atividades de maneira coordenada, conseguindo assim suprimir o vetor (KITTAYAPONG et al., 2008).

Por conseguinte, os exemplos acima elencados ilustram como o apoio da comunidade e o acesso a informação são as chaves do sucesso para o controle do vetor. Na Inglaterra, o *National Health System* (NHS), que é o serviço público de saúde da Inglaterra, decidiu publicar os dados de mortalidade em cirurgias cardíacas de todos os hospitais do país. A partir desta iniciativa, as pessoas passaram a escolher hospitais que tinham uma taxa de mortalidade menor. Isto causou uma revolução no sistema de saúde e acabou levando as unidades de saúde com altos índices de mortalidade à mudança em seus procedimentos, gestão, profissionais e isto melhorou os indicadores de mortalidade em cirurgias cardíacas nacionais em um curto prazo (KEOGH et al., 1998). Portanto, o envolvimento social e a

análise de dados são ferramentas, que trabalhando em conjunto, melhoram a qualidade de vida de uma comunidade.

No Brasil, os casos de dengue são registrados pelo Sistema de Informação de Agravos de Notificação (SINAN). Esse sistema foi desenvolvido no início da década de 90, tendo como objetivo a coleta e o processamento dos dados sobre agravos de notificação em todo o território nacional, fornecendo informações para a análise do perfil da morbidade e contribuindo, dessa forma, para a tomada de decisões nos níveis municipal, estadual e federal (LAGUARDIA et al., 2004). O SINAN é hoje uma das principais fontes de dados para vigilância da dengue e a notificação se baseia na comunicação de casos confirmados e suspeitos, não apenas da dengue, mas de outras doenças que ocorrem no Brasil.

Com base nos dados do SINAN são tomadas as decisões de controle epidemiológico no Brasil. Estes dados são disponibilizados pelo MS, para as secretarias de saúde estaduais que por sua vez disponibilizam boletins mensais e ou semanais sobre casos suspeitos e confirmados de dengue de cada localidade. Um dos objetivos da disponibilização destes dados para a população é que uma vez as pessoas acompanhando a quantidade de casos e o cenário da dengue em sua localidade ela possa se engajar ainda mais no combate à dengue eliminando os criadouros do mosquito. Além disso, o volume de dados gerado pelo SINAN é algo considerável, analisando apenas o ano de 2017, teve-se aproximadamente 255.655 casos suspeitos, segundo o Ministério da Saúde do Brasil. Lidar com este volume de casos é tarefa computacional complexa. Processos simples de análises, com contagens, agrupamentos estatísticos e comparações podem levar muito tempo devido ao grande volume de dados (WHO, 2009).

Segundo Segel e Heer (2010), a história é contada através de fatos e dados. Cada vez mais jornalistas e cientistas usam técnicas de visualização de dados e análises de dados para dar sentido a suas histórias e pesquisas. Os gráficos conseguem de maneira muito mais eficaz demonstrar variações e padrões que seriam impossíveis de perceber se analisando tabelas e números, permitindo que centenas de milhares de dados se transformem em conhecimento útil para as pessoas no dia a dia.

Neste contexto de um alto volume de dados para serem analisados e explorados, uma das tecnologias mais utilizadas é o *Business Inteligence*. De acordo com Primak (2008), esta tecnologia teve sua origem na Inglaterra, onde a rainha Elizabeth I ordenou que o filósofo

Francis Bacon inventasse um sistema dinâmico de informação. O *Gartner Group* na década de 80 definiu a tecnologia *Businness Intelligence* como, processo inteligente de coleta, organização, análise, compartilhamento e monitoramento de dados, gerando suporte a tomada de decisões no ambiente de negócios (PRIMAK, 2008).

1.2 OBJETIVOS

1.2.1 Objetivo geral

Este trabalho tem por objetivo principal coletar e analisar dados dos casos de dengue, contribuir para uma melhor compreensão do comportamento da dengue em Brasília, no período de 2010 a 2015.

Objetivos específicos

Os objetivos deste trabalho podem ser detalhados segundos dois aspectos ou áreas de interesse: mineração e apresentação dos dados.

Quanto à mineração de dados, este trabalho se propõe a:

- Minerar os microdados do SINAN de casos de dengue no Distrito Federal.
- Georreferenciar os dados do SINAN oriundos do processo de mineração.

Quanto à apresentação dos dados, este trabalho se propõe a:

- Criar painéis para análise dos dados SINAN, utilizando o *Business Intelligence* ,analisar os casos de dengue sob a perspectiva de diversos indicadores geográficos e epidemiológicos.

1.3 REVISÃO DA LITERATURA

Para esta pesquisa a escolha foi pela narrativa bibliográfica. Connelly e Clandinin (2000) lembram que esse tipo de pesquisa acontece na interação do passado, presente e futuro. Criando uma noção de tempo e espaço que ajuda a marcar a situação. A investigação narrativa, se dá por meio do entendimento dos indivíduos e seu contexto social. Transformando o termo “experiência” em termo de pesquisa.

A narrativa é o método de pesquisa e ao mesmo tempo o fenômeno pesquisado. O texto é modelado pelo processo de interpretação do pesquisador, do participante e da relação entre

eles e é contextualizado devido às circunstâncias particulares da situação. A análise de dados da pesquisa narrativa é realizada, geralmente, a partir da tematização dos dados. Não aplicando estratégias rebuscadas, conforme elucidado a seguir:

> "Não utiliza critérios explícitos e sistemáticos para a busca e análise crítica da literatura. A busca pelos estudos não precisa esgotar as fontes de informações. Não aplica estratégias de busca sofisticadas e exaustivas. A seleção dos estudos e a interpretação das informações podem estar sujeitas à subjetividade dos autores. É adequada para a fundamentação teórica de artigos, dissertações, teses, trabalhos de conclusão de cursos" (USP, 2018).

A base bibliográfica pesquisada e utilizada neste trabalho considerou a busca por meio de livros, teses, monografias e artigos nas seguintes fontes especializadas: *Scientific Eletronic Library Online* (Scielo), Comissão de Aperfeiçoamento de Pessoal do Nível Superior (CAPES), *Institute of Electrical and Electronics Engineers* (IEEE), Biblioteca Central da Universidade de Brasília (BCE/UNB).

Foram utilizadas palavras-chaves (*Aedes aegypti*, *Business Intelligence*, Inteligência de Negócio, *Data Mining*, mineração de dados, SINAN) para realização da pesquisa nas bases de dados eletrônicas. Dessa forma, seguem breves descrições sobre aquelas que possuem relação com o objeto de estudo desta pesquisa.

O PubMed é uma base de dados que permite a pesquisa bibliográfica de artigos publicados em revistas de grande circulação da área médica. Essa base foi desenvolvida pelo *National Center for Biotechnology Information* (NCBI), sendo mantido pela *National Library of Medicine* (NLM).

A pesquisa realizada com a palavra-chave — *Aedes aegypti* + *Business Intelligence* nas diversas bases de dados retornou 23 trabalhos. Já a pesquisa com o argumento — *SINAN* + *Business Intelligence* implicou em 0 trabalhos. — *Aedes aegypti* + *Data mining* implicou em 294 trabalhos. Outra pesquisa com o argumento — *SINAN* + *Mineração de dados* implicou em 2 trabalhos. As outras bases bibliográficas pesquisadas seguiram os mesmos parâmetros retornando em vários casos 0 trabalhos. O Quadro 1 mostra os diversos resultados da pesquisa bibliográfica nas bases de dados.

Quadro 1: Palavras-chave utilizadas para busca de artigos na língua portuguesa, inglesa.

PALAVRAS-CHAVE	CAPES	PUBMED	IEEE	UNB	SCIELO	TOTAL
Aedes aegypti + business intelligence	0	0	23	0	0	23
Aedes aegypti + Inteligência de Negócio	0	0	0	0	0	0
Aedes aegypti + Data Mining	3	8	283	0	1	294
Aedes aegypti + Mineração de Dados	3	0	0	0	0	3
SINAN + Business Intelligence	0	0	0	0	0	0
SINAN + Inteligência de Negócio	0	0	0	0	0	0
SINAN + Data Mining	0	0	0	2	0	2
SINAN + Mineração de Dados	2	0	0	0	0	2

Fonte: Autoria própria.

Após um estudo sobre os artigos que tinham maior relevância para a pesquisa, foram escolhidos onze artigos que se assemelham com o tema proposto, embora outros materiais para consulta também foram utilizados. No Quadro 2 seguem os artigos relevantes para o tema proposto.

Quadro 2: Artigos relevantes para o tema proposto.

TITULO	RESUMO	REFERÊNCIA
Plano de Contingência Nacional para Epidemias de Dengue. Diretrizes nacionais para prevenção e controle de epidemias de dengue.	Os informes do MS e o Manual de Normas e Diretrizes nacionais para prevenção e controle de epidemias de dengue, abordam as principais manifestações, sintomas, prevenção e cuidados sobre a Dengue, *Chikungunya* e *Zika*, indicando meios e instrumentos para potencializar a prevenção e o controle desses vírus nas comunidades locais	BRASIL$_2$. Plano de Contingência Nacional para Epidemias de Dengue. Brasília: Ministerial da Saúde: 2015. BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância Epidemiológica. Diretrizes nacionais para prevenção e controle de epidemias de dengue / Ministério da Saúde, Secretaria de Vigilância em Saúde, Departamento de Vigilância Epidemiológica. – Brasília: Ministério da Saúde, 2009. 160 p. – (Série A. Normas e Manuais Técnicos).
Rastreamento do foco do Aedes Aegypti utilizando	Utilizou o processamento de informações baseadas em Ovitrampas e SIG para a detecção do foco do *Aedes*	

processamento de imagens e sistema de informações geográficas no Distrito Federal.	*aegypti* utilizando processamento de imagens no Distrito Federal. Em outra dissertação de mestrado, o autor lida com os SIGs para o desenvolvimento de bibliotecas digitais geográficas distribuídas os quais podem potencializar a comunicação no enfrentamento de problemas diversos.	SILVA, M. M. Rastreamento do foco do Aedes Aegypti utilizando processamento de imagens e sistema de informações geográficas no Distrito Federal. 2013.
Geocodificação de endereços urbanos com indicação de qualidade	Endereços urbanos são uma das principais formas de expressão da localização geográfica em cidades. Muitos sistemas de informação incluem atributos para receber endereços e, assim, contam com uma referência espacial indireta. A obtenção de coordenadas a partir de endereços é um dos métodos de geocodificação mais importantes, mas é dificultada por variações comuns no endereço, como abreviações e omissão de componentes. O artigo apresenta um método de geocodificação de endereços urbanos, que reconhece fragmentos do endereço na entrada e realiza buscas em um banco de dados geográfico de referência, para retornar coordenadas. O resultado é acompanhado de um indicador de certeza geográfica, que indica a expectativa de acerto. Uma avaliação experimental do método é apresentada	DAVIS, C. A.; FONSECA, F. T. Geocodificação de endereços urbanos com indicação de qualidade. GeoInformatica, v. 11, n. 1, p. 103–129, 2007.
Projeto de análise de dados para implantação de data mart como ferramenta para tomada de decisão em combate aos vírus da dengue, zika e Chikungunya	Com o objetivo de fazer uso de ferramentas e técnicas de Business Intelligence (BI) para análise e mapeamento de bancos de dados em geral com ênfase em saúde pública, criar um sistema de apoio à decisão para combater doenças epidêmicas como Dengue, Zika e Chikungunya em Recife. Será o resultado pretendido através deste projeto. Analisando os bancos de dados públicos disponíveis na prefeitura de Recife, um Data Mart como repositório de dados e ferramenta de business intelligence será implantado para apoiar soluções de prevenção e combater o foco do mosquito Aedes Aegypti. Com isso, este projeto trata da consolidação de dados que começa com a migração de dados para sistemas informatizados, criando um histórico online que está disponível para um Sistema de Apoio à Decisão (SAD) para autoridades públicas e consultável por todos os envolvidos, cidadãos, equipe de saúde e gestores governamentais ou organizações privadas de saúde	NIVALDO MARIANO CARVALHO,DAVID GALDÊNCIO FERREIRA,MOISÉS ERICKISON BRITO DE ARAÚJO, 2017 PROJETO DE ANÁLISE DE DADOS PARA IMPLANTAÇÃO DE DATA MART COMO FERRAMENTA PARA TOMADA DE DECISÃO EM COMBATE AOS VÍRUS DA DENGUE, ZIKA E CHIKUNGUNYA. Interscientia, v. 5, p. 106–123, 2017
Nation-Wide, Web-Based, Geographic Information System for the Integrated Surveillance and Control of Dengue Fever in Mexico	A incidência da dengue e sua distribuição geográfica estão aumentando em todo o mundo. Qualidade e informação é essencial para sua prevenção e controle. Um sistema de vigilância integral da dengue, baseado na web e geograficamente habilitado (Dengue-GIS), foi desenvolvido para a coleta, integração, análise e elaboração de relatórios de dados de intervenções epidemiológicas, epidemiológicas, entomológicas e de controle.	HERNÁNDEZ-ÁVILA, J. E. et al. Nation-Wide, Web-Based, Geographic Information System for the Integrated Surveillance and Control of Dengue Fever in Mexico. PLoS ONE, v. 8, n. 8, p. e70231, 6 ago. 2013
Sustained Reduction of the Dengue Vector Population Resulting from an Integrated Control Strategy Applied in Two Brazilian Cities	O Aedes aegypti desenvolveu adaptações orientadas pela evolução para sobreviver no habitat humano doméstico. Vários modelos de armadilhas foram projetados considerando essas estratégias e testados para monitorar esse vetor eficiente da dengue. Aqui, relatamos uma avaliação em escala real de um sistema de monitoramento e controle de populações de mosquitos com base na amostragem de ovos, juntamente com a tecnologia de sistemas de informação geográfica. O SMCP-Aedes, sistema baseado em tecnologia aberta e padrões de dados abertos, foi criado de março / 2008 a outubro / 2011 como um teste piloto em duas cidades de Pernambuco - Brasil: Ipojuca (10.000 residentes) e Santa Cruz (83.000), num esforço conjunto das autoridades e funcionários de saúde, e uma rede de cientistas fornecendo apoio científico.	REGIS, L. N. et al. Sustained Reduction of the Dengue Vector Population Resulting from an Integrated Control Strategy Applied in Two Brazilian Cities. PLoS ONE, v. 8, n. 7, 2013
	Em 1977, Sir Terence English estabeleceu o registro cirúrgico cardíaco do Reino Unido, que coleta dados de atividade e mortalidade de todos os procedimentos cirúrgicos cardíacos realizados em cada unidade de cirurgia cardíaca do NHS, totalizando 35.000 procedimentos por ano. Embora aparentemente simples em conceito, o processo representou a primeira tentativa na Grã-Bretanha, de coletar dados de resultados e atividades nacionais. Criando um indicador de desempenho da taxa de mortalidade em Cirurgias	KEOGH, B. E. et al. Public confidence and cardiac surgical outcome. BMJ, v. 316, p. 1759–1760, 1998

Public confidence and cardiac surgical outcome	cardíacas por hospital, os pacientes passaram a realizar procedimentos preferencialmente em hospitais com baixos índices de mortalidade.	
Suppression of dengue transmission by application of integrated vector control strategies at sero-positive GIS-based foci	Uma pesquisa sorológica realizada em crianças de escolas primárias de seis escolas na província de Chachoengsao, Tailândia, no final de um surto de transmissão da dengue. A análise geográfica dos casos soropositivos foi realizada para determinar os focos de transmissão. A implementação do controle de vetores foi realizada nos focos e também num raio de 100 metros em torno dos focos nas áreas tratadas contando com a participação da comunidade em colaboração com o governo local. As estratégias de controle de vetores incluíram a redução de fontes junto com o uso de coberturas de tela, uma combinação de Bacillus thuringiensis subsp. Termociclopoides de Israelensis e de Mesocyclops, e ovitrampas letais.	KITTAYAPONG, P. et al. Suppression of dengue transmission by application of integrated vector control strategies at sero-positive GIS-based foci. American Journal of Tropical Medicine and Hygiene, v. 78, n. 1, p. 70–76, 2008
Big data for infectious disease surveillance and modeling	Considerando uma ampla definição de big data para a saúde pública, e abrangendo informações de pacientes reunidas de registros de saúde eletrônicos de alto volume e sistemas de vigilância participativa, bem como a mineração de traços digitais, como mídias sociais, pesquisas na Internet e registros de telefones celulares. Destacam-se várias áreas transversais que exigem mais pesquisas, incluindo representatividade, vieses, volatilidade e validação, e a necessidade de análises robustas, estatísticas e baseadas em hipóteses. Otimistas de que a revolução da tecnologia do bigdata melhorará enormemente a granularidade e a agilidade das informações epidemiológicas disponíveis, com sistemas híbridos aumentando em vez de suplantando os sistemas de vigilância tradicionais e melhores perspectivas de modelos e previsões precisas de doenças infecciosas.	BANSAL, S. et al. Big data for infectious disease surveillance and modeling. Journal of Infectious Diseases, v. 214, n. November, p. S375–S379, 2016
Avaliação do dado sobre endereço no Sistema de Informação de Agravos de Notificação utilizando georreferenciamento em nível local de casos de tuberculose por dois métodos no município do Rio de Janeiro	O presente trabalho objetivou avaliar a qualidade dos dados referentes ao endereço dos casos notificados de tuberculose (TB) ao Sistema de Informação de Agravos de Notificação (SINAN), de 2005 a 2008, no município do Rio de Janeiro. Criou-se um indicador a partir da razão entre a taxa de incidência da TB calculada a partir do bairro de moradia declarado pelo paciente e a taxa de incidência do bairro obtida após o georreferenciamento do endereço do mesmo indivíduo, com o intuito de medir possíveis impactos que inconsistências no endereço podem causar nos cálculos de indicadores da doença por bairro. Foram utilizadas duas técnicas de georreferenciamento de dados a partir do endereço de residência. Foi realizada uma correção no campo endereço que modificou 27% dos registros. Houve uma diferença no resultado do georreferenciamento entre os dois métodos de 64 e 69%.	MAGALHÃES, M. DE A. F. M.; MATOS, V. P. DE; MEDRONHO, R. DE A. Avaliação do dado sobre endereço no Sistema de Informação de Agravos de Notificação utilizando georreferenciamento em nível local de casos de tuberculose por dois métodos no município do Rio de Janeiro. Cadernos Saúde Coletiva, v. 22, n. 2, p. 192–199, 2014

Durante a pesquisa, pode-se observar que o uso de *Business Intelligence* ainda é algo pouco difundido no contexto do estudo epidemiológico. Levando em consideração os poucos artigos encontrados, tanto na área de *Business Intelligence* aplicado ao estudo dos casos de dengue bem como a mineração dos dados do SINAN para os casos da dengue. Justifica-se a pesquisa do tema proposto, uma vez que o *Business Intelligence* é uma ferramenta que permite analisar e detectar padrões em grandes quantidades de dados e o SINAN é a única fonte de dados oficial dos casos de dengue no Brasil.

1.4 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado em seis capítulos, incluindo este capítulo.

No capítulo dois, é apresentada uma visão geral do referencial teórico, objetivando a compreensão das tecnologias, conceitos, e padrões utilizados na área da saúde e tecnologia para o apoio à tomada de decisão e mineração de dados. Logo, são abordados os seguintes temas: 1. *Aedes aegypti* e *Aedes albopictus*, 2. Sistema de Informação de Agravos de Notificação, 3. Mineração de dados, 4. Business Intelligence, 5. Código de endereçamento postal brasileiro, 6. Sistema de informação geográfico, 7. Sigeson-Dengue, 8. AedesMaps, 9. Tableau.

O capítulo três detalha a metodologia utilizada no estudo.

O capítulo quatro descreve os resultados obtidos através do processo de mineração de dados e os painéis utilizando técnicas de *business intelligence* construídos a partir dos dados minerados e georreferenciados do SINAN.

O capítulo cinco discute os pontos de maior importância envolvendo o tema deste estudo e apresenta as conclusões finais do trabalho.

Por fim, o capítulo seis apresenta os trabalhos futuros que podem ser desenvolvidos a partir das ideias apresentadas neste documento.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 AEDES AEGYPTI E AEDES ALBOPICTUS

2.1.1 Ciclo Biológico

A dengue é transmitida por mosquitos do gênero *Aedes*, sendo o *Aedes aegypti* seu principal vetor. O mosquito é encontrado, principalmente, no meio urbano, colonizado em depósitos de armazenamento de água. A infestação do mosquito torna-se mais intensa durante o verão, decorrente do aumento da temperatura e intensificação das chuvas. Sua presença ocorre principalmente em locais com maior densidade populacional. Nesse cenário, as fêmeas do mosquito *Ae. aegypti* encontram alimento e diversos criadouros para o seu desenvolvimento biológico (BRITO et al., 2016).

O *Aedes aegypti* é a espécie mais conhecida, entretanto *Ae. albopictus* também pode transmitir os vírus. As duas espécies são muito semelhantes: *Ae. Aegypti* tem listas brancas no dorso que se assemelham a uma lira, enquanto o *Ae. albopictus* é bem mais escuro e possui um risco longitudinal no local. A Figura 1 mostra os aspectos físicos das duas espécies. À esquerda (*Ae. aegypti*) e à direita (*Ae. albopictus*).

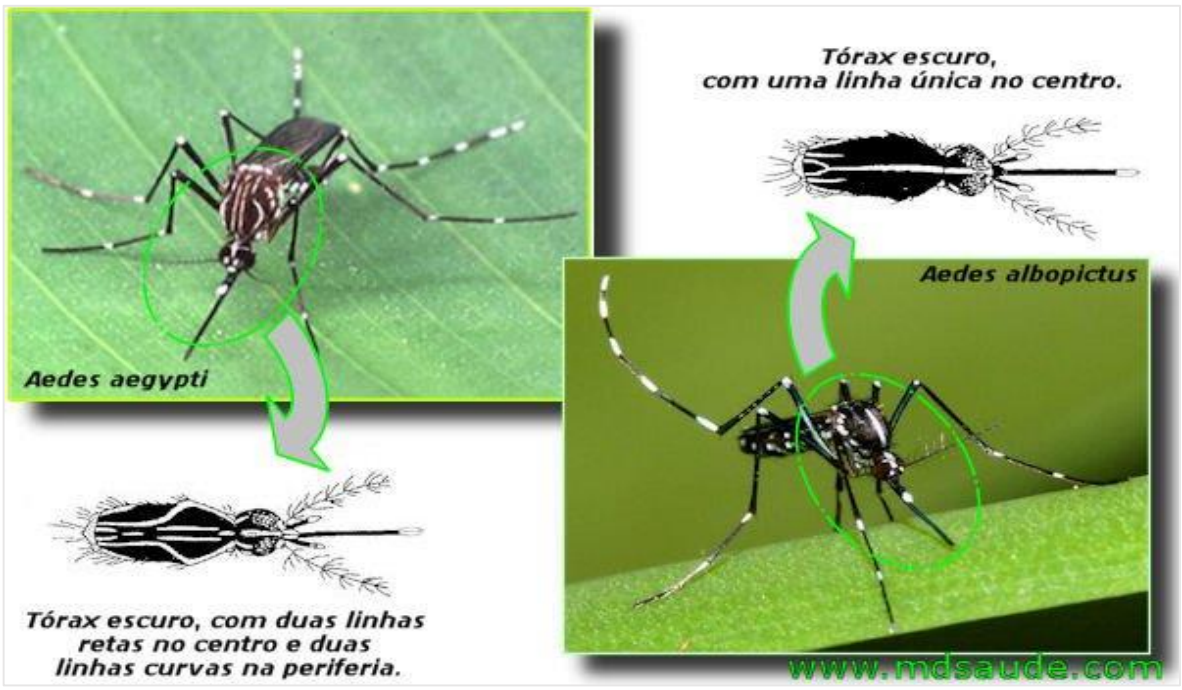


Figura 1: Mosquito da dengue *(Ae. aegypti e Ae. albopictus)* – (MDSaude, 2012).

Ao se analisar seus hábitos, podemos observar que se diferenciam em alguns aspectos como: *Ae. aegypti* geralmente vive no interior das residências e dos imóveis, próximo ao ser humano. Já *Ae. albopictus* vive em ambientes com vegetação, tais como praças, parques e matas (BRASIL, 2009).

Ae. aegypti geralmente tem o tamanho menor que 1 cm de diâmetro. Apresenta vôo baixo, geralmente abaixo de meio metro, pica preferencialmente os pés, tornozelos e as pernas. Na natureza a sua alimentação geralmente é baseada em néctar e seivas, apenas a fêmea é capaz de realizar a hematofagia – capacidade de picar o homem para sugar o seu sangue. Três dias depois de realizar a hematofagia as fêmeas estão prontas para por seus ovos (BRITO et al., 2016). O ciclo de vida do *Ae. aegypti* pode variar com a alimentação, temperatura e quantidade de larvas no criadouro. Seu ciclo ocorre em duas fases: uma aquática (ovo, larva, pupa) e outra terrestre (adulto). Os ovos inicialmente se apresentam na coloração branca e com o amadurecimento se tornam escuros. Há relatos que o ovo pode resistir até a 450 dias, permitindo sobreviverem a ambientes secos ou chuvosos até sua eclosão. O desenvolvimento do embrião até a fase madura leva em média 48 horas. Após a eclosão do ovo, o desenvolvimento até a vida adulta do mosquito pode chegar a um período de 10 dias. Uma fêmea pode originar até 1500 mosquitos durante sua vida que podem estar infectados pelo vírus da dengue, em um processo chamado de transmissão vertical. Os ovos são depositados próximo à superfície da água limpa e parada. O *Ae. aegypti* possui preferência por depósitos artificiais, sendo um fator determinante para sua crescente proliferação nos centros urbanos das regiões tropicais e subtropicais do planeta (BESERRA et al., 2009).

2.1.2 Distribuição Geográfica no Mundo

Originário do Egito, na África, vem se espalhando de maneira gradativa pelas regiões tropicais e subtropicais do planeta desde o século 16. Descrito cientificamente pela primeira vez em 1762, quando foi denominado *Culex aegypti*. O nome definitivo – *Aedes aegypti* – foi estabelecido em 1818, após a descrição do gênero *Aedes*. Relatos da Organização Pan-Americana de Saúde (OPAS) mostraram que a primeira epidemia de dengue no continente americano ocorreu no Peru, no início do século 19, com surtos no Caribe, Estados Unidos, Colômbia e Venezuela (NELSON, 1986).

Praticamente todas as regiões tropicais e subtropicais são quase que completamente tomadas pelos mosquitos bem como regiões mais frias como o norte da Argentina, partes da

Europa, sul dos Estados Unidos, sul da China também sofrem com sua presença. Na Figura 2 podemos observar a distribuição geográfica do *Ae. aegypti* e *do Ae. albopictus* no mundo.

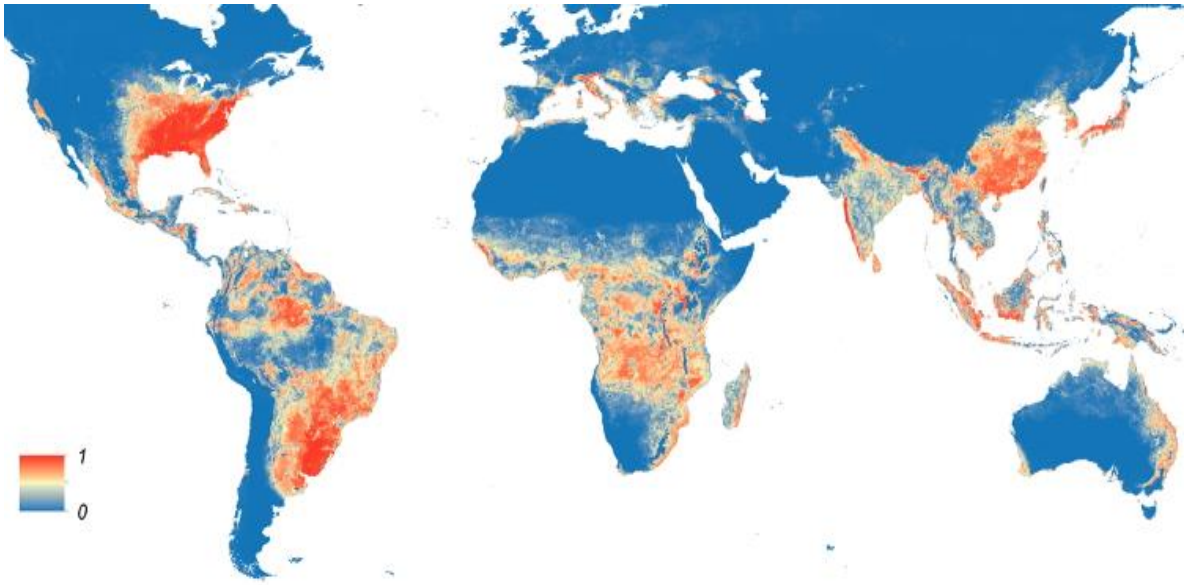


Figura 2: Distribuição mundial da população *Aedes albopictus* (KRAEMER *et al.*, 2015).

Nas Américas a dengue tem sido relatada há mais de 200 anos. Na década de 50, a Febre Hemorrágica da Dengue (FHD) foi descrita, pela primeira vez, nas Filipinas e Tailândia. Após a década de 60, a circulação do vírus da dengue intensificou-se nas Américas. A partir de 1963, houve circulação comprovada dos sorotipos 2 e 3, em vários países. Em 1977, o sorotipo 1 foi introduzido nas Américas, inicialmente pela Jamaica. A partir de 1980, foram notificadas epidemias em vários países, aumentando consideravelmente a magnitude do problema. Brasil (1982), Bolívia (1987), Paraguai (1988), Equador (1988), Peru (1990) e Cuba (1977/1981). A FHD que afetou Cuba, em 1981, é considerada como evento de extrema importância na história da dengue nas Américas. Essa epidemia foi causada pelo sorotipo 2, tendo sido o primeiro relato de febre hemorrágica da dengue, ocorrido fora do sudeste asiático e do pacífico ocidental. O segundo surto ocorreu na Venezuela, em 1989 (SILVA, 2013).

2.1.3 Distribuição Geográfica no Brasil

Sua introdução no Brasil pode ter ocorrido entre os séculos XVI e XIX, durante o comércio de escravos. Tal disseminação foi registrada com a ocorrência da doença em Curitiba (PR) no final do século XIX e em Niterói (RJ) no início do século XX (BRASIL, 2017b).

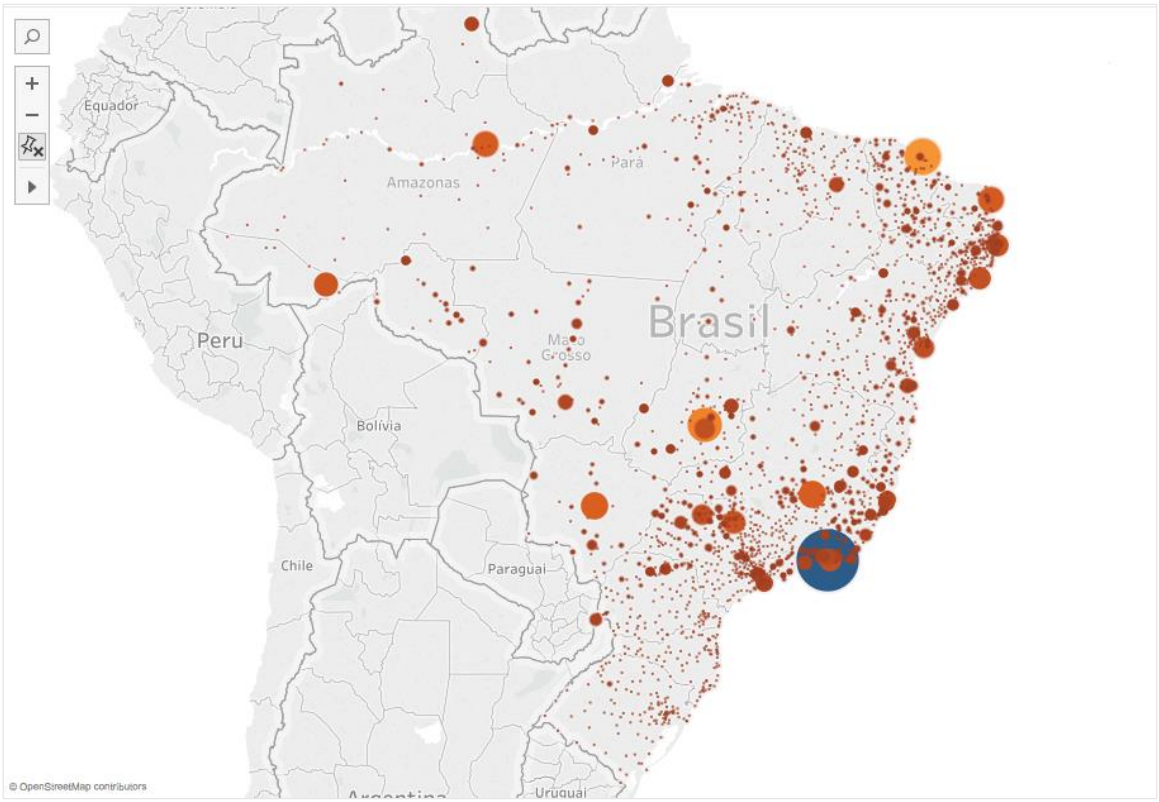


Figura 3: Distribuição dos casos no Brasil entre 2001 e 2012. (BRASIL, 2014)

No Brasil, o primeiro registro de *Ae. albopictus* ocorreu em 1986, no estado do Rio de Janeiro, posteriormente em Minas Gerais e em São Paulo e, no ano seguinte, no Espírito Santo. Em 2014, foi relatado sua presença em 3.285 municípios brasileiros, excluindo quatro estados: Sergipe, Acre, Amapá e Roraima. Embora existam semelhanças entre o comportamento do *Ae. aegypti* e do *albopictus*, as diferenças entre eles são determinantes para a dinâmica de transmissão. Hoje o *Ae. aegypti* está presente em todas as Unidades da Federação, distribuído em, aproximadamente, 4.523 municípios. Observa-se que todas as regiões tropicais e subtropicais são quase que completamente tomadas pelos mosquitos e que regiões mais frias como o norte da Argentina, partes da Europa, sul dos Estados Unidos, sul da China também sofrem com sua presença.(KRAEMER et al., 2015)

2.2 SISTEMA DE INFORMAÇÃO DE AGRAVOS DE NOTIFICAÇÃO

O Sistema de Informação de Agravos de Notificação (SINAN) foi desenvolvido no início da década de 90, pelo MS e um dos seus principais objetivos era a coleta e processamento dos dados sobre agravos de notificação em todo o território nacional, fornecendo informações para a tomada de decisão na esfera federal bem como a análise do perfil da morbidade de várias doenças no Brasil (LAGUARDIA et al., 2004).

A implantação do aplicativo SINAN-DOS iniciou-se em 1993, tendo sido precedida por testes-piloto realizados em Santa Catarina e Pernambuco. Os resultados e observações derivados desses testes não foram disponibilizados para todos os usuários ou registrados em documentos oficiais. Essa implantação foi realizada de forma gradual, em virtude do caráter voluntário de adesão das Secretarias de Estado e Municipais de Saúde, delineando um padrão irregular, tanto no uso dos formulários padronizados para os agravos de notificação compulsória, quanto na operação do programa informatizado do SINAN-DOS e análise dos dados coletados. Somente em 1998, o uso do SINAN foi regulamentado por meio de portaria ministerial, tornando obrigatória a alimentação regular da base de dados nacional pelos Municípios, Estados e Distrito Federal, designando a FUNASA, por meio do extinto CENEPI – atual Secretaria de Vigilância em Saúde, do MS – como a gestora nacional do sistema (BRITO, 1993).

O aplicativo SINAN foi concebido, originalmente, para armazenar, a partir de instrumentos e códigos de acesso padronizados em nível nacional, as informações das doenças de notificação compulsória, com suas respectivas fichas de notificação e investigação, sendo permitido às unidades federadas incluir notificações de outros agravos, adequando o sistema ao perfil epidemiológico de populações distintas (BRITO, 1993).

2.3 MINERAÇÃO DE DADOS

2.3.1 Conceito

Atualmente, vive-se em uma sociedade com um alto grau de acesso a informação, computadores e sistemas informatizados que geram milhares de dados sobre os mais diversos aspectos do nosso trabalho e nossa sociedade. Centenas de milhares de dados são gerados a cada segundo em todo mundo, entretanto apenas uma pequena parte destes dados são estruturados de uma maneira que podem ser facilmente analisados e transformados em conhecimento útil para o dia a dia das pessoas (PRIMAK, 2008).

Os dados podem fornecer soluções para os mais diversos problemas do cotidiano. Segundo Pang-Ning (2006), a mineração de dados é o processo de descoberta automática de informações úteis em grandes depósitos de dados. Descobrir informações nos dados é o principal propósito desta tecnologia, que muitas vezes é confundida com a recuperação de dados. Pesquisas na *internet* e consultas a bancos de dados são processos de recuperação de dados geralmente realizados em fontes de dados que já estão estruturadas. Em muitos casos

a mineração de dados até é usada para melhorar um processo de recuperação de dados, mas são tecnologias distintas. No decorrer dos anos, várias áreas do conhecimento tem utilizado a mineração de dados para a extração de informações em grandes volumes de dados. Uma aplicação desta tecnologia é a detecção de padrões em pagamentos, compras, *etc* de clientes em um supermercado, visando melhorar a experiência do cliente, sugerindo a compra de produtos que ele comumente utiliza ou sugerindo produtos que ele precisará comprar em breve pois a vida útil está proxima do fim. Também podemos citar a sua aplicação em pesquisas cientificas em estudos comportamentais de sites acessados por usuários, ou em estudos dos textos que são postados nas redes sociais diarimente. A aplicação desta tecnologia é praticamente ifinita e proporciona estudar comportamentos e padrões que antes eram praticamente impossíveis de se observar devido a grande volume de dados. Com o crescente aumento no poder computacional e o uso de softwares estatísticos ao longo dos anos, a precisão dos resultados tem aumentando sensivelmente e os custos tem diminuido em mesma proporção (PANG-NING, et al., 2006).

2.3.2 Mineração de Dados e a Descoberta do conhecimento

A descoberta de conhecimento é o objetivo final do processo de mineração de dados, é a etapa mais importante do processo, onde somos deparados com as reflexões decorrentes do processo de análise das informações. Na figura 4 podemos visualizar como este processo ocorre na prática.

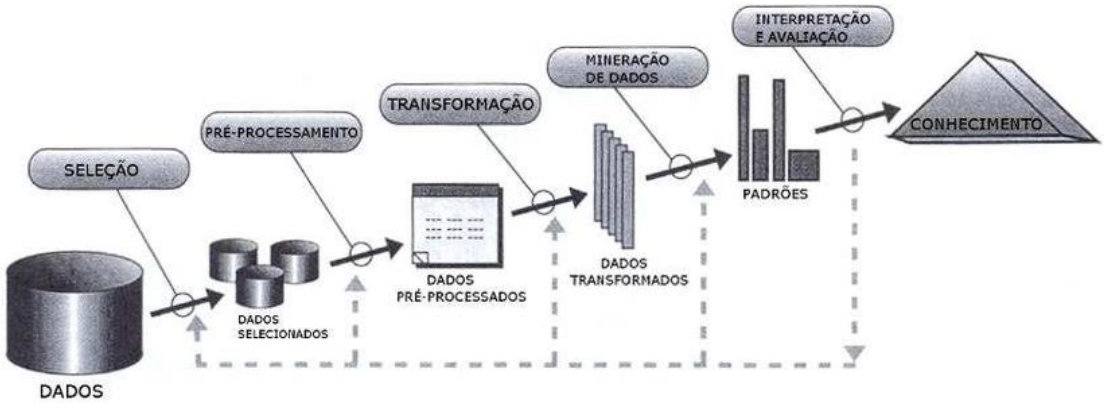


Figura 4: Fluxo básico do processo de mineração de dados. (FAYYAD et al. 1996)

Os dados de entrada podem ser dos mais variados tipos, como planilhas, arquivos de texto, estruturados e não estruturados, podem estar em um repositório centralizados ou distribuídos em localidades diversas. Estes dados são chamados de dados brutos. Mais tarde, na etapa de pré-processamento eles sofrerão uma transformação no processo de

normalização, este processo procura inconsistências e dados aleatórios que não seguem nenhum tipo de padrão. Isto ajuda diminuir o ruído nos dados, produzindo informações com maior acurácia e melhor qualidade. No pré-processamento, os dados são tratados e agrupados de uma maneira que possam ser trabalhados nas etapas subsequentes. Esta etapa é vital para o processo, colocando ordem nos grandes volumes de dados. Nesta etapa podem ser utilizadas várias técnicas, uma das mais utilizadas é a fusão de dados, ela consiste em agrupar dados que foram previamente limpos e estão com qualidade suficiente para serem analisados. (PRIMAK, 2008).

Em muitos casos, dados faltantes, campos não preenchidos ou informações que não seguem nenhum padrão, só atrapalham a análise e a visualização, por isto se faz necessário limpar estes dados, retirando linhas duplicadas, substituindo abreviações entre outros. Na etapa de pós-processamento, os dados já estão estruturados e, neste momento, já podem ser utilizados, por exemplo, para se dar carga em um banco de análises, ou serem analisados por uma ferramenta especializada em análise de dados como o *Business Intelligence*. Uma vez passada por estas etapas o conhecimento pode finalmente ser produzido (PANG-NING, et al., 2006).

2.3.3 Distância Levenshtein

Neste trabalho a principal dimensão de análise é o endereço do paciente. Mesmo com os endereços limpos e tratados é necessário checar se endereço existe e se segue padrões de alguma base de endereços. Uma base que é amplamente utilizada e por sua vez é confiável, é a base de endereços postais, fornecida pela empresa de correios do Brasil. Usamos esta base em um processo chamado *matching*, que consiste em procurar em uma base de dados confiável, dados parecidos ou iguais, permitindo desta maneira validar que o dado de fato existe e é confiável.

O algoritmo escolhido foi a *distância levenshtein*, usado maioria dos processos de mineração que envolve a etapa de *matching*, ele permite calcular de maneira relativamente simples a similaridade entre duas sequências de caracteres. Desenvolvida pelo russo Vladimir Iosifovich Levenshtein em 1965 na universidade de estadual de Moscou, o algoritmo usando técnicas matemáticas de design combinatório, consegue de maneira simples e com baixo esforço computacional determinar rapidamente o custo matemático para

criar uma réplica do texto a ser comparado. Matematicamente, a distância Levenshtein entre duas *STRINGS a,b* (do tamanho |a| e |b|, respectivamente) é dado por $\text{lev}_{a,b}(|a|,|b|)$ onde:

$$\text{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) \\ min\begin{cases} lev_{a,b}(i-1\,,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1\,,j-1\,)+1_{(a_q \neq b_j)} \end{cases} \end{cases}$$

(1)

Quando $1(a_q \neq b_j)$ é o indicador igual a 0 quando $a_1 = b_j$ é igual a 1 , e $lev_{a,b}(i,j)$ a distância entre o primeiro caractere de a e o primeiro j caracteres de b.

De maneira geral, a distância Levenshtein ou distância básica de edição entre duas STRING, é dada pelo número mínimo de operações necessárias para transformar a STRING A na STRING B. As operações são as seguintes.

- copiar um caractere da STRING A para a STRING B (custo 0)
- excluir um caractere da STRING A (custo 1)
- substituir um caractere por outro (custo 1)

Na Figura 5, pode-se analisar uma matriz que demonstra quantos passos seriam necessários para transformar a palavra “meilenstein ” em “levenshtein”. De acordo com a Figura 5, tem-se “caminhos” possíveis, grifados de amarelo, que permitiram uma solução de custo mais baixo para reproduzir a palavra original. Como o algoritmo opera de maneira matemática, torna-se assim uma ferramenta útil para aplicações que precisam determinar quão semelhantes duas sequências de caracteres são (LEVENSHTEIN, 1966).

		m	e	i	l	e	n	s	t	e	i	n
	0	1	2	3	4	5	6	7	8	9	10	11
l	1	1	2	3	3	4	5	6	7	8	9	10
e	2	2	1	2	3	3	4	5	6	7	8	9
v	3	3	2	2	3	4	4	5	6	7	8	9
e	4	4	3	3	3	3	4	5	6	6	7	8
n	5	5	4	4	4	4	3	4	5	6	7	7
s	6	6	5	5	5	5	4	3	4	5	6	7
h	7	7	6	6	6	6	5	4	4	5	6	7
t	8	8	7	7	7	7	6	5	4	5	6	7
e	9	9	8	8	8	7	7	6	5	4	5	6
i	10	10	9	8	9	8	8	7	6	5	4	5
n	11	11	10	9	9	9	8	8	7	6	5	4

Figura 5: Matriz de caminhos possíveis de reprodução de duas palavras.

2.3.4 Data Mart

O resultado de cada uma das etapas de mineração precisa ser armazenado de maneira estruturada, afim que se possa acompanhar o progresso e avaliar a qualidade das informações. Em uma parte das empresas estes dados são armazenados em grandes bancos de dados chamados DataWare House (DW), que por sua vez é composto de vários *Data Mart* (DM) que são subconjunto de um (DW), que normalmente consiste em uma única área temática, podendo ser dependente ou independente. Caso seja dependente, ele é criado diretamente a partir do DW, tendo a vantagem de usar um modelo de dados consistente e apresentar dados de qualidade. Por outro lado, um DM pode ser independente e pequeno, podendo ser projetado para uma unidade estratégica de negócios específica (PRIMAK, 2008).

As diferenças entre DM e DW são apenas com relação ao tamanho e ao escopo do problema a ser resolvido. Os DMs atendem as necessidades de unidades específicas de negócio da empresa, auxiliando no tratamento de um problema departamental ou local. Já os DWs atendem as necessidades de toda a empresa, auxiliando no suporte à decisão de todos os níveis empresariais. (PRIMAK, 2008)

2.3.5 Extração, Transformação e Carga

Para alimentar de dados o DM, descrito acima é necessário o uso de ferramentas especializadas no envio de dados para os bancos de dados onde os DataMarts estão hospedados. O processo de extração transformação e carga, ou em inglês *Extraction, Transform and Load* (ETL) é um componente integral de qualquer projeto centrado em dados. Ele consiste em extração (leitura dos dados de um ou mais bancos de dados), transformação (conversão dos dados extraídos de sua forma anterior na forma em que precisam estar para que sejam colocados em um DM ou apenas em outro banco de dados) (TURBAN et al., 2009).

ETL refere-se a uma coleção de ferramentas que desempenham um papel crucial para ajudar a descobrir e corrigir problemas de qualidade de dados e carregar eficientemente grandes volumes de dados no DM. A acurácia e a pontualidade de relatórios, consultas locais e análises preditivas dependem da capacidade de obtenção dados de alta qualidade no DM de bases de dados operacionais e fontes de dados externas (RUBIO SERRANO, 2014).

Para Primak (2008) o processo ETL é uma das etapas mais críticas de um projeto para criar um banco de análises, pois envolve a fase de movimentação de dados (PRIMAK, 2008). Já Turban (2009) considera uma peça extremamente importante na integração de dados, sendo o seu principal objetivo realizar o carregamento de dados integrados e limpos. (TURBAN et al., 2009).

No ETL é onde os dados são movimentados de fato dentro do processo de mineração, ele acontece em diversas etapas da mineração dos dados. Em muitos casos alimentando os motores de mineração, em outras armazenando o resultado da mineração. Etapa crucial e vital para qualquer projeto de mineração de dados.

2.4 BUSINESS INTELLIGENCE

2.4.1 Conceito

Segundo Silva et al. (2016), uma das definições do *Business Intelligence* é que funciona como uma ferramenta para processamento de uma grande quantidade de dados para o apoio à tomada de decisões gerenciais. Este processo é baseado na transformação de dados em informações, decisões e finalmente em ações. Inclui dessa forma, arquiteturas, ferramentas, bancos de dados, aplicações e metodologias. Permite ainda, acesso interativo aos dados (às vezes em tempo real), proporciona a manipulação dos dados e fornece aos gerentes e analistas de negócios a capacidade de realizar uma análise mais adequada.

Entretanto o *Business Inteligence*, é uma das peças de um processo mais complexo que é a Análise de Dados ou *Data Analytics* como é conhecido em inglês. O objetivo final da tecnologia de *Business Inteligence* é ajudar o usuário a compreender e analisar o passado, permitindo dessa maneira compreender as interações entre os dados e criar estratégias para solucionar problemas. Na figura 6 podemos visualizar como ferramentas e processos se conectam para atingir o objetivo principal da análise de dados que é poder prever o que pode acontecer, permitindo assim traçar cenários e simular o risco envolvido em cada uma das decisões a ser tomada.

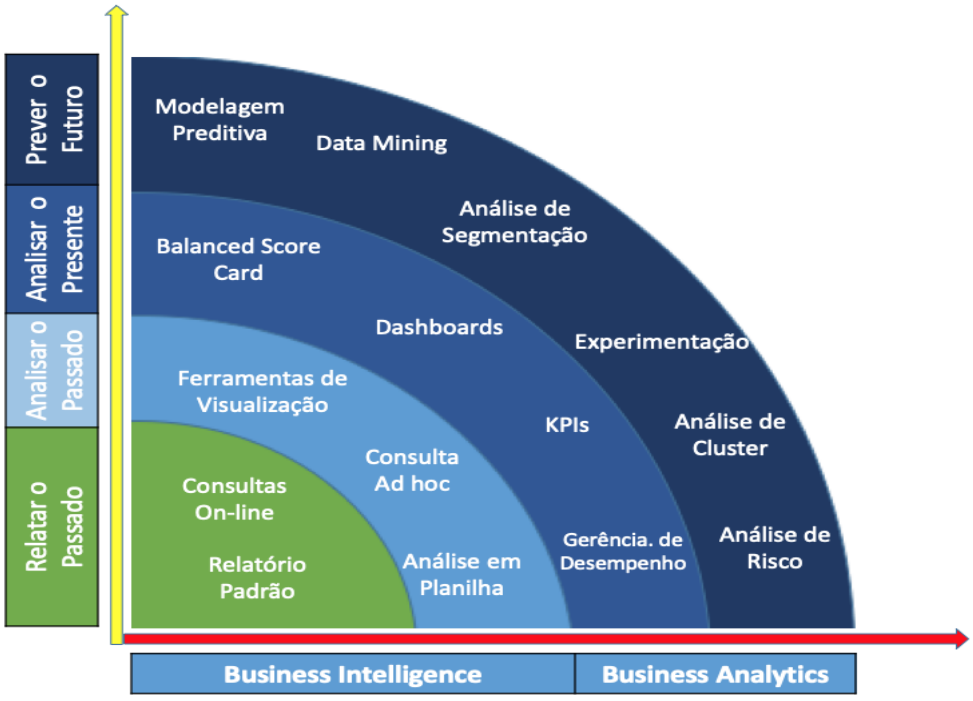


Figura 6: Maturidade dos processos de análise de dados. Autoria Própria

As ferramentas de *Business Intelligence* já existem a algum tempo e tem entre os seus principais usuários os gestores de empresas privadas e públicas, sendo muito utilizada na análise de dados de campanhas de marketing e análises de vendas. Há algum tempo, vem também se tornando peça fundamental para o ambiente acadêmico, com sua facilidade de operação, processos de importação de dados mais simples e criação iterativa de gráficos, estas ferramentas vêm proporcionando novas descobertas em todos os campos.

Outro fator relevante é que as ferramentas atuais permitem a criação de painéis que podem ser disponibilizados na web. Isto permitiu que várias pessoas passassem a compreender os dados de maneira gráfica e isto tem criado uma revolução neste segmento. De maneira geral o principal objetivo é proporcionar aos usuários uma experiência com dados, para que uma vez analisando o passado, possam compreender padrões nos dados e assim prever acontecimentos futuros.

2.4.2 Dashboards

Os *Dashboards* ou painéis gráficos são um subconjunto de relatórios que incluem a habilidade formal de publicar os dados em relatórios baseados na *Web* por meio de interfaces gráficas interativas, que exibem informações mediante um painel de instrumentos. Estes painéis exibem a situação de um determinado momento da empresa, sendo gerados com base em indicadores, permitindo uma visualização rápida, dinâmica e interativa dos pontos mais relevantes que envolvem uma parte específica do negócio ou de todo o seu conteúdo. (SALLAM et al., 2011).

Os *Dashboards* são componentes comuns, se não de todos, da maior parte dos sistemas de gerenciamento, de medição e de suítes. Esses painéis proporcionam exibições visuais de informações importantes que são consolidadas e organizadas em uma tela única para serem analisadas e exploradas facilmente, proporcionando agilidade e precisão na tomada de decisão (TURBAN et al., 2009).

As aplicações de *dashboards* por meio de gráficos coloridos permite o agrupamento *roll-up* , onde os registros similares são agrupados gerando operações de contagem ou soma destes agrupamentos, ou detalhamento *drill-down,* que se consiste no detalhamento das informações agrupadas pelo processo de *roll-up,* fazendo com que o usuário possa “explorar” o processo de construção daquela informação (COUTO, 2012).

2.5 CÓDIGO DE ENDEREÇAMENTO POSTAL BRASILEIRO

Segundo Aranha (1997) o CEP foi inicialmente projetado para acelerar e reduzir o custo da entrega de correspondências, permitindo a separação automatizada dos objetos postais. Entretanto os códigos de endereçamento acabaram desempenhando uma função muito mais ampla de servir como uma linguagem hierarquizada e bastante consistente de descrição do espaço geográfico. Neste papel, seu uso disseminou-se para outras organizações não necessariamente postais, e em particular pelas áreas de planejamento e marketing das empresas privadas. Combinando características da estratégia de descrição nominal com elementos atenuados da estratégia de sequências de estruturas hierarquizadas.

Atualmente o CEP está estruturado segundo o sistema decimal, sendo composto por 5 variáveis principais de Região, Sub-região, Setor, Subsetor, Divisor de Subsetor e Identificadores de Distribuição. Os correios dividiram o Brasil em dez regiões postais para fins de codificação postal, utilizando como parâmetro o desenvolvimento socioeconômico e fatores de crescimento demográfico de cada Unidade da Federação ou conjunto delas. A distribuição do CEP foi feita no sentido anti-horário a partir do estado de São Paulo, pelo primeiro algarismo, conforme podemos visualizar na figura 7. (BRASIL, 2017a)

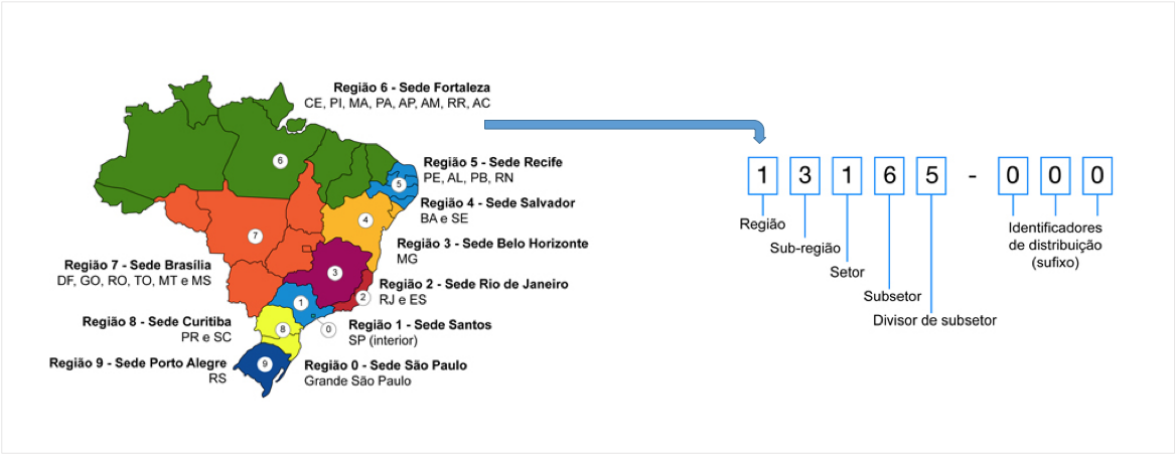


Figura 7: Estrutura do CEP.(BRASIL, 2017a)

Esta estrutura numérica permite de maneira simples identificar, qual posição geográfica a quem um determinado número de CEP se refere, uma vez que oferece uma classificação lógica para agrupar as regiões geográficas. O CEP fornece um nível de detalhamento maior, quando comparado ao sistema utilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) de macrorregiões e microrregiões. Fazendo um paralelo entre os dois sistemas, as macrorregiões e as microrregiões seriam equivalentes aos níveis de Região e Sub-região do

CEP. Como vantagem CEP permite mais três níveis de detalhamento de uma localidade. Nas análises geográficas realizadas nesta pesquisa o nível de detalhamento do CEP se mostrou mais adequado, permitindo demonstrar aos usuários dos painéis, quais são as ruas e quadras que tem mais casos e um maior nível de detalhe, aproximando as pessoas de uma realidade geográfica que faz parte do seu cotidiano, ampliando a compreensão do indicador e trazendo sentido para a realidade do usuário. Este objetivo não poderia ser atingido em sua totalidade utilizando-se apenas a microrregião que hoje é o sistema utilizado pela maioria das secretarias de saúde do Brasil. Na figura 8 podemos visualizar de maneira visual como o detalhamento do CEP, consegue chegar a um nível mais detalhado geográfico, tomando como exemplo uma pequena área de uma cidade do estado de São Paulo.

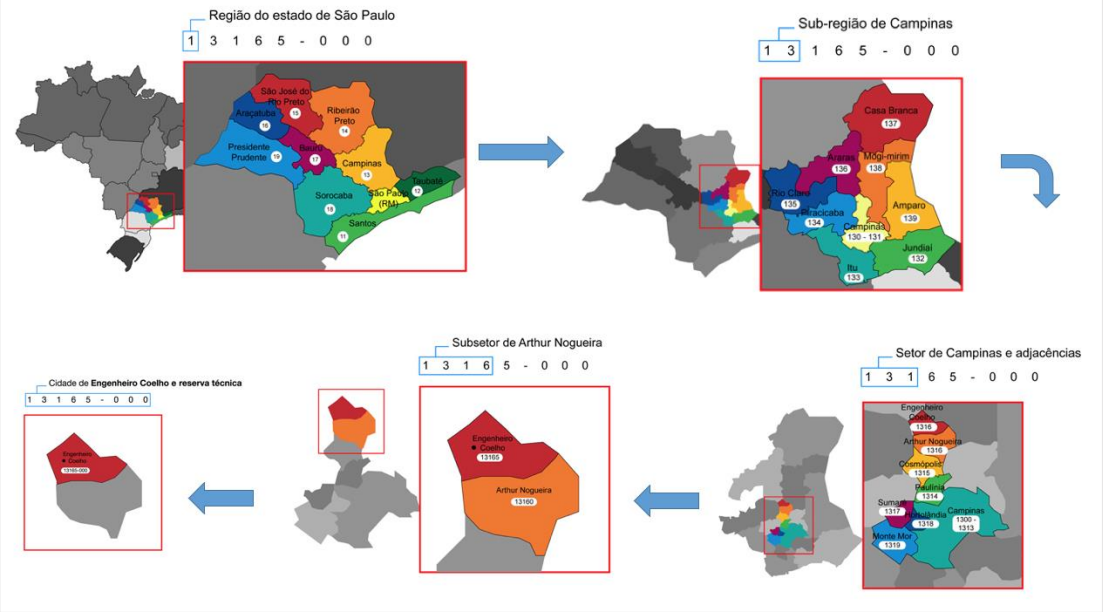


Figura 8: Estrutura detalhada das variáveis do CEP. (BRASIL, 2017a)

O código de endereçamento postal, não é um sistema exclusivo do Brasil e sim um modelo internacionalizado utilizado em aproximadamente 194 países em todo mundo. A União Postal Universal (UPU) é o organismo internacional, responsável pela padronização e interoperabilidade entre os diversos sistemas de endereçamento postal dos países membros, (LYALL, 2013).

2.6 SISTEMA DE INFORMAÇÃO GEOGRÁFICO

Um sistema de informação geográfico, é um sistema que permite a representação espacial de informações de um sistema computacional em um mapa cartográfico. Permite de maneira simples e visual compreender padrões sob a perspectiva geográfica e tem sido um recurso amplamente utilizado na área de logística, segurança e combate epidemiológico. O uso das tecnologias de informação geográficas em alguns setores da gestão pública tem se tornado cada vez mais comum na tomada de decisão por parte dos gestores públicos e possibilita análise cada vez mais detalhadas , integrando dados de áreas e fontes diversas (LITE, 2010).

O geoprocessamento é uma tecnologia interdisciplinar que permite a convergência de diferentes disciplinas científicas para o estudo de fenômenos ambientais e urbanos. Enquanto o Georreferenciamento, descreve um conjunto de objetos que possuem representação espacial e estão associados a regiões da superfície da terra, representando a visão de campos e de objetos. (CÂMARA et al., 2002)

Dessa forma, o geoprocessamento se utiliza de ferramentas computacionais, que automatizam a produção de documentos cartográficos, realizando análises complexas quando integradas aos dados de diversas fontes. (NETO et al., 2014)

Além disso, os SIGs resultam da combinação entre três tipos de tecnologias distintos: 1. O sensoriamento remoto – utiliza-se de satélites e radares para coleta de dados sobre um objeto ou fenômeno sem que ocorra contato físico entre o mesmo e o coletor. 2. O GPS-tecnologia de localização por satélite, onde se envia informações sobre a posição de algo, por meio, do sistema de navegação por satélite a partir de um dispositivo móvel, independente do horário e condição climática. É possível ainda, informar endereços, rotas, posições de latitude e longitude. 3. Geoprocessamento - consiste em usar softwares especialmente programados para essa função, ou seja, através do sensoriamento remoto e do GPS são obtidas informações e estas são manipuladas e passam a produzir mapas, cartogramas, gráficos e sistematizações em geral (PENA, 2017).

2.7 AEDESMAPS

A Plataforma *AedesMaps* foi construída para conscientizar e permitir que a população tenha um papel mais ativo no combate a proliferação do vetor da dengue. Basicamente se trata de um conjunto de aplicações que permitem que os usuários possam analisar dados históricos de casos de dengue segmentados em microrregiões do Distrito Federal e postar possíveis focos do mosquito. O grande diferencial desta plataforma é a sua integração com as redes sociais, onde cada possível foco é postado também na rede social Facebook, permitindo assim que governo e população tenham conhecimento e possam concentrar seus esforços onde é realmente mais necessário (LIMA, 2018).

A principal aplicação dentro da plataforma *AedesMaps* é o mapa de casos. Este mapa foi desenvolvido a partir dos dados minerados do SINAN. Ele é a materialização prática do processo de mineração de dados descrito neste trabalho. Ele permite agrupar os casos por CEP. Atuando em conjunto com o sistema de mapas da *Google*, permite que o usuário visualize imagens de satélite juntamente com pontos marcados de cores diferentes, que podem variar de acordo com a quantidade de casos (LIMA, 2018).

A aplicação permite várias iterações com o usuário, por exemplo ao clicar em um ponto qualquer do mapa, permitindo ele visualizar na barra lateral direita uma aba com mais informações sobre a localidade em questão. Nesta aba o usuário pode visualizar quantidade de casos por ano, logradouro da localidade, CEP e outras informações relevantes. Outra funcionalidade interessante no projeto é que ele diferencia através das cores, casos de doença e possíveis focos do aedes, onde a cor azul do ponto no mapa é o indicador de um possível caso. Usar dados georreferenciados obtidos a partir do processo de mineração desenvolvido por esta pesquisa é o grande diferencial deste projeto. Na Figura 7 podemos visualizar de maneira detalhada como se dá o aspecto gráfico e as funcionalidades descritas, como um produto final disponível na *internet* através do endereço eletrônico http://www.aedesmaps.com.br .

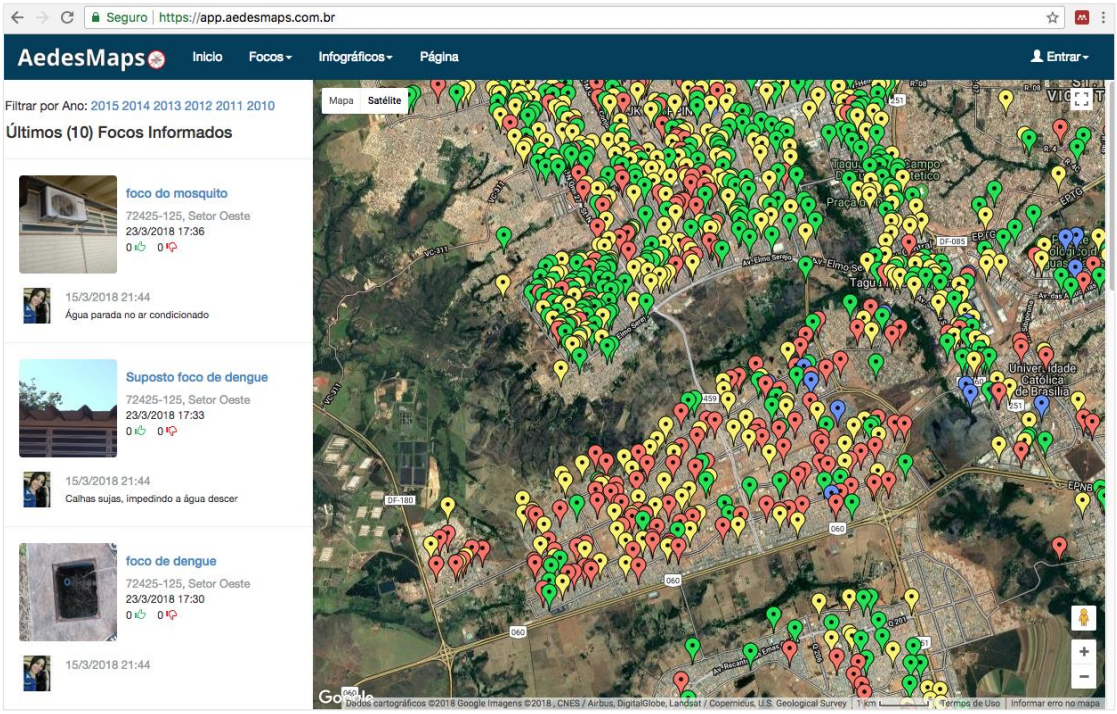


Figura 9: Aplicativo *AedesMaps.* (LIMA, 2018)

2.8 TABLEAU

O *Tableau Desktop* é uma ferramenta de *Businnes Intelligence e Businnes Analytics*, ela permite ao usuário criar *dashboards*, que permitem analisar grandes volumes de dados e implementar uma vasta análise de indicadores de performance. O *Tableau Desktop* é uma das soluções mais comercializadas do mercado para o segmento de visualização de dados. O *Tableau Desktop* auxilia a criação de painéis interativos, que permite vários tipos de análises dos dados. Combina uma linguagem descritiva para renderização de gráficos denominada *Visual Query Language* (*VizQL*) que é uma linguagem de consulta estruturada para bancos de dados. Na Figura 10 pode-se visualizar a sua tela principal onde campos de fontes de dados são agrupados em dimensões e medidas, permitindo que o usuário possa de maneira simples utilizar recursos de arrastar e soltar movendo campos para linhas e colunas. Permitindo, desta maneira, fazer análises e cruzar informações de modo simples e eficiente.(PIMENTEL et al., 2016)

Um ponto positivo na utilização do Tableau é que não é necessário grande conhecimento na área de programação de computadores ou de gerenciamento de bancos de dados relacionais. Com uma interface gráfica intuitiva ele conduz o usuário a uma solução rápida para visualização e interpretação de dados, em contraposição à lentidão e rigidez de soluções legadas (TABLEAU, 2014).

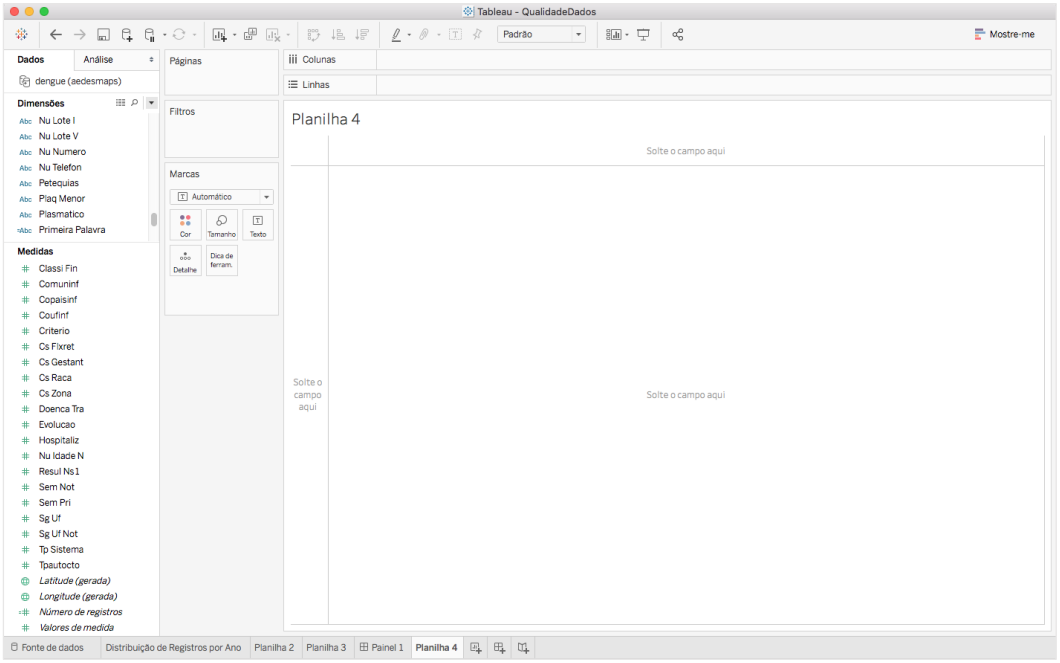


Figura 10: Tela principal de análises do *tableau desktop.* Autoria Própria

3 METODOLOGIA

3.1 AMBIENTE DO ESTUDO

Para o desenvolvimento deste trabalho foram utilizados os microdados extraídos do sistema SINAN, disponibilizados pelo Núcleo de Medicina Tropical (NMT) da Universidade de Brasília (UnB) com casos de dengue, do período de 2010 a 2015. Nestes microdados utilizou-se a informação de endereço, contida no campo de logradouro do endereço fornecido pelo paciente nas fichas de notificação individual, que foram enviadas ao SINAN.

Utilizando técnicas de mineração de dados, foi melhorada a qualidade destes endereços, corrigindo abreviações incorretas e informações incompletas que dificultam o processo de geocodificação. O trabalho de mineração foi dividido em três fases. Sendo a primeira, de melhoria da qualidade dos endereços, a segunda, de geocodificação e georreferenciamento e a terceira, de agregação dos dados utilizando o CEP dos endereços georreferenciados como dimensão de agregação. Após os dados serem georreferenciados, foi possível utilizar a ferramenta de *Business Intelligence*, Tableau, para analisar os dados e criar painéis que permitem estudar correlações e procurar padrões nos dados. Na Figura 11 é possível visualizar o processo de maneira global.

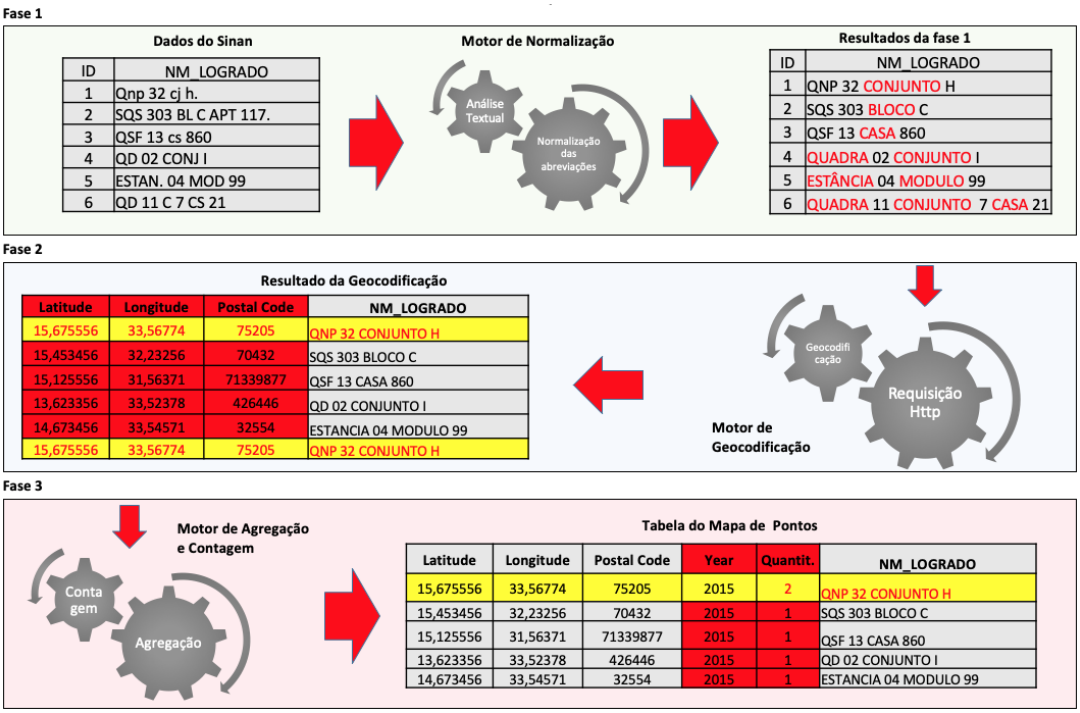


Figura 11: Fluxo de dados de mapeamento de casos. Autoria Própria

3.1.1 Microdados do SINAN

Os dados fornecidos para estudo encontravam-se em formato DBF separados por ano. Para trabalhar com estes dados foi preciso inicialmente importá-los para um banco de dados relacional (Mysql). Nesta primeira fase, não foi realizada nenhum tipo de filtragem ou tratamento nos dados. Para o armazenamento destes dados foi criada uma tabela chamada SINAN contendo os 101 campos originários do arquivo DBF, mantendo-se os tipos de dados e tamanhos correspondentes ao dicionário de dados fornecido pelo MS. A carga inicial dos dados foi realizada utilizando um *script* desenvolvido em PHP, que mapeia cada um dos campos do arquivo DBF, importando os dados para a tabela SINAN do banco de dados.

Os dados fornecidos compreendem casos confirmados e suspeitos reportados no Distrito Federal referentes aos anos de 2010 a 2015. Os dados referentes ao ano de 2016, 2017 não serão computados, pois ainda não foram disponibilizados para o NMT-UnB. Alguns campos necessitaram ser adicionados para o funcionamento mais eficiente do motor de georreferenciamento e viabilizar futuras análises e agrupamentos dos dados. Estes novos campos são apresentados na Tabela 3.

Quadro 3: Campos adicionados a tabela de casos do SINAN. Autoria Própria

Campo	Tamanho	Tipo de Dados	Obrigatório	Descrição
ID	11	Inteiro	SIM	Chave primária da tabela, identifica cada registro com um número de maneira única
CRAW	1	Texto	NÃO	Usado para determinar o status do registro no decorrer dos vários processos de análise do registro.
LATITUDE	45	Texto	NÃO	Latitude do endereço
LONGITUDE	45	Texto	NÃO	Longitude do endereço
RESULT	255	Texto	NÃO	Logradouro completo do endereço retornado do motor de georreferenciamento
CEP	45	Texto	NÃO	CEP do endereço georreferenciados.
CEP_PROVAVEL	45	Texto	NÃO	CEP do Resultado da primeira fase de mineração usando o algoritmo de LEVENSHTEIN
LOG_PROVAVEL	255	Texto	NÃO	Logradouro do Resultado da primeira fase de mineração usando o algoritmo de LEVENSHTEIN
LEVENSHTEIN	45	Texto	NÃO	Resultado do teste do algoritmo de LEVENSHTEIN

Após os novos campos serem adicionados à tabela de casos no banco relacional, pode-se verificar que a quantidade de casos importados e agrupados por ano confere com a quantidade de linhas encontradas nos arquivos DBF. No processo de importação foram importados **87.096** registros. Após análise dos registros, foram excluídos **2.567** registros sem logradouro. Apresentado assim no final uma quantidade de **84.529** registros com o logradouro preenchido na tabela.

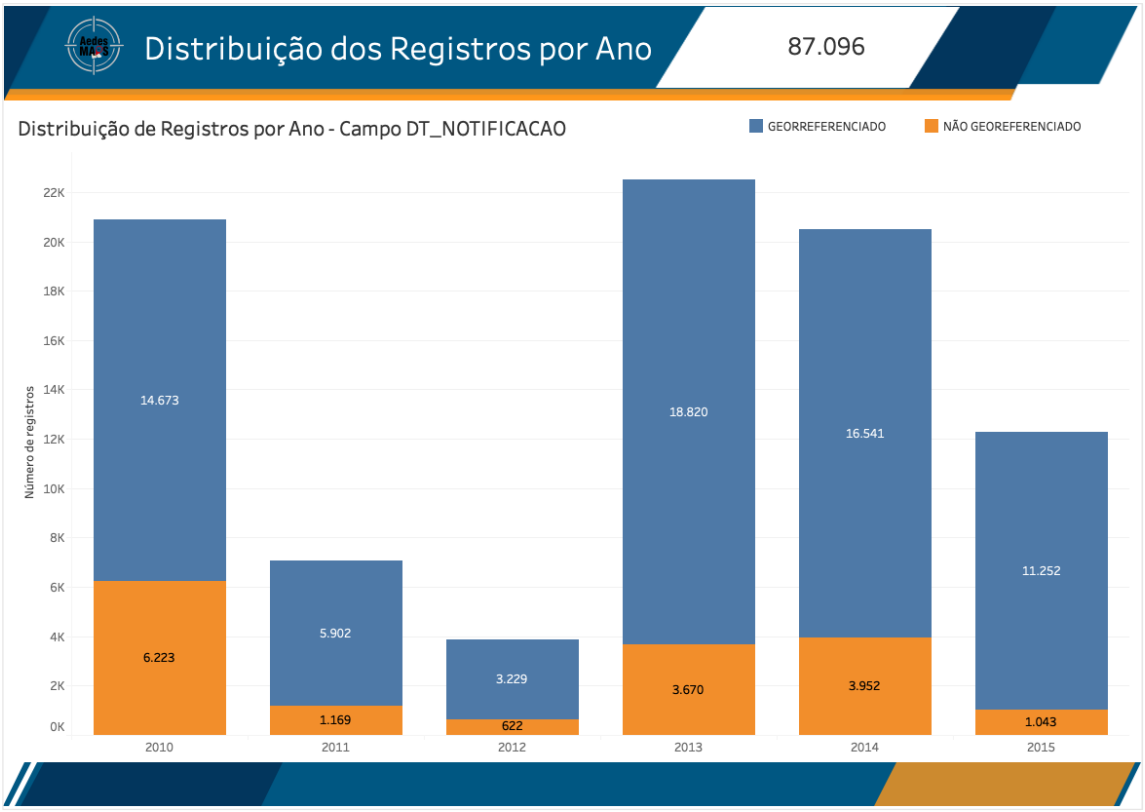


Figura 12: Distribuição dos registros por ano. Autoria Própria

Na Figura 12 é possível visualizar os registros agrupados por ano, onde pode-se perceber que o ano com menor quantidade de registros foi o ano de 2012. Segundo o Ministério Saúde a diminuição dos casos de dengue não foi um movimento acontecido unicamente no Distrito Federal. Em todo o brasil houve uma queda de aproximadamente 78% dos casos de Dengue. A tendência à queda foi atribuída a diversos fatores como a organização da rede pública e uma maior conscientização da população na adoção hábitos de prevenção. De todos os estados, o Distrito Federal e outros 23 apresentaram reduções importantes de casos graves este ano, com exceção de Alagoas, Mato Grosso e Goiás, que registraram aumento. O estado com maior redução foi o Amazonas, com queda de 96% em relação ao mesmo período do ano passado; seguido pelo Acre e Roraima, ambos com 94%; Paraná, com 93%; São Paulo, com 83%; Espírito Santo, 78%; e Rio de Janeiro, 76%. Em números absolutos, o estado do Rio de Janeiro foi o que apresentou a maior redução de casos graves, registrando 891 casos graves de janeiro ao início de novembro de 2012, contra 3.783 no mesmo período do ano anterior.

3.1.2 Motor de normalização

A normalização dos logradouros foi dividida em duas fases. Na primeira foi utilizada a base de dados nacional de CEP dos correios que foi importada ao banco de dados MYSQL em uma tabela chamada CEP, com a finalidade de permitir a utilização do algoritmo da *distância Levenshtein*, onde são comparadas as similaridades entre duas sentenças. Assim, foi possível verificar quais logradouros tinham maior similaridade com os da base de dados nacional de CEP. Na segunda etapa, o foco do trabalho foi os registros onde a similaridade foi baixa ou inexistente. Utilizando um dicionário de abreviações criado a partir dos logradouros do SINAN, os logradouros foram atualizados e submetidos novamente ao processo de comparação com os logradouros da base nacional de CEP.

Para tornar o processo de mineração mais eficiente e a pesquisa na base de CEP mais rápida foi criado na tabela CEP no banco de dados um índice do tipo FULLTEXT. Este índice permite pesquisas por palavras muito rápidas, devido ao fato de utilizar um algoritmo otimizado de procura em grandes volumes de textos armazenados em bancos de dados. Com o intuito de tornar o processo mais eficiente foi implementada uma função interna no banco de dados com o algoritmo de Levenshtein. Esta decisão melhorou de sobremaneira a velocidade e simplificou o processo. Na Tabela 4, pode-se visualizar o código fonte da função que implementa o algoritmo de Levenshtein.

Quadro 4: Código fonte do Algoritmo de Levenshtein para MYSQL (LENTZ, 2013).

```
DELIMITER $$
DROP FUNCTION IF EXISTS LEVENSHTEIN $$
CREATE FUNCTION LEVENSHTEIN(s1 VARCHAR(255) CHARACTER SET utf8, s2 VARCHAR(255) CHARACTER SET
utf8)
  RETURNS INT
  DETERMINISTIC
  BEGIN
    DECLARE s1_len, s2_len, i, j, c, c_temp, cost INT;
    DECLARE s1_char CHAR CHARACTER SET utf8;
    -- max strlen=255 for this function
    DECLARE cv0, cv1 VARBINARY(256);
    SET s1_len = CHAR_LENGTH(s1),
        s2_len = CHAR_LENGTH(s2),
        cv1 = 0x00,
        j = 1,
        i = 1,
        c = 0;
    IF (s1 = s2) THEN
      RETURN (0);
    ELSEIF (s1_len = 0) THEN
      RETURN (s2_len);
    ELSEIF (s2_len = 0) THEN
      RETURN (s1_len);
    END IF;
    WHILE (j <= s2_len) DO
      SET cv1 = CONCAT(cv1, CHAR(j)),
          j = j + 1;
    END WHILE;
    WHILE (i <= s1_len) DO
      SET s1_char = SUBSTRING(s1, i, 1),
          c = i,
          cv0 = CHAR(i),
```

```
          j = 1;
      WHILE (j <= s2_len) DO
        SET c = c + 1,
            cost = IF(s1_char = SUBSTRING(s2, j, 1), 0, 1);
        SET c_temp = ORD(SUBSTRING(cv1, j, 1)) + cost;
        IF (c > c_temp) THEN
          SET c = c_temp;
        END IF;
        SET c_temp = ORD(SUBSTRING(cv1, j+1, 1)) + 1;
        IF (c > c_temp) THEN
          SET c = c_temp;
        END IF;
        SET cv0 = CONCAT(cv0, CHAR(c)),
            j = j + 1;
      END WHILE;
      SET cv1 = cv0,
          i = i + 1;
    END WHILE;
    RETURN (c);
  END $$
DELIMITER ;
```

Para exemplificar o processo de normalização, a Figura 13 demonstra o passo a passo do processo de normalização dos dados, onde o objetivo final do processo é determinar o CEP de onde ocorreu o caso, ou pelo menos um CEP aproximado do caso. O método de Levenshtein fornece uma maneira relativamente rápida e sem custos de conexão com a *internet* e o uso da *api* do *Google* de encontrar um endereço similar e assim determinar a região aproximada onde aconteceu o caso.

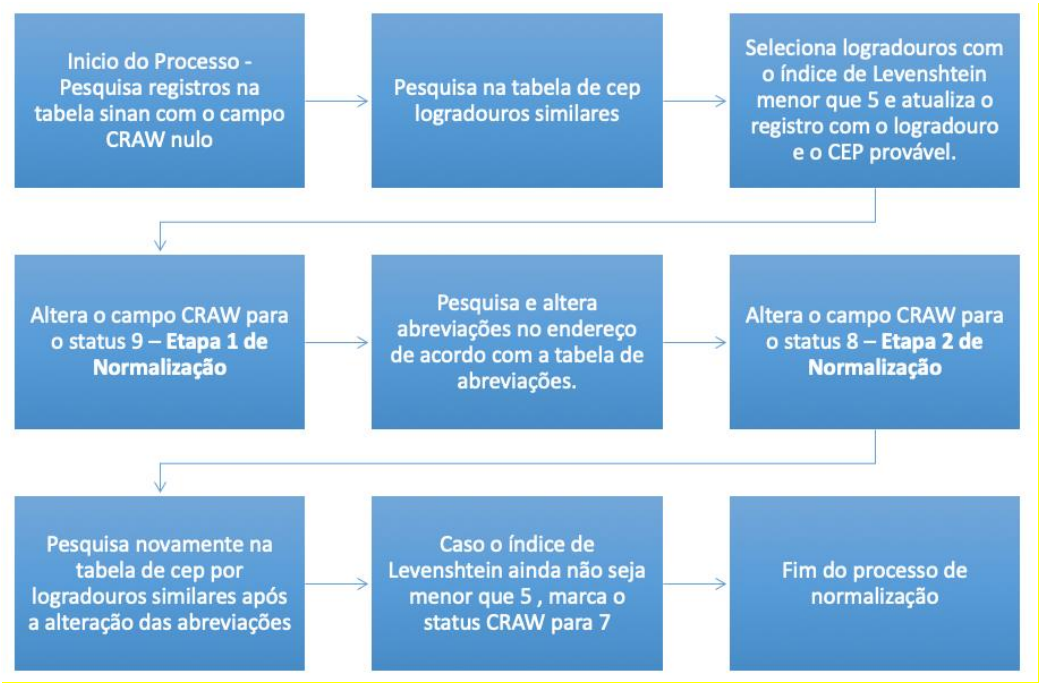


Figura 13: Fluxo de execução do motor de normalização. Autoria Própria

Através de métodos matemáticos e computacionais por aproximação, é possível determinar o CEP mais provável ou próximo de uma localidade. Isto ocorre através da similaridade de logradouros pesquisados na base nacional de CEP, que já está armazenada

no banco de dados. Na Tabela 6 em uma pequena amostra dos registros é possível observar a efetividade do método aplicado a um conjunto de endereços do banco de dados. O indicador de similaridade de Levenshtein, demonstrado na Tabela 5, demonstra que quanto menor o valor do índice, mais acurado será o resultado. Sendo que o grau 0 demonstra que as duas sentenças são praticamente idênticas

Quadro 5: Dados resultantes das etapas do motor de normalização. (LUSTOSA et al., 2017)

NM_LOGRADO	LOG_PROVAVEL	CEP_PROVAVEL	LEVENSHTEIN
SQS 406 BLOCO F	SQS 406 BLOCO F	70255060	0
SQS 215 BL J	SQS 215 BLOCO J	70294100	3
SQS 411 BL J	SQS 411 BLOCO J	70277100	3
SQS 313 BL F	SQS 313 BLOCO F	70382060	3
SQS 214 BL C	SQS 214 BLOCO C	70293030	3
SQS 207 BLOCO K APT	SQS 207 BLOCO K	70253110	4
SQS 206 BLOCO J APT	SQS 206 BLOCO J	70252010	4
QR 606 CONJ 02	QR 606	72322200	8

Fonte: SINAN

Na segunda fase de normalização, os registros que não obtiveram uma classificação de similaridade satisfatória usando o algoritmo de Levenshtein, passam por um processo de qualificação, onde abreviações incorretas são substituídas por uma segunda rodada de normalização ou em alguns casos o logradouro é simplificado. No estudo em questão, por exemplo, não se faz necessária a informação de qual casa ou apartamento aconteceu o caso, apenas o CEP já é o suficiente para se agregar os casos. Ainda assim, foi removido a informação de casa e de apartamento e, ainda, foi submetido à uma nova fase para análise pelo algoritmo Levenshtein onde se chamou de Rodada 2 de normalizações. Esgotado estes passos, o registro é alterado com o *status* 7 para uma última pesquisa, desta vez submetida à base cartográfica do *Google Maps* e não a base de CEP.

3.1.3 Motor de geocodificação dos logradouros

O motor de geocodificação, basicamente é responsável por duas tarefas: a primeira é encontrar um par de coordenadas geográficas para registros onde não foi possível determinar um logradouro similar na base de CEP da fase de normalização. A segunda é localizar as coordenadas geográficas do CEP de outros registros, onde o motor de normalização conseguiu encontrar um logradouro provável na base de CEP. Em ambos os processos, o funcionamento é basicamente o mesmo, mudando apenas nos casos onde não foi possível se encontrar um logradouro provável, nestes casos o campo utilizado pelo processo de georreferenciamento é o NM_LOGRADO, tratado e limpo pela última fase de normalização dos registros. Nos demais casos o campo utilizado é o LOG_PROVAVEL que contém o

logradouro similar da base de CEP. Na Figura 12 é possível observar o fluxo do processo de geocodificação.

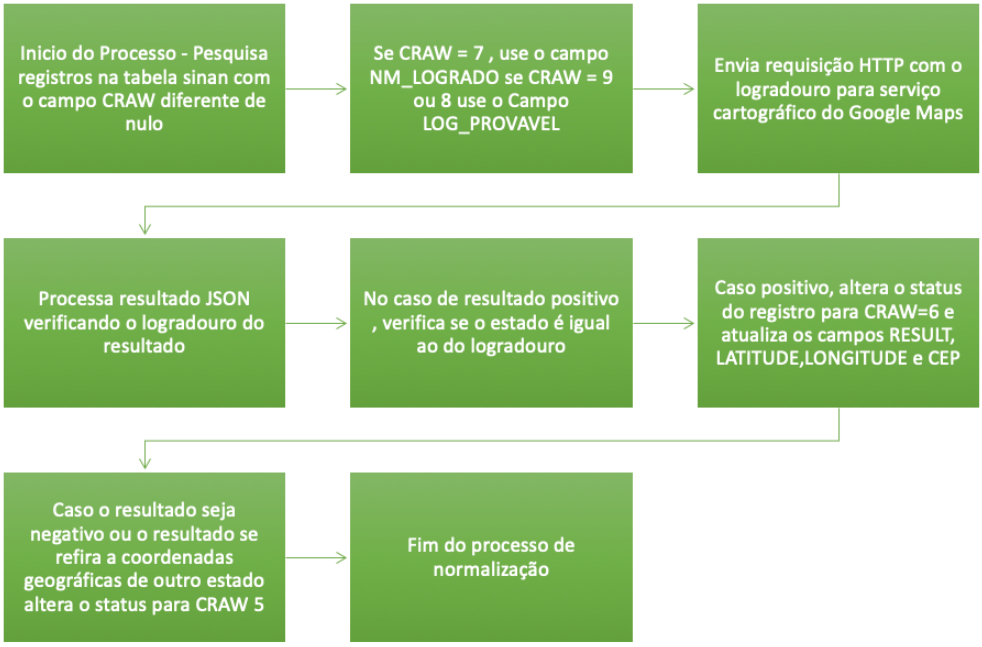


Figura 14: Fluxo de execução do Motor de Geocodificação. Autoria Própria

Para consultar os registros na base cartográfica do *Google*, foi desenvolvido um programa na linguagem PHP, que consulta cada um dos registros no banco de dados e usa o protocolo HTTP para enviar o logradouro como um parâmetro de requisição aos servidores do *GoogleMaps*. Os servidores do *GoogleMaps*, por sua vez, utilizam algoritmos proprietários e muito avançados, que conseguem localizar com mais precisão os logradouros onde o motor de normalização não obteve sucesso comparando com a base de CEP. Na Figura 15 pode-se observar de maneira simplificada como o motor de georreferenciamento acessa os servidores do *Google* e como é tratada a resposta obtida por este serviço.

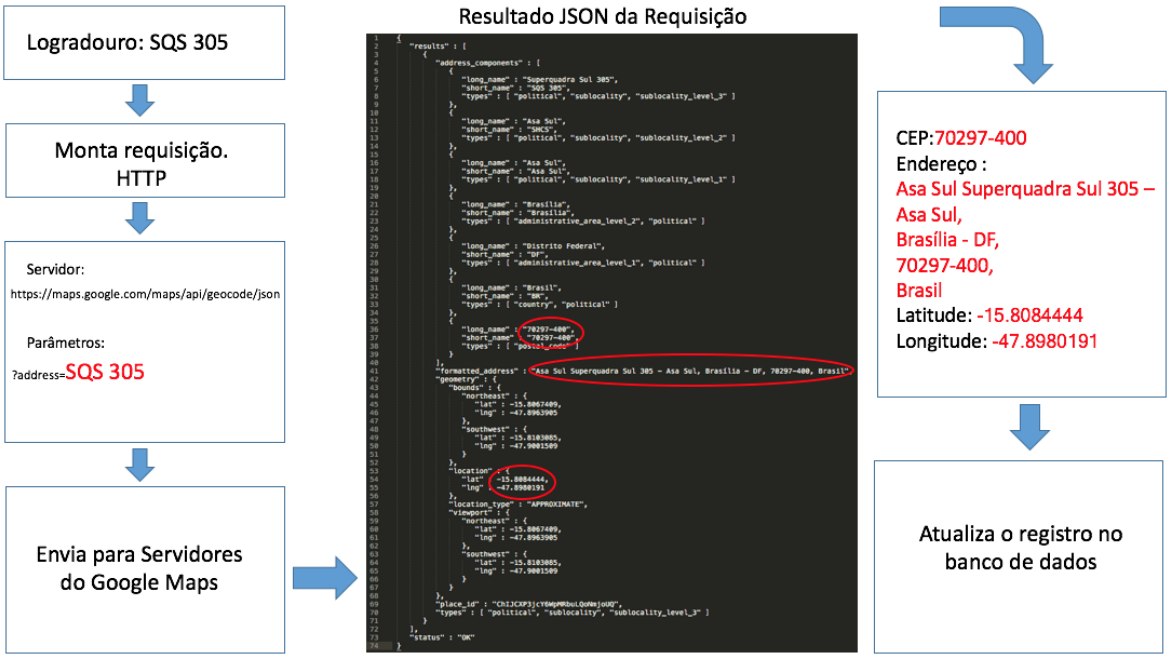


Figura 15: Fluxo de funcionamento do geocodificação dos registros. Autoria Própria

Para otimizar a geocodificação dos logradouros foi adicionada uma camada de consistência que elimina os logradouros que já foram geocodificados em algum momento pelo motor, diminuindo a quantidade de requisições ao serviço do *GoogleMaps*. Esta camada é muito importante, uma vez que o serviço do *GoogleMaps* apresenta custo financeiro quando mais de 1500 requisições são enviadas no mesmo dia, por conta desta limitação foi criada esta funcionalidade, bem como um limitador que impede que mais de 1500 registros sejam enviados no mesmo dia.

3.1.4 Motor de agregação e contagem

O motor de agregação e contagem é o mecanismo mais simples no processo de georreferenciamento, na qual utiliza as funções nativas do banco de agregação e contagem de registros e insere os dados resultantes do processo na tabela do banco dados chamada *points* que foi criada com informações básicas para análises geográficas. Na Tabela 6 pode-se observar a estrutura dos campos utilizados para a execução da etapa em estudo.

Quadro 6: Campos da tabela *point.*

Campo	Tamanho	Tipo de Dados	Obrigatóri	Descrição
ID	11	Inteiro	SIM	Identificador único do registro
POSTALCODE	45	Texto	SIM	Cep do Logradouro
LATITUDE	45	Texto	SIM	Coordenada de Latitude do cep
LONGITUDE	45	Texto	SIM	Coordenada de Longitude do cep
CITY	255	Texto	NÃO	Cidade onde o caso foi reportado
STATE	45	Texto	SIM	Estado
NEIGHBORHOOD	255	Texto	NÃO	Bairro
YEAR	10	Inteiro	SIM	Ano
QUANTITY	45	Inteiro	SIM	Quantidade

Uma vez que o motor de geocodificação, termina o processo, ele atualiza o *status* do registro na tabela sinan. Assim o motor de agrupamento e contagem quando acionado passa a procurar outros registros com o mesmo CEP, e passa a incrementar a quantidade de casos por ano na tabela *points*.

3.1RECURSOS TECNOLÓGICOS

O sistema foi desenvolvido inteiramente nos sistemas operacionais *Ubuntu* 16 e *MacOsx Mojave*, utilizando linguagem PHP na versão 7. Foram também utilizados os frameworks *Synfony* na sua versão 3.0, bibliotecas *javascript jquery* na versão 3.0 e as bibliotecas de CSS *bootstrap*. Foram também utilizados no desenvolvimento deste projeto licença educacional do software PHP STORM na sua versão 2016.3 bem como a licença educacional do software TABLEAU DESKTOP na versão 2018.1, gentilmente cedidas pelas fabricante.

3.2 DELIMITAÇÃO DO ESTUDO

O presente trabalho visou a elaboração de um conjunto de painéis com dados de casos do Distrito Federal. As análises dos dados foram limitadas aos dados do SINAN. Aqui não se objetivou fazer análise semântica sobre a tecnologia do *Business Intelligence*. A metodologia do trabalho também não objetivou descrever detalhadamente como foi construído cada painel, bem como o uso da ferramenta de *Business Intelligence* Tableau.

4. RESULTADOS

4.1.1 Visão Geral

O estudo foi realizado no período de fevereiro de 2017 a agosto de 2018 gerando resultados finais. Os primeiros resultados foram publicados na plataforma *AedesMaps*. Onde pode ser acessada na rede mundial de computadores através do seguinte endereço: http://www.aedesmaps.com.br/. Foi enviado um artigo intitulado **Geocoding Dengue Cases for Spatial Analysis** (anexo 1) utilizado como referência nesta pesquisa, sendo aceito, apresentado e publicado no Congresso Internacional de Engenharia Clínica e Gestão da Tecnologia da Saúde – II (ICEHTMC 2017).

4.1.2 Geocodificação

Após o final do processo de geocodificação dos registros do SINAN, foi realizado o agrupamento dos registros pela dimensão de ano. Chamou atenção o fato dos registros não demonstrarem uma melhora significativa de qualidade ao decorrer dos anos. Eles mantiveram uma taxa de geocodificação média de 82 %. Este fato demonstra que os endereços nas fichas de notificação continuam ano após ano sendo preenchidas de maneira incorreta com dados incompletos. Entretanto no ano de 2015 este indicador subiu para 91% o que pode indicar uma conscientização por parte dos profissionais de saúde em relação a qualidade das informações envidadas. Na tabela 7 podemos acompanhar como foi a taxa de geocodificação em cada um dos anos.

Quadro 7: Percentual de Geocodificação do motor de mineração por ano.

Ano	Total de Registros	Registros Geocodificados	Taxa de Geocodificação
2010	20.896	14.633	70,0%
2011	7.071	5.888	83,3%
2012	3.851	3.224	83,7%
2013	22.490	18.777	83,5%
2014	20.493	16.488	80,5%
2015	12.295	11.218	91,2%

Fonte: SINAN

Neste estudo também foi analisado a perspectiva do tamanho do logradouro, onde foi estudado se existe alguma relação entre o tamanho do logradouro com sucesso da geocodificação. Podemos observar que os logradouros com tamanho 10 e 15 caracteres, foi o grupo de registros que obteve o melhor resultado, com 90,8% de sucesso na geocodificação. Já os registros com tamanho entre 26-30 caracteres, obtiveram uma taxa de

sucesso de apenas 73,8%, bem distante da média global do processo que foi de 82%. Em teoria os logradouros maiores deveriam fornecer mais informações e isto facilitaria sua localização na base cartográfica, na prática não é isto o que acontece.

Uma hipótese é que os endereços do Distrito Federal seguem um padrão diferente do adotado no resto do Brasil, utilizando siglas e separando os logradouros por setores, aparentemente isto torna mais fácil a localização dos logradouros pela API de geocodificação do *Google Maps*. Na Tabela 8 as taxas de geocodificação foram detalhadas de acordo com seu tamanho de caracteres.

Quadro 8: Taxa de Geocodificação pelo tam. de caracteres do logradouro.

Tamanho de caracteres Logradouro	Registros	Taxa de Geocodificação
5	692	84,1%
6	628	94,3%
7	200	75%
8	740	80,5%
9	997	92,7%
10	1481	86,9%
11	2495	90,9%
12	2372	91,6%
13	3119	94,1%
14	4872	93,8%
15	3015	88%
16	4212	79,6%
17	7029	87,4%
18	5294	87,3%
19	3516	86,2%
20	3499	90,3%
21	4210	92%
22	6323	93,3%
23	4620	84,4%
24	3020	79,4%
25	3368	82,7%
26	2532	78,9%
27	1769	69,2%
28	1449	73,8%
29	1370	73,9%
30	1776	73,4%
31	1433	66,4%
32	1234	53,7%
33	1158	69,4%

Fonte: SINAN

4.1.3 Painel de Casos de dengue no Brasil

Com o objetivo de permitir a população e pesquisadores uma visão mais ampla dos casos de dengue em âmbito nacional, o painel de casos de dengue no Brasil, permite análises quantitativas dos casos pela perspectiva de tempo e pela perspectiva geográfica. Com dados já agregados do SINAN e extraídos no período de 2001 a 2012 o usuário pode acompanhar a evolução dos casos por período, ou se desejar, aplicando os filtros de ano, estado e cidade. Neste painel foi utilizado o recurso chamados de *mapa de símbolos* do software Tableau, ele permite diferenciar de maneira gráfica a quantidade de casos. Utilizando um círculo, que fica maior ou menor de acordo com a quantidade de casos em uma determinada cidade, o usuário pode visualizar com maior facilidade as localidades com mais casos de dengue.

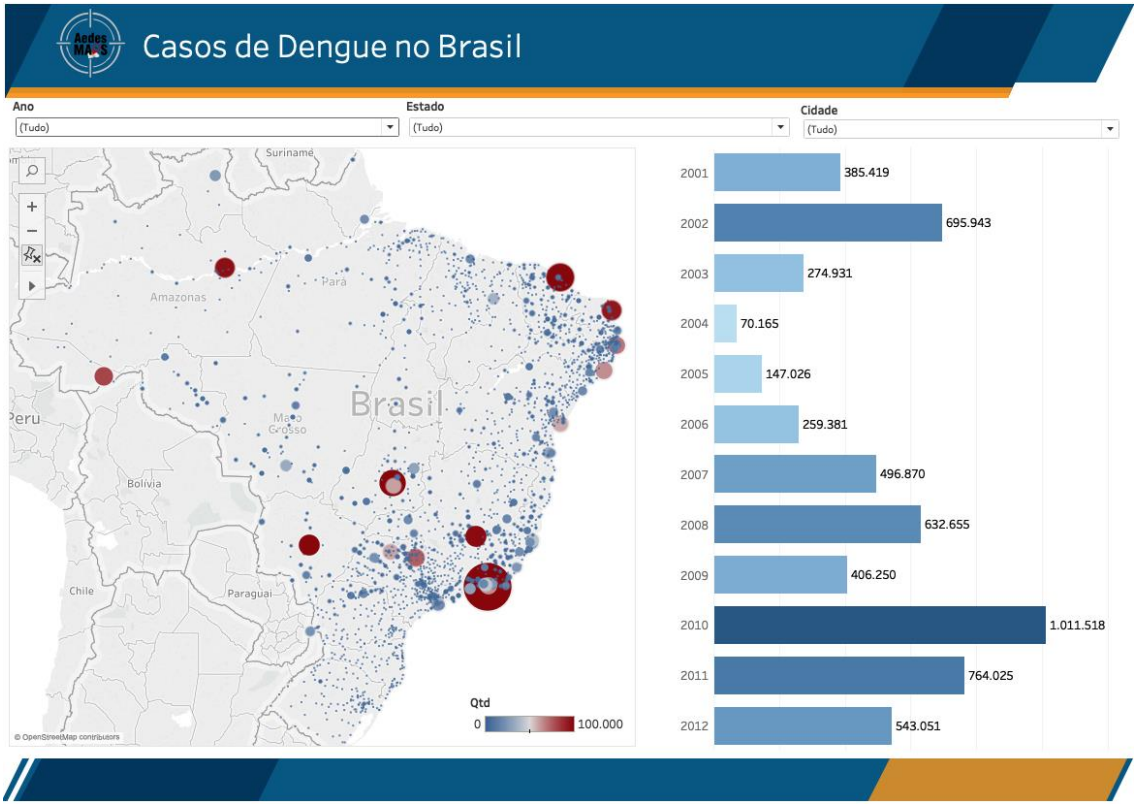


Figura 16: Casos de Dengue no Brasil. Autoria Própria

Foi implementado neste painel a funcionalidade de escala de cores nos símbolos do mapa, onde localidades com mais de 50.000 casos de dengues mudam a tonalidade de azul para vermelho, que se torna mais escura a medida que a quantidade de casos aumenta. Usando os estudos de Bamz (1965) sobre a psicologia das cores e as emoções que elas despertam nas pessoas o nosso objetivo foi aguçar os sentidos e a curiosidade dos usuários focando nos maiores pontos de interesse do painel dentro do mapa geográfico.

4.1.4 Painel de Análises de Indicadores Epidemiológicos

Já no painel de análises de indicadores epidemiológicos o objetivo é permitir ao usuário investigar do ponto de vista geográfico como se comportam algumas variáveis epidemiológicas dos dados do SINAN. Os filtros disponibilizados permitem que o usuário cruze vários tipos de informações. Também pode por exemplo filtrar o sexo do paciente em um determinado ano e visualizar em quais regiões mais homens ou mulheres foram mais afetados pela dengue.

Utilizando técnicas avançadas de visualização de dados, o painel exibe um circulo no mapa maior ou menor de acordo com a densidade de casos. Isto permite que o usuário encontre padrões nos casos, possibilitando assim investigações cada vez mais sofisticadas. Na barra superior, a cada nova seleção que o usuário realiza, o sistema atualiza automaticamente a quantidade de registros afetados pelos filtros ou seleções geográficas realizadas pelo usuário. Este recurso ajuda o acompanhamento do indicador da quantidade de casos de maneira muito mais simples e dinâmica. Um outro facilitador acrescentado ao painel é a janela de detalhes que abre um gráfico secundário ao usuário rolar o ponteiro do mouse sobre um ponto do mapa. Este gráfico exibe o quantitativo de casos por cada centro de saúde de notificação. Esta informação permite ao usuário analisar onde os pacientes foram tratados.

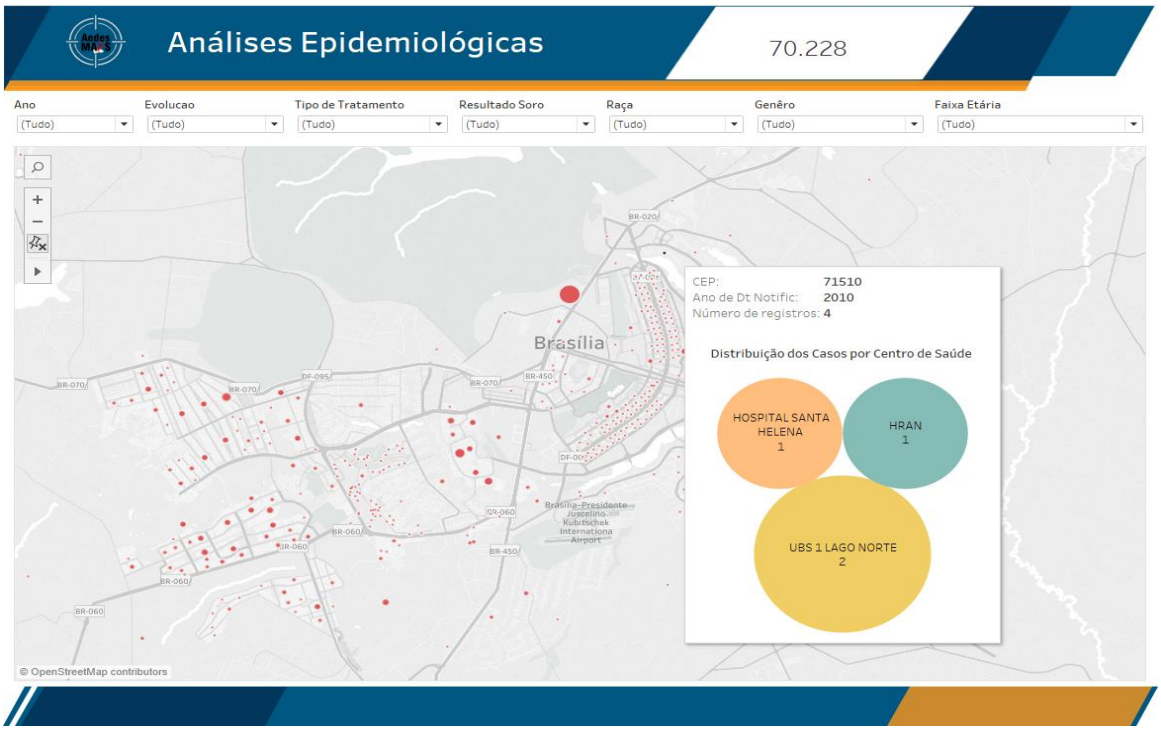
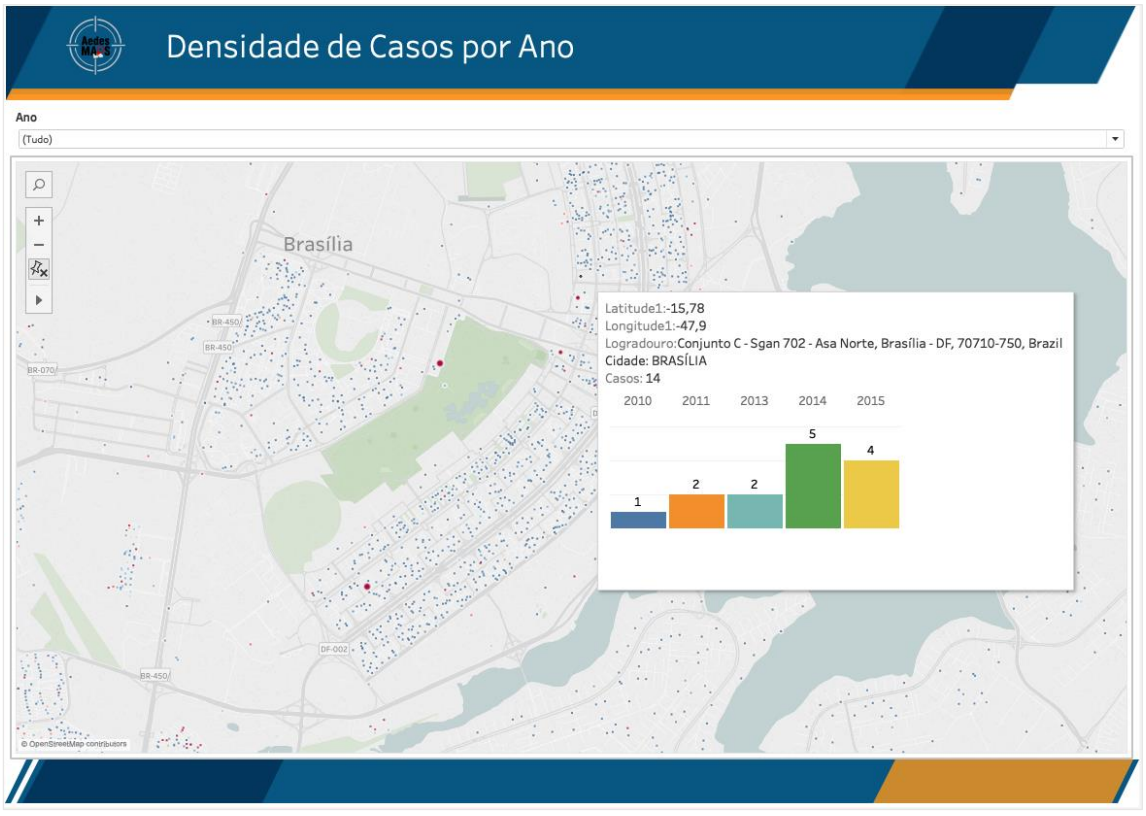


Figura 17: Análises Epidemiológicas. Autoria Própria

Uma outra funcionalidade deste painel é que o agrupamento dos casos é realizado pelo CEP até divisor de subsetor, ou seja, os 5 primeiros números do CEP são usados como chave de agrupamento dos casos. Esta configuração permite por exemplo agrupar todas as ruas de uma determinada quadra ou setor. Isto auxilia a identificar os pontos mais críticos dentro de uma determinada localidade.

4.1.5 Painel de Densidade de Casos por ano

No painel de densidade o objetivo é ir ainda mais fundo na análise dos casos, aqui é possível analisar quantos casos tiveram em um bloco ou rua em particular ao longo do tempo. Neste mapa o usuário pode selecionar um ou mais anos e utilizando a mesma tecnologia de visualização de dados do painel anterior. Também permite que cada ponto geográfico aumente ou diminua de tamanho de acordo com a quantidade de casos. Um diferencial neste painel é que além do ponto geográfico aumentar de acordo com a quantidade, também é sensibilizado através de uma gradação entre tons de azul e vermelho, onde permite o usuário visualizar rapidamente distorções na quantidade de casos em alguma localidade.



.

Figura 18: Densidade de Casos por Ano. Autoria Própria

Ao rolar o mouse sobre qualquer um dos pontos geográficos o painel, exibe ao usuário um subpainel com informações detalhadas do ponto, como quantidade acumulada de casos e um gráfico com histórico de casos ao longo do tempo, como podemos visualizar na figura 18. Ao analisar os dados deste painel alguns fatos nos chamaram atenção, como o caso da cidade satélite de Planaltina, onde ao se analisar os casos de uma localidade chamada de setor tradicional no CEP 73330, encontramos 211 casos acumulados ao longo dos 5 anos. Ao analisar esta região geográfica no mapa de densidade de casos que é mais detalhado, vimos que 40% dos casos se concentraram em apenas três ruas desta localidade, todas próximas ao Centro de Ensino Fundamental 02 de Planaltina. Isto é preocupante uma vez que uma escola tem grande densidade de crianças que passam o período diurno, geralmente com as pernas e braços descobertos, que são justamente os locais e horário onde o mosquito *aedes* geralmente pica as pessoas. Comportamento similar também acontece na cidade satélite de Candangolândia, onde dos 320 casos analisados para o CEP 71725, 85% dos casos estão em até 3 ruas de distância de escolas de ensino fundamental, Escola Classe 1 de Candangolândia, Centro de Ensino Fundamental 1 de Candangolândia e Escola Classe 2 de Candangolândia. Na cidade satélite da Ceilândia no CEP 72236 com 301 casos, localizado no setor P.sul o percentual foi um pouco menor, com aproximadamente 31% dos casos, próximos a Escola Classe 52 de Ceilândia.

4.1.6 Análises de Distribuição de Indicadores Epidemiológicos por faixa etária

Investigar padrões é um dos principais objetivos deste trabalho, analisar como os casos se comportam do ponto de vista geográfico é muito importante, mas também é vital analisar como os casos se encontram distribuídos nos grupos populacionais do Distrito Federal. Claramente ao se estudar a distribuição dos casos pelo sexo dos pacientes podemos notar que de maneira geral os homens são o grupo populacional mais afetado pela dengue. Pudemos observar nas análises que o grupo masculino é em quase todas as faixas etárias aproximadamente 90% maior que o grupo feminino. Outro fato interessante é que a distribuição dos casos através da pirâmide populacional do IBGE para o Distrito Federal apresenta um percentual da população feminina muito próxima da população masculina, em maioria dos casos é até maior. Ainda assim os homens são mais infectados. Em estudo conduzido por Anker (2011) na Ásia em países tropicais muito similares ao nosso, a taxa de casos por gênero apresenta uma taxa de casos com diferença de gênero mínima.(ANKER, 2011)

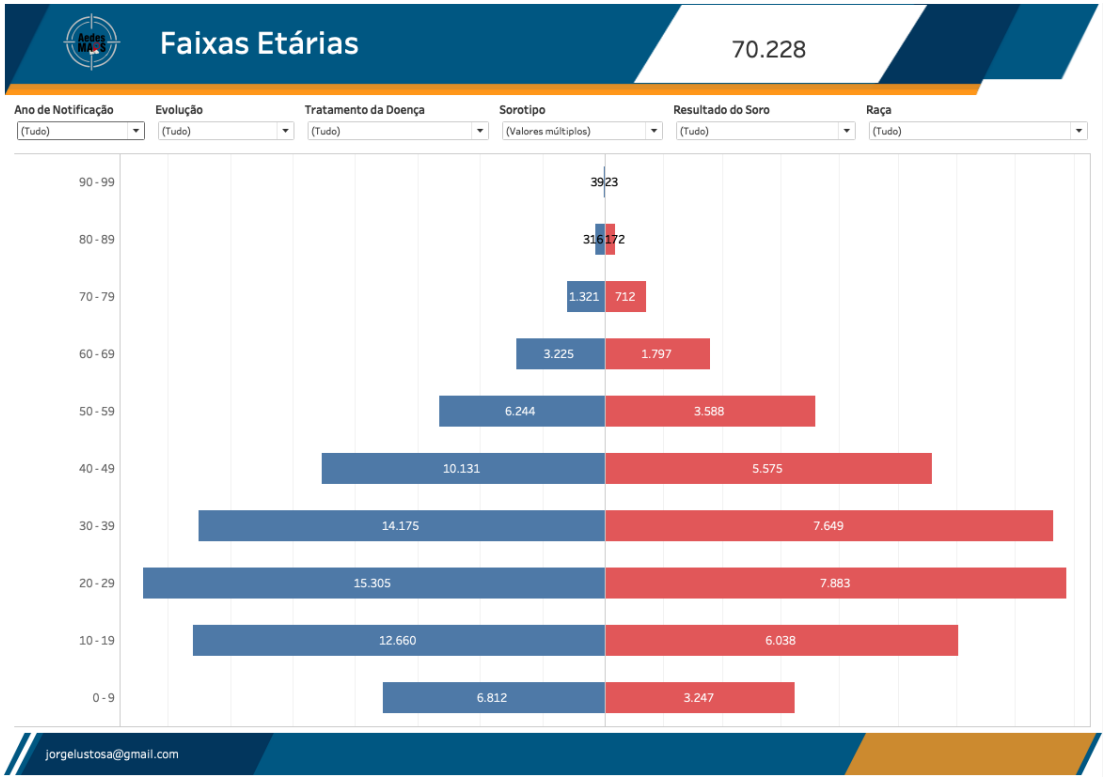


Figura 19: Painel de Indicadores Epidemiológicos por faixa etária. Autoria Própria

4.1.7 Painel do processo de mineração

Este painel tem um aspecto mais técnico e o seu propósito é demonstrar o processo de mineração dos dados. O painel demonstra a distribuição dos registros minerados usando o algoritmo de Levenshtein que já foi descrito com mais detalhes na metodologia. Neste indicador quanto menor o valor numérico melhor, indicando um alto grau de similaridade entre o logradouro do SINAN pesquisado em relação aos logradouros da base nacional de CEP.

No processo de mineração usando o algoritmo de Levenshtein 26,75% dos registros não apresentaram similaridade alguma, 22.614 casos de dengue não puderam ser localizados neste processo. Mas em segundo processo de mineração, onde foi utilizado um processo de melhoria dos dados, através da aplicação de um dicionário de troca de abreviações e análise pela base cartográfica do google, 8.313 registros puderam ser georreferenciados com alta qualidade. Isto representa que 36,76% dos registros de baixa qualidade ainda puderam de alguma maneira ser aproveitados no georreferenciamento.

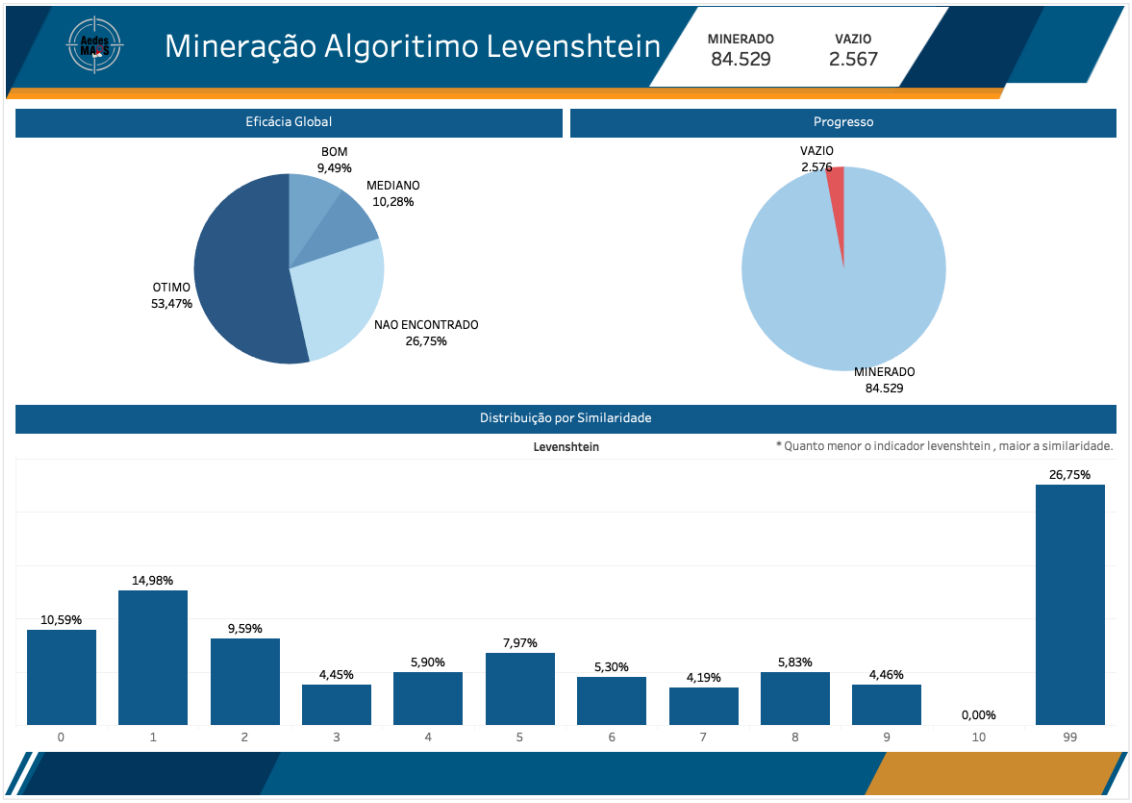


Figura 20: Mineração utilizando algoritmo de Levensthein. Autoria Própria

4.1.8 Painel de Nuvem de Palavras

A nuvem de palavras ou *cloud tags* como é conhecido no idioma inglês. Trata-se de uma técnica de análise de dados popular que teve uma grande adoção na análise grandes quantidades de dados não estruturados. Neste painel foi utilizado o campo de observações da ficha de notificação para demonstrar as principais reclamações e observações relatadas pelos profissionais de saúde no momento do atendimento ao paciente. Neste painel podemos observar que cefaleia, dor, mialgia e febre são as observações mais frequentes detectadas pelos profissionais de saúde ao se atender os pacientes de dengue, como é demonstrado na figura 21.

Um outro fato que chamou atenção durante o processo de análises é que nas observações inseridas na ficha de notificação individual a maioria dos profissionais de saúde que preenchem as fichas, não se preocupa com a pontuação correta das palavras, isto produz em muitos casos palavras iguais com escritas diferentes. Provavelmente padronizar os termos ou as palavras poderia ser uma solução interessante e permitiria por exemplo analisar dentro de uma região geográfica ou do ponto de vista epidemiológico quais são as observações e sintomas mais recorrentes.

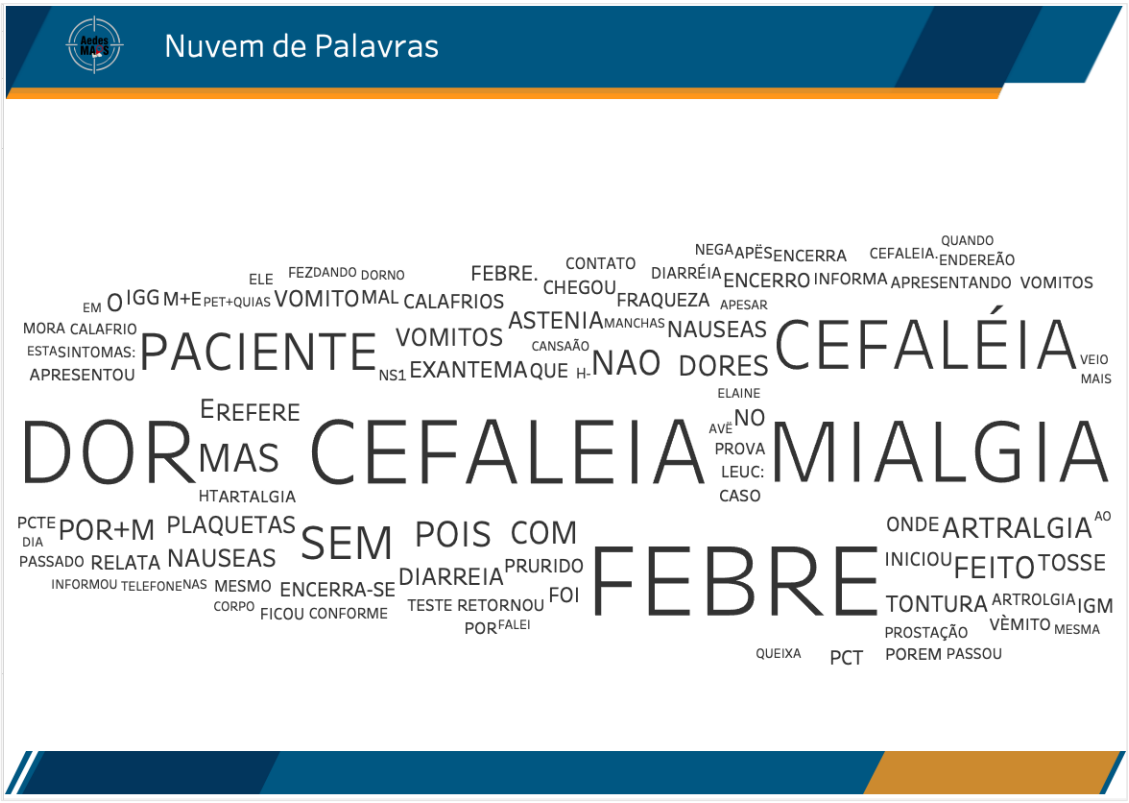


Figura 21: Painel de Nuvem de palavras. Autoria Própria

5. DISCUSSÃO E CONCLUSÃO

A problemática do mosquito *Ae. aegypti* deve ser encarado como uma questão social e não apenas de saúde pública, deve ser uma prioridade para o governo e para todas as outras áreas. Várias ferramentas e ações no combate ao vetor não devem ser deixadas de lado ou negligenciadas. A população deve continuar sendo parceira, se organizando por meio da educação ambiental que começa em sua própria moradia. Os dados epidemiológicos demonstram que o problema está em cada rua, em cada quadra, em cada bloco e que cada esforço empreendido faz diferença.

Kosonen (2014) em seu artigo intitulado "*User Motivations and knowledge sharing in Idea crowdsourcing*", relata que o *Crowdsourcing* (contribuição colaborativa ou colaboração coletiva), é um método complementar no auxílio a métodos especializados acerca da resolução de problemas. Destaca a contribuição da população abordando experiências que não poderiam ser mostradas apenas com a pesquisa cientifica, demostrando os potenciais benefícios da coletivização.

Outra grande contribuição para os estudos na área da saúde são os dados disponibilizados pelo SINAN. São dados extremamente importantes, muito ricos e cheios de detalhamentos. Esta base permite uma série de análises, possibilitando o cruzamento de informações principalmente associadas a Dengue. Todavia, devido lacunas no preenchimento dos dados de logradouro, dificulta-se as análises geográficas, necessitando realizar o processo de mineração de dados. Os dados estão em sua forma bruta tem muitas inconsistências advindas da escassez ou ineficiência de recursos humanos para a coleta desses dados.

Davis e Fonseca (2007) em seu artigo "*Assessing the certainty of locations produced by an address geocoding system*" propuseram uma metodologia para análise de endereço em texto livre, considerando padrão de localização brasileiro. Em seu artigo, os autores propuseram três etapas de análise. O primeiro estágio consiste em identificar os componentes do endereço e organizá-los em uma estrutura de dados apropriada. Estágio de correspondência que consiste em procurar o valor no banco de dados de endereços para procurar uma correlação melhor entre os valores identificados e os dados presentes no banco de dados. A última etapa extrai as coordenadas correspondentes do banco de dados de referência, gerando três indicadores: Indicador de certeza de análise (PCI), Indicador de certeza de correspondência (MCI) e Indicador de certeza de localização (LCI). Neste estudo foi utilizada o conceito de vários estágios para mineração dos endereços, e a base de referência foi base do CEP, com o diferencial do uso do algoritmo de Levenshtein para a pesquisa de endereços similares na base de referência.

Eisen e Lozano-Fuentes (2009), defendem o uso de um SIG como solução para o mapeamento e modelagem para uso em áreas endêmicas, especificamente relacionada a dengue. Eles avaliaram vários tipos de modelagem espacial e mapeamento utilizando dados epidemiológicos de fontes oficiais. E obtiveram uma resposta positiva sobre SIG e mapeamento de áreas de risco, reforçando que seu uso facilita o combate ao *Ae. aegypti*, por fazer esse monitoramento.

Um exemplo prático é o Sistema de Informação Geoestatístico e Sonoro da Dengue (SIGESON-DENGUE) que é um sistema construído por meio de um algoritmo que, obtém e processa as imagens das Ovitrampas (armadilhas onde são depositados os ovos do mosquito) e automatiza a contagem desses ovos georreferenciando-os (BRASIL, 2015). Anterior ao SIGESON-DENGUE, tem-se o Sistema de Aquisição e Processamento de Imagens de Ovitrampas (SAPIO) e o Sistema de Informação Geográfica para Ovitrampas da

Dengue (SIGO-DENGUE) (SILVA, 2013b). Nesse projeto foi apresentado a automatização da contagem dos ovos depositados pelo mosquito Aedes aegypti (portador do Dengue) em Ovitrampas (armadilhas) por meio de um Sistema para o Processamento Digital de Imagens (SPDI), bem como a implementação de um SIG para auxiliar no acompanhamento estatístico da proliferação da Dengue, na Capital Federal (SILVA et al., 2013).

Silva, (2016) em seu estudo constatou que no ano de 2013, cidades como Ceilândia, Varjão, Fercal, Brazlândia, (Scia) Estrutural, foram às cidades que estiveram mais frequentemente no grupo de pior situação social e econômica, indicando também uma maior taxa de incidência de casos de dengue nessas regiões. Com exceção da Ceilândia e Fercal as demais apresentam boa cobertura de esgotamento sanitário, coleta de lixo e abastecimento de água, em comparação a cidades como Park Way e Jardim Botânico, onde no caso há pouco serviço de saneamento básico oferecido pelo Estado, porém, nessas regiões consideradas de maior poder aquisitivo a incidência e casos de Dengue são menores, e por possuírem um nível populacional menor e pouco déficit de saneamento ser mais fácil de controlar devido as melhores infraestruturas de moradia. Entretanto, outras regiões consideradas de alta renda também tiveram altas taxas de incidência da doença.

Um outro ponto observado durante fase de análises geográficas dos dados foi a questão da grande quantidade de casos próximos as escolas, em alguns casos concentrando entre 30% e 80% dos casos de uma localidade num raio de aproximadamente 500 metros destas escolas. O caso da cidade satélite de Planaltina, onde uma localidade chamada de setor tradicional no CEP 73330, foi encontrado 211 casos acumulados ao longo dos 5 anos. Ao analisar esta região geográfica no mapa de densidade de casos, foi constatado que 40% dos casos se concentraram em apenas três ruas desta localidade, todas próximas ao Centro de Ensino Fundamental 02 de Planaltina. Comportamento similar também acontece na cidade satélite de Candangolândia, onde dos 320 casos analisados para o CEP 71725, 85% dos casos estão em até 3 ruas de distância de escolas de ensino fundamental, Escola Classe 1 de Candangolândia, Centro de Ensino Fundamental 1 de Candangolândia e Escola Classe 2 de Candangolândia. Na cidade satélite da Ceilândia no CEP 72236 com 301 casos, localizado no setor P.sul o percentual foi um pouco menor, com aproximadamente 31% dos casos, próximos a Escola Classe 52 de Ceilândia. Durante a pesquisa não foram encontrados outros estudos sobre este tema, isto dado preocupante levando-se em consideração que no período estudado quase 30.000 crianças e adolescentes tiveram casos de dengue.

Minerar os dados dos casos através da perspectiva geográfica, analisar os dados minerados e georreferenciados, usando ferramentas de *Business Intelligence*. Construindo painéis que utilizam técnicas avançadas de visualização de dados, permite investigar padrões volumétricos e geográficos dos casos de dengue. Permitindo assim, um combate muito mais efetivo ao mosquito *Ae. Aegypti*.

6. TRABALHOS FUTUROS

- Otimizar os recursos disponíveis para automatização de Ovitrampas e integrá-las à plataforma *AedesMaps.*
- Analisar os dados advindos das ovitrampas sejam elas manuais ou automatizadas;
- Melhoria dos dados do Sinan, através da conscientização dos profissionais de saúde, sobre a importância de se preencher o endereço corretamente.
- Minerar dados dos anos de 2016 e 2017.
- Criar um novo mapa para casos próximos as escolas
- Melhorar algoritmo de mineração de dados para identificar os 11.000 casos que apresentaram falha no processo de *matching*.
- Minerar os microdados de todos os estados do Brasil.
- Criar mapa com casos de todo Brasil.

REFERÊNCIAS BIBLIOGRÁFICAS

ANKER, M.; ARIMA, Y. Male-female differences in the number of reported incident dengue fever cases in six Asian countries. **Western Pacific Surveillance and Response**, 2011.

ARANHA, F. Atlas dos setores postais: uma nova geografia a serviço da empresa. **RAE-revista de administração de empresas**, v. 37, n. 3, p. 20–27, 1997.

BAMZ, J. **Arte y ciencia del color**. [s.l.] Las Ediciones del Arte, 1965.

BANSAL, S. et al. Big data for infectious disease surveillance and modeling. **Journal of Infectious Diseases**, v. 214, n. November, p. S375–S379, 2016.

BESERRA, E. B. et al. Ciclo de vida de Aedes (Stegomyia) aegypti (Diptera: Culicidae) em águas com diferentes características. **Iheringia Sér Zool**, v. 99, n. 3, p. 281–285, 2009.

BOWMAN, L. R.; RUNGE-RANZINGER, S.; MCCALL, P. J. **Assessing the Relationship between Vector Indices and Dengue Transmission: A Systematic Review of the EvidencePLoS Neglected Tropical Diseases**, 2014.

BRASIL, C. **Estrutura do CEP**. Disponível em: <https://www.correios.com.br/precisa-de-ajuda/o-que-e-cep-e-por-que-usa-lo/estrutura-do-cep>. Acesso em: 31 out. 2018a.

BRASIL, M. DA S. Boletim Epidemiológico. **Boletim Epidemiológico Secretaria**, v. 48, n. 11, p. 1–10, 2017b.

BRITO, L. S. F. DE. Sistema de informaçöes de agravos de notificação-SINAN. **Anais do Seminário Nacional de Vigilância Epidemiológica**, p. 145–146, 1993.

CÂMARA, G. et al. TerraLib, Tecnologia Brasileira de Geoinformação: para quem e para quê? **Informática Pública**, v. 4, n. 1, 2002.

COUTO, A. V. DO. Uma abordagem de gerenciamento de redes baseado no monitoramento de fluxos de tráfego netflow com suporte de técnicas de business intelligence. 2012.

DA SILVA, D. et al. INTELIGÊNCIA DE NEGÓCIO. **Maiêutica-Tecnologias da Informação**, v. 1, n. 01, p. 746, 2016.

DAVIS, C. A.; FONSECA, F. T. Assessing the certainty of locations produced by an address geocoding system. **GeoInformatica**, v. 11, n. 1, p. 103–129, 2007.

DOS SANTOS PIMENTEL, B. et al. Um sistema de suporte à decisão para análise dos resultados do Enade. 2016.

EISEN, L.; LOZANO-FUENTES, S. Use of Mapping and Spatial and Space-Time Modeling Approaches in Operational Control of Aedes aegypti and Dengue. **Plos Neglected Tropical Diseases**, v. 3, n. 4, p. 7, 2009.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **Al Magazine**, 1996.

FERREIRA-DE-BRITO, A. et al. First detection of natural infection of Aedes aegypti with Zika virus in Brazil and throughout South America. **Mem Inst Oswaldo Cruz Rio de Janeiro**, v. 111, n. 10, p. 655–658, 2016.

HAGERTY, J.; SALLAM, R. L.; RICHARDSON, J. Magic Quadrant for Business Intelligence Platforms. **Gartner for Business …**, n. February, p. 1–52, 2011.

HERNÁNDEZ-ÁVILA, J. E. et al. Nation-Wide, Web-Based, Geographic Information System for the Integrated Surveillance and Control of Dengue Fever in Mexico. **PLoS ONE**, v. 8, n. 8, p. e70231, 6 ago. 2013.

HORSTICK, O.; MORRISON, A. C. Dengue Disease Surveillance: Improving Data for Dengue Control. **PLoS Neglected Tropical Diseases**, v. 8, n. 11, 2014.

KEOGH, B. E. et al. Public confidence and cardiac surgical outcome. **BMJ**, v. 316, p. 1759–1760, 1998.

KITTAYAPONG, P. et al. Suppression of dengue transmission by application of integrated vector control strategies at sero-positive GIS-based foci. **American Journal of Tropical Medicine and Hygiene**, v. 78, n. 1, p. 70–76, 2008.

KRAEMER, M. U. G. et al. The global distribution of the arbovirus vectors Aedes aegypti and Ae. Albopictus. **eLife**, v. 4, n. JUNE2015, 2015.

LAGUARDIA, J. et al. Sistema de Informação de Agravos de Notificação (Sinan): desafios no desenvolvimento de um sistema de informação em saúde. **Epidemiologia e Serviços de Saúde**, v. 13, n. 3, p. 135–147, 2004.

LENTZ, A. **LEVENSHTEIN MySQL stored function | Open Query Pty Ltd**. Disponível em: <https://openquery.com.au/blog/levenshtein-mysql-stored-function>. Acesso em: 22 jul. 2018.

LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions, and reversals. **Soviet Physics Doklady**, v. 10, n. 8, p. 707–710, 1966.

LIMA, D. F. **AedesMaps : Uma Proposta para o controle de doenças transmitidas pelo mosquito Aedes aegypti**, 2018.

LITE, M. E. ANÁLISE DA CORRELAÇÃO ENTRE DENGUE E INDICADORES SOCIAIS A PARTIR DO SIG-ANALYSIS OF CORRELATION BETWEEN DENGUE AND SOCIAL INDICATORS FROM GIS. **Hygeia**, v. 6, n. 11, 2010.

LUSTOSA, J. L. S. et al. **Geocoding dengue cases for spatial analysis**. International Clinical Engineering and Health Technology Management Congress - ICEHTMC. **Anais**...São Paulo: ICEHTMC, 2017Disponível em: <http://www.icehtmc.com/>

LYALL, F. **International communications: The international telecommunication union and the universal postal union**. [s.l: s.n.].

MARTINS, D.; DAVIS, C. A.; FONSECA, F. T. Geocodificação de endereços urbanos com indicação de qualidade. **Proceedings of the Brazilian Symposium on GeoInformatics**, p. 36–41, 2012.

MINISTÉRIO DA SAÚDE. Diretrizes Nacionais para a Prevenção e Controle de Epidemias de Dengue. **Secretaria de Vigilancia em Saúde**, p. 162, 2009.

NELSON, M. J. Aedes aegypti: biologia y ecologia. 1986.

NETO, V. C. et al. Desenvolvimento e integração de mapas dinâmicos georreferenciados para

o gerenciamento e vigilância em saúde. **Journal of health informatics**, v. 6, n. 1, 2014.

NIVALDO MARIANO CARVALHO,DAVID GALDÊNCIO FERREIRA,MOISÉS ERICKISON BRITO DE ARAÚJO, R. R. DE L. PROJETO DE ANÁLISE DE DADOS PARA IMPLANTAÇÃO DE DATA MART COMO FERRAMENTA PARA TOMADA DE DECISÃO EM COMBATE AOS VÍRUS DA DENGUE, ZIKA E CHIKUNGUNYA. **Interscientia**, v. 5, p. 106–123, 2017.

PANG-NING, T.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. [s.l: s.n.].

PENA, R. F. A. **SIG. Sistema de Informações Geográficas - Brasil Escola**. Disponível em: <https://brasilescola.uol.com.br/geografia/sig.htm>. Acesso em: 26 jul. 2018.

PRIMAK, F. V. **Decisões com bi (business intelligence)**. [s.l.] Fabio Vinicius Primak, 2008.

REGIS, L. N. et al. Sustained Reduction of the Dengue Vector Population Resulting from an Integrated Control Strategy Applied in Two Brazilian Cities. **PLoS ONE**, v. 8, n. 7, 2013.

SILVA, M. M. DA. Rastreamento Do Foco Do Aedes Aegypti De Informações Geográficas No Distrito Federal. 2013.

SILVA, J. L. DE S. Dengue : Uma análise de casos no Distrito Federal e suas condições de Saneamento Básico no ano de 2013 Jéssika Luana de Souza Silva Dengue : Uma análise de casos no Distrito Federal e suas condições de Saneamento Básico no ano. 2016.

TABLEAU. Visual Analysis Best Practices. **Tableau Software**, p. 41, 2014.

TURBAN, E. et al. **Business Intelligence: um enfoque gerencial para a inteligência do negócio**. [s.l.] Bookman Editora, 2009.

WHO. Global Strategy for Dengue Prevention and Control 2012–2020. **World Health Organiszation**, p. 43, 2012.

WORLD HEALTH ORGANIZATION (WHO) REGIONAL OFFICE FOR SOUTH-EAST ASIA. Dengue: guidelines for diagnosis, treatment, prevention, and control. **Special Programme for Research and Training in Tropical Diseases**, p. 147, 2009a.

WORLD HEALTH ORGANIZATION (WHO) REGIONAL OFFICE FOR SOUTH-EAST ASIA. **Dengue: guidelines for diagnosis, treatment, prevention, and controlSpecial Programme for Research and Training in Tropical Diseases**. [s.l: s.n.].

SÍRIO-LIBANÊS
ENSINO E PESQUISA

II ICEHTMC
SEP 21st - 22nd | 2017
SÃO PAULO - BRAZIL

Certificamos que o trabalho

GEOCODING DENGUE CASES FOR SPATIAL ANALYSIS

dos autores: JORGE; DELMIRA FERREIRA; LUCIANA CáSSIA ARAúJO DE; LOURDES MATTOS; MARCOS TAKASHI, foi apresentado, na modalidade Apresentação Oral, no evento II Congresso Internacional de Engenharia Clínica e Gestão de Tecnologia em Saúde ocorrido de 21 a 23 de setembro de 2017 no Instituto Sírio-Libanês de Ensino e Pesquisa em São Paulo/SP.

São Paulo, 23 de setembro de 2017

Ernesto Iadanza
Chair of IFMBE/CED

Dr. Saide Jorge Calil
Congress Chair

Dr. Marcello Dias Bonfim
Manager - Clinical Engineering
Hospital Sírio-Libanês

SÍRIO-LIBANÊS
ENSINO E PESQUISA

II ICEHTMC
SEP 21st - 22nd | 2017
SÃO PAULO - BRAZIL

Certificamos que o trabalho

GEOCODING DENGUE CASES FOR SPATIAL ANALYSIS

foi apresentado na modalidade Apresentação Oral, por Jorge Luis da Silva Lustosa, no evento II Congresso Internacional de Engenharia Clínica e Gestão de Tecnologia em Saúde ocorrido de 21 a 23 de setembro de 2017 no Instituto Sírio-Libanês de Ensino e Pesquisa em São Paulo/SP.
São Paulo, 23 de setembro de 2017

Ernesto Iadanza
Chair of IFMBE/CED

Dr. Saide Jorge Calil
Congress Chair

Dr. Marcello Dias Bonfim
Manager - Clinical Engineering
Hospital Sírio-Libanês

Certificamos que o trabalho “SISTEMA ESPECIALISTA PARA DIRECIONAMENTO NA DIFERENCIAÇÃO ENTRE AS PATOLOGIAS CAUSADAS PELO AEDES AEGYPTI” (n.737) de Rodrigo Amaral, Luciana de Sousa, Jorge Lustosa, Henderson Sanchez e Lourdes Brasil foi apresentado no XXV Congresso Brasileiro de Engenharia Biomédica, realizado na cidade de Foz do Iguaçu, Paraná, de 17 a 20 de outubro de 2016.

Prof. Dr. Joaquim Miguel Maia
Presidente do Congresso Brasileiro de Engenharia Biomédica CBEB 2016
Universidade Tecnológica Federal do Paraná - UTFPR

Prof. Dr. Sérgio Santos Mühlen
Presidente da Sociedade Brasileira de Engenharia Biomédica - SBEB
Universidade Estadual de Campinas - UNICAMP

Prof. Dr. Eduardo Tavares Costa
Presidente da Comissão Científica do CBEB 2016
Universidade Estadual de Campinas - UNICAMP

REALIZAÇÃO: SBEB UTFPR

APOIO: FUNTEF-PR, CAPES, CNPq, FUNDAÇÃO ARAUCÁRIA, Unimed, IGUASSU, metrologia, sbis

PATROCÍNIO: ITAIPU, Fiep, IBMP, daan serviços, COPEL, FLIR

PRESTADORAS DE SERVIÇO: equalzee, Naipi Travel, Stand

CÓDIGO FONTE DO DICIONÁRIO DE NORMALIZAÇÃO

```
<?PHP

NAMESPACE ENGINEBUNDLE\COMMAND;

USE SYMFONY\BUNDLE\FRAMEWORKBUNDLE\COMMAND\CONTAINERAWARECOMMAND;
USE SYMFONY\COMPONENT\CONSOLE\INPUT\INPUTARGUMENT;
USE SYMFONY\COMPONENT\CONSOLE\INPUT\INPUTINTERFACE;
USE SYMFONY\COMPONENT\CONSOLE\INPUT\INPUTOPTION;
USE SYMFONY\COMPONENT\CONSOLE\OUTPUT\OUTPUTINTERFACE;
USE GEOCODER\GEOCODER;

CLASS CRAWLOGRADOUROCOMMAND EXTENDS CONTAINERAWARECOMMAND
{
    PROTECTED FUNCTION CONFIGURE()
    {
        $THIS
            ->SETNAME('CRAWLOGRADOURO')
            ->SETDESCRIPTION('...')
            ->ADDARGUMENT('ARGUMENT', INPUTARGUMENT::OPTIONAL, 'ARGUMENT DESCRIPTION')
            ->ADDOPTION('OPTION', NULL, INPUTOPTION::VALUE_NONE, 'OPTION DESCRIPTION')
        ;
    }

    PROTECTED FUNCTION EXECUTE(INPUTINTERFACE $INPUT, OUTPUTINTERFACE $OUTPUT)
    {
        $ARGUMENT = $INPUT->GETARGUMENT('ARGUMENT');


        $EM = $THIS->GETCONTAINER()->GET('DOCTRINE')->GETMANAGER();


       $SQL = "SELECT * FROM DENGUE  WHERE  NM_LOGRADO IS NOT NULL  AND LOGRADOURO_PROVAVEL IS NULL  AND CEP_PROVAVEL IS NULL ORDER BY ID ASC  " ;
        $CONNECTION = $EM->GETCONNECTION();
        $STATEMENT = $CONNECTION->PREPARE($SQL);
        $STATEMENT->EXECUTE();
        $RESULTS = $STATEMENT->FETCHALL();
        $KEY = "AIZASYBYTN-XXXXXXXXXXXXXXXXXXXX" ;


FOREACH ($RESULTS AS $R) {

            $BAIRRO =  "" ;

            $R['NM_LOGRADO'] = PREG_REPLACE ('/[^\P{L}\P{N}]/U', ' ', $R['NM_LOGRADO']);

            // QUADRA E SUAS VARIACOES
            $R['NM_LOGRADO'] = STR_REPLACE("QD ","QUADRA ", $R['NM_LOGRADO'])  ;
            $R['NM_LOGRADO'] = STR_REPLACE("Q ","QUADRA ", $R['NM_LOGRADO'])  ;
            $R['NM_LOGRADO'] = STR_REPLACE("Q. ","QUADRA ", $R['NM_LOGRADO'])  ;
            $R['NM_LOGRADO'] = STR_REPLACE("QD. ","QUADRA ", $R['NM_LOGRADO']) ;
            $R['NM_LOGRADO'] = STR_REPLACE("QD.","QUADRA ", $R['NM_LOGRADO'])  ;

            // CONJUNTO E SUAS VARIACOES
            $R['NM_LOGRADO'] = STR_REPLACE(" CONJ-"," CONJUNTO ", $R['NM_LOGRADO'])  ;
            $R['NM_LOGRADO'] = STR_REPLACE(" CONJ"," CONJUNTO ", $R['NM_LOGRADO'])  ;
            $R['NM_LOGRADO'] = STR_REPLACE(" CJ"," CONJUNTO ", $R['NM_LOGRADO'])  ;
            $R['NM_LOGRADO'] = STR_REPLACE(" COJ "," CONJUNTO ", $R['NM_LOGRADO'])  ;
            $R['NM_LOGRADO'] = STR_REPLACE(" UNTO ","", $R['NM_LOGRADO'])  ;

            // RUA E SUAS VARIACOES
            $R['NM_LOGRADO'] = STR_REPLACE(" R "," RUA ", $R['NM_LOGRADO'])  ;

            //AVENIDA
            $R['NM_LOGRADO'] = STR_REPLACE("AV ","AVENIDA ", $R['NM_LOGRADO'])  ;
            $R['NM_LOGRADO'] = STR_REPLACE("AV. ","AVENIDA ", $R['NM_LOGRADO'])  ;

            //BLOCO
            $R['NM_LOGRADO'] = STR_REPLACE(" BL "," BLOCO ", $R['NM_LOGRADO'])  ;
            $R['NM_LOGRADO'] = STR_REPLACE("BL. ","BLOCO ", $R['NM_LOGRADO'])  ;
            $R['NM_LOGRADO'] = STR_REPLACE(" BL. "," BLOCO ", $R['NM_LOGRADO'])  ;

            //CHACARA
            $R['NM_LOGRADO'] = STR_REPLACE("CHAC ","CHÁCARA ", $R['NM_LOGRADO'])  ;
            $R['NM_LOGRADO'] = STR_REPLACE("CH. ","CHÁCARA ", $R['NM_LOGRADO'])  ;
            $R['NM_LOGRADO'] = STR_REPLACE("CHAC. ","CHÁCARA ", $R['NM_LOGRADO'])  ;
            $R['NM_LOGRADO'] = STR_REPLACE("CHACARA ","CHÁCARA ", $R['NM_LOGRADO'])  ;

            // LIMPA DE ONDE FOI ENCONTRADO PARA FRENTE
            $R['NM_LOGRADO'] = PREG_REPLACE('/ CASA.*/', '', $R['NM_LOGRADO']);
            $R['NM_LOGRADO'] = STR_REPLACE("CASA","", $R['NM_LOGRADO'])  ;

            $R['NM_LOGRADO'] = PREG_REPLACE('/ CS.*/', '', $R['NM_LOGRADO']);
            $R['NM_LOGRADO'] = STR_REPLACE("CS","", $R['NM_LOGRADO'])  ;

            $R['NM_LOGRADO'] = PREG_REPLACE('/ APT.*/', '', $R['NM_LOGRADO']);
            $R['NM_LOGRADO'] = STR_REPLACE("APT","", $R['NM_LOGRADO'])  ;

            $R['NM_LOGRADO'] = PREG_REPLACE('/ APARTAMENTO.*/', '', $R['NM_LOGRADO']);
            $R['NM_LOGRADO'] = STR_REPLACE("APARTAMENTO","", $R['NM_LOGRADO'])  ;

            $R['NM_LOGRADO'] = PREG_REPLACE('/ AP.*/', '', $R['NM_LOGRADO']);
            $R['NM_LOGRADO'] = STR_REPLACE("AP","", $R['NM_LOGRADO'])  ;

            $R['NM_LOGRADO'] = PREG_REPLACE('/ APTO.*/', '', $R['NM_LOGRADO']);
            $R['NM_LOGRADO'] = STR_REPLACE("APTO","", $R['NM_LOGRADO'])  ;

            $R['NM_LOGRADO'] = PREG_REPLACE('/ LOTE.*/', '', $R['NM_LOGRADO']);
            $R['NM_LOGRADO'] = STR_REPLACE("LOTE","", $R['NM_LOGRADO'])  ;

            $R['NM_LOGRADO'] = PREG_REPLACE('/ LT.*/', '', $R['NM_LOGRADO']);
            $R['NM_LOGRADO'] = STR_REPLACE("LT","", $R['NM_LOGRADO'])  ;

            $R['NM_LOGRADO'] = PREG_REPLACE('/ CH.*/', '', $R['NM_LOGRADO']);
            $R['NM_LOGRADO'] = STR_REPLACE("CH","", $R['NM_LOGRADO'])  ;


            // TRATAMENTO PARA BAIRRO NO LOGRADOURO

            // ARAPOANGA
            IF(STRPOS($R['NM_LOGRADO'],"ARAPOANGA") || STRPOS($R['NM_LOGRADO'],"ARAP.") || STRPOS($R['NM_LOGRADO'],"ARAPOANGA.") )
            {

                $R['NM_LOGRADO'] = STR_REPLACE("ARAPOANGA ","", $R['NM_LOGRADO'])  ;
                $R['NM_LOGRADO'] = STR_REPLACE("ARAP. ","", $R['NM_LOGRADO'])  ;
                $R['NM_LOGRADO'] = STR_REPLACE("ARAPOANGA. ","", $R['NM_LOGRADO'])  ;

                $BAIRRO = "ARAPOANGA (PLANALTINA)" ;

            }
```

```
            // VALE DO AMANHECER
            IF(STRPOS($R['NM_LOGRADO'],"AMANHECER")  )
            {


                $R['NM_LOGRADO'] = STR_REPLACE("VALE DO AMANHECER ","", $R['NM_LOGRADO'])  ;
                $R['NM_LOGRADO'] = STR_REPLACE("V AMANHECER ","", $R['NM_LOGRADO'])  ;

                $R['NM_LOGRADO'] = STR_REPLACE("CR ","CONJUNTO RESIDENCIAL ", $R['NM_LOGRADO'])  ;

                $BAIRRO = "VALE DO AMANHECER (PLANALTINA)" ;

            }

            // BURITIS
            IF(STRPOS($R['NM_LOGRADO'],"BURITIS")  )
            {


                $R['NM_LOGRADO'] = STR_REPLACE("V BURITIS ","", $R['NM_LOGRADO'])  ;

                $R['NM_LOGRADO'] = STR_REPLACE(" Q ","QUADRA ", $R['NM_LOGRADO'])  ;
                $R['NM_LOGRADO'] = STR_REPLACE(" QD ","QUADRA ", $R['NM_LOGRADO'])  ;

            }

            $SQL_LOGRADOURO  = "SELECT LOGRADOURO ,
                                       MATCH (LOGRADOURO)
                                       AGAINST ('".STR_REPLACE("'","", STR_REPLACE("-","", $R['NM_LOGRADO'])."' IN BOOLEAN MODE) AS `SCORE` ,  CEP
, LEVENSHTEIN(LOGRADOURO,'".STR_REPLACE("'","", STR_REPLACE("--","", $R['NM_LOGRADO'])."') AS LEVENSHTEIN
                                FROM CEP
                                 WHERE
                                MATCH (LOGRADOURO)
                                       AGAINST ('".STR_REPLACE("'","", STR_REPLACE("-","", $R['NM_LOGRADO'])."' IN BOOLEAN MODE)
                                 AND  LEVENSHTEIN(LOGRADOURO,'".STR_REPLACE("'","", STR_REPLACE("-","", $R['NM_LOGRADO'])."') < 10

                                 " ;

                                IF($BAIRRO != "")
                                {

                                    $SQL_LOGRADOURO  .=  "AND BAIRRO = '".$BAIRRO."' "  ;
                                }

                                $SQL_LOGRADOURO  .= "

                                 AND ESTADO = 'DF'
                                 ORDER BY LEVENSHTEIN ASC  LIMIT 1


                                  " ;

            $STMT_LOGRADOURO = $CONNECTION->PREPARE($SQL_LOGRADOURO);
            $STMT_LOGRADOURO->EXECUTE();
            $RESULTS_LOGRADOURO = $STMT_LOGRADOURO->FETCHALL();


            IF(SIZEOF($RESULTS_LOGRADOURO) > 0 )
            {

                FOREACH ($RESULTS_LOGRADOURO AS $RL) {


                    IF ( ISSET($RL['CEP']) &&  $RL['CEP'] != "")
                    {

                        ECHO "ORIGINAL:".$R['NM_LOGRADO']." - SCORE: ".$RL['SCORE']." - PROVAVEL:".$RL['LOGRADOURO']." CEP: ".$RL['CEP']." -
LEVENSHTEIN:".$RL['LEVENSHTEIN']."\N" ;

                        ECHO $SQL_UPDATE = "UPDATE DENGUE_LIMPA SET LOGRADOURO_PROVAVEL = '".$RL['LOGRADOURO']."' , CRAW = '9'  , CEP_PROVAVEL =
'".$RL['CEP']."' , LEVENSHTEIN= '".$RL['LEVENSHTEIN']."'  WHERE ID = ".$R['ID']."  " ;
                        ECHO "\N" ;
                        $STMT_UPDATE = $CONNECTION->PREPARE($SQL_UPDATE);
                        $STMT_UPDATE->EXECUTE();
                    }
                    ELSE
                    {
                        $SQL_UPDATE = "UPDATE DENGUE_LIMPA SET  CRAW = '9'  , LEVENSHTEIN= '99'  WHERE ID = ".$R['ID']."  " ;
                        $STMT_UPDATE = $CONNECTION->PREPARE($SQL_UPDATE);
                        $STMT_UPDATE->EXECUTE();
                        ECHO "NAO ENCONTRADO NO BANCO SIMILARIDADE\N" ;

                    }

                }

            }
            ELSE
            {

                $SQL_UPDATE = "UPDATE DENGUE_LIMPA SET  CRAW = '9'  , LEVENSHTEIN= '99'  WHERE ID = ".$R['ID']."  " ;
                $STMT_UPDATE = $CONNECTION->PREPARE($SQL_UPDATE);
                $STMT_UPDATE->EXECUTE();

                ECHO "ENDERECO: ".$R['NM_LOGRADO']." SIMILARIDADE DE LOGRADOURO NAO ENCONTRADO NO BANCO \N" ;
            }


        }
    }
}
```

Geocoding dengue cases for spatial analysis

J. L. S. Lustosa[1], L. M. Brasil[1,2], M. T. Obara[3], D.F. Lima[1] and L.C.A de Sousa[1]

[1] Universidade de Brasília (UnB)/Programa de Pós-Graduação em Engenharia Biomédica (PPGEB), Brasília, Brazil

[2] Universidade de Brasília (UnB)/Lato Sensu em Engenharia Clínica, Brasília, Brazil

[3] Universidade de Brasília (UnB)/Núcleo de Medicina Tropical, Brasília, Brazil

Abstract: **Brazil goes through a phase of great epidemic possibilities, mainly by the *Aedes aegypti* vector. The necessity to control these disease agents, the absence of antiviral drugs or vaccines for treatment contributes to proliferation. The cases of the disease in Brazil are recorded in *Information System on Notifiable Diseases* (Sistema de Informação de Agravos de Notificação - SINAN). Developed in the early 1990's and the main objective is collecting and processing data to about diseases in the country. At SINAN we can find a lot of information like sorology, date of hospitalization, state, address, and others. Analyze spatial data is the main objective and investigate geographical patterns in the determined region of dengue cases. However, analyzing a sample of the data we are faced with a series of address information that is not properly informed. The patient address data usually abbreviated irregularly in the address field, this study tries to improve the quality of these unstructured data, geocoding this information to obtaining the latitude and longitude coordinates of dengue cases, for posterior analysis in geographical information's system.**

Keywords: ***Spatial Analysis, Geocoding, SINAN, Google Geocoding API.***

I. Introduction

Arboviruses (from "*arthropod borne viruses*") are viruses that can be transmitted to humans by arthropod vectors, like *Flaviviridae*: *Febre Amarela*, *Dengue*, *Zika* and others. The main vector of dengue is *Aedes aegypti* and *Aedes* (*Stegomyia*) *albopictus*, which has great anthropophilia, adaptability, environmental transformations and domiciliation. These attributes contribute to increasing the breeding sites of the mosquito [1]. The dengue has characteristics that amplify the spread of the disease and increase the possibility of large and explosive epidemics. Viral replication in the *Aedes albopictus* mosquito besides *Aedes aegypti* increases the geographical extent of regions with viral circulation potential [2]. According to the article Zika Virus Outbreak, Bahia [3] cases of infection with *Dengue*, *Chikungunya* and Zika *Virus* (DENV, CHIKV and ZIKV) in Brazil and other places where the diagnosis is based on clinical and epidemiological reasons are not reliable, and all these questions show the need for laboratory confirmation of these arboviruses.

Due to the absence of antiviral drugs or vaccines to treat or prevent these diseases agents, the only available option for prevention is, control the reproduction of *Aedes aegypti*, reduction of the breeding source and using larvicides (immatures) and insecticides to adults [4]. In the last two decades, the incidence of dengue in the Americas has shown an upward trend, more than 30 countries reporting cases of the disease, despite the numerous eradication or control programs that have been implemented [5]. Concerning the cases of the disease here in Brazil, is used the SINAN. It was developed at the beginning of the 1990s, and the main objective is the collection and processing of data on reporting diseases on all country, providing information for the analysis of the morbidity profile and contributing to decision making for all level of public health administration [6]. SINAN is the main data source for surveillance of dengue and the notification is based on the communication of confirmed and suspected cases, not only of dengue but of other diseases that occur in our country. SINAN data are sent to the system through the *Individual Notification Form* (Ficha individual de Notificação - FIN), filled when the patient arrives at hospital and there is a suspected case of dengue [7]. After, this information's are sent electronically to the Ministry of Health.

The epidemiological bulletins are based on SINAN data, publish by the Ministry of Health and give a global view of dengue cases in Brazil. For more effective prevention, we need more details of all these cases, need to know the neighborhood and street, it is not enough that we only take care of our house if there is an abandoned build or even an abandoned swimming pool. They become a breeding and growing dengue cases in the region. Geocoding the address data of dengue cases is one solution that enables a more effective analysis of vectorial control. To analyze all these data, we will need a *Geographic Information System* (GIS), is the tool most used for spatial analysis. The GIS use spatial data like latitude and longitude information to georeferencing of the addresses, the name of this process is geocoding. The address is a textual information, needs a treatment to be used in a computer system, incorporated in a GIS to serve

like spatial statistic component for the analyses. This preparation consists in normalization (treat abbreviations, spaces, special characters), separation (in type of public place, title, name, number, complement, neighborhood, city and other references) and standardization (match the database format Base) [12].

The SINAN data still the largest official source for diseases cases information in Brazil [11] and according to DAVIS [8], 80% of the information used by local governments in the area of health, safety and education is associated with the geographical location, especially to addresses.

For this reason, it is important to geocode the data and analyze on the GIS to observe statistical and geographic patterns in the data.

II. MATERIAL AND METHODS

For the development, the data extracted from the SINAN *dBASE format* (DBF) [5], will be used a staging database. A relational database, faster and cheaper prepared for a large number of SINAN records is MariaDB [9].

To geocoding theses address from SINAN database we going to use, Google Maps Geolocation API [13], this tool can be accessed from internet, by http protocol and will receive by params the address information. Geocoding is a set of methods capable of transforming descriptions into geographic coordinates. All these data are, in general, place names, relative positional expressions or addresses.

To build the algorithm that will geocode the data, we used the Personal Hypertext Preprocessor (PHP) programming language. PHP is one of the most widely used programing languages in the world. Its popularity is due to the ease in creating dynamic applications, with support to the majority of existing databases and to the set of functions that, by means of a flexible structure, easy code construction [10]. The data geocoding engine searches each address in the Google map database using a web service and returns a set of geographic information in JavaScript Object Notation (JSON) format. Like in Figure 1.

Figure 1. JSON response retrieved from google maps request . Own authorship

The data returned by the service must be treated, sorted, classified, and properly stored for future reference. We avoided creating new fields in the SINAN main table, but we need to create new field like a primary key to identify the records like unique. I other table we store the field created in SINAN cases table and the latitude, longitude data. In order to group new cases and allow more optimized queries in the database, we created three additional columns in this additional table, which are: neighborhood, city and postal code.

These fields allow us to store the high-quality data from the Google Maps API, in addition to allowing you to quickly count records and generate statistical information easily. A piece of the source code that performs this operation can be seen in Figure 2.

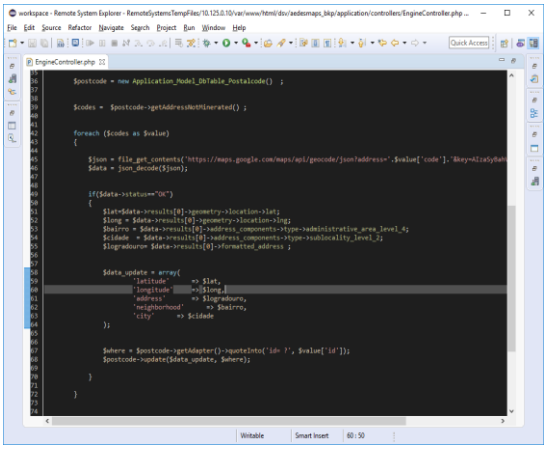


Figure 2. Piece of geolocation engine source code .Own authorship

The main objective is spatial analyses, using the geolocation we initially pretend evaluate which locations have more cases and how they are

distributed, like in Figure 3. To view these cases, we use the geographic information system Google Maps. In this map we group the cases by geographical coordinates and change the color of the marker according to the number of cases.

To display the map in Figure 3, we export the records stored in the database in Keyhole Markup Language (KML) format, our future objective is to allow data to be directly accessed from the database without one middle tier.

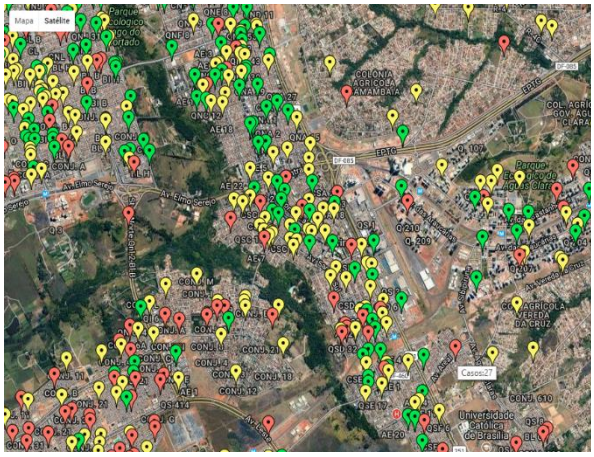


Figure 3. Cases by geographical coordinates. Own authorship

Was used in this research, data manipulation tools like databases. Any information that can identify the patients was previously suppressed from reports. For legal reasons, the data has been registered under CAAE 53662416.2.0000.0030.

III. RESULTS

The period studied included 70,525 suspected and confirmed cases of dengue, divided in the years 2010 to 2015 as shown in Table 2. We had a low percentage of records that wasn't useful to be used. The percentual of empty address records was 3.21%, and the percentual of addresses with less than 5 characters was 0,38%. In the studied location, there aren't addresses with less than 5 characters.

Table 1. Address Sample Distribution

Situation	Records	Total
Empty Address	2.266	3,21%
Small Address*	268	0,38%
Full Address	67.991	96,53%

* Address with five or less characters.

The total of addresses consulted in Google geolocation service was 67,991 records. 54,774 of this records, Google geolocation service was able to find the coordinates of latitude and longitude, with an efficiency of 80.56%. In Table 2 we can see geocoding rate was uniform over the years, maintaining an average of 81.48%, too close to the overall geocoding rate of all data analyzed.

Table 2. Mineration Engine Geocoding Rate

Year	Records	Successful	Geocoding Rate
2010	10.746	9.610	89,42%
2011	6.980	5.521	79,09%
2012	2.755	2.325	84,39%
2013	19.401	14.995	77,28%
2014	17.340	14.838	79,80%
2015	10.857	8.571	78,94%

We also analyze if exists any relationship to the size of the address. And it was possible to observe that Google geolocation service presented difficulties to find addresses coordinates with the increase of addresses characters, as we can see in Table 3.

The geocoding process obtained a high success rate in relation to the cases studied by Magalhães, Matos and Medronho [7].We believe that this rate is due to the fact that Brasilia addresses uses predictable pattern of organization, distributed in blocks and sets that exist in few cities of Brazil and the world. Which makes Google's geolocation service much efficient because creates unique address combinations.

Table 3 Geocode Rate by address character count

Character Count	Records	Geocoding Rate
5	585	99,83%
6	478	99,58%
7	166	98,19%
9	891	95,17%
12	1945	93,93%
13	2479	92,78%
8	673	90,79%
22	5327	89,32%
10	1298	88,75%
11	2131	88,36%
21	3496	88,07%
14	3835	87,25%
23	3753	87,21%
20	2869	85,19%
24	2526	80,64%
26	2067	78,71%
25	2885	78,37%
17	6084	77,84%
19	2804	76,93%
27	1518	75,36%

15	2589	73,70%
29	8059	73,61%
28	1201	71,61%
16	3636	65,62%
18	4350	65,61%
30	131	54,20%
32	9	44,44%
31	198	26,77%
33	8	0,00%

IV. Discussion

The previous bibliographical survey performed for this research, several related works were found. Each of these with different methodologies and techniques.

Davis and Fonseca [8], proposed an address analysis methodology for free text, considering Brazilian localization standard. In their article, the authors proposed three stages analyses. The Parsing Stage, consists in identify the address components and organizing into an appropriate data structure. Matching stage that consists searching the value in address database in order to search for a better correlation between the identified values and the present data in the database. The last stage, extracts the corresponding coordinates from the reference database, generating three indicators: Parsing Certainty Indicator (PCI), Matching Certainty Indicator (MCI) and Locating Certainty Indicator (LCI).

Each stage, these indicators receive a value between 0 and 1, where 0 represents total uncertainty in the result, while 1 represents maximum certainty. This value is based on several rules, involving the approximate search of components of the address with bonus of hits and discount of errors among the results searched. The final Geocoding Certainty Indicator (GCI) is obtained through the product of the indicators of each stage. This indicator according to the author allows a faster and reliable result, with low computational cost. Magalhães, Matos and Medronho [7] used the urban addresses present in the SINAN data from the Tuberculosis cases of Rio de Janeiro to analyze the incidence rate by neighborhood using georeferencing.

In the matching stage, they used the commercial software ARCGIS, which is a set of applications for geographic information management and the *Google Maps Geolocation API* to generate the pair of coordinates necessary to show the cases on the map. They have compared the two methods and Google Maps found more coordinates [8].

Google Maps has a very easy-to-use graphical user interface, but for a more effective analysis, we'll need to apply specific data filters like time period, cases confirmed, types of viruses and other information that is available in the SINAN database.

V. Conclusions

Like best results of the study of Magalhães, Matos and Medronho [7] was obtained using the *Google Geocoding API,* we chose this platform as the best tool for mining data.

However, the SINAN data quality is very precarious and many addresses need a human to understand, in next study's we pretend use artificial intelligence, more specifically neural networks for patterns recognition [15] and learning of these patterns, giving more precise results.

The *Google Geocoding API* [13] can understand many unstructured addresses, but each region has its own toponymical dictionary [14], in second moment will be necessary develop a previous layer to help minerating engine, translating some specific names places to a name that *Google API* can translate.

VI. References

1. CUNHA, R.V.; NOGUEIRA, R.M.R. Dengue e dengue hemorrágico. In: Coura, J.R. — Dinâmica das doenças infecciosas e parasitárias (Rio de Janeiro, Guanabara-Kooganp. 1767-81. 2006.
2. DONALISIO, M. R.; FREITAS, A. R. R. Chikungunya in Brazil: anemerging challenge. Revista Brasileira de Epidemiologia, 18(1), 283-285, 2015.
3. BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância das Doenças Transmissíveis Plano de Contingência Nacional para Epidemias de Dengue / Ministério da Saúde, Secretaria de Vigilância em Saúde, Departamento de Vigilância das Doenças Transmissíveis. – Brasília: Ministério da Saúde, 2015.
4. BARRERA, R. Control de los mosquitos vectores del dengue y delchikunguña:¿ es necesario reexaminar las estratégias actuales?. Biomédica,35(3), 297-9, 2015.
5. BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde, Departamento de Vigilância Epidemiológica. Diretrizes Nacionais para a Prevenção e Controle da Dengue. Brasília, 2009.
6. LAGUARDIA, J. et al . Sistema de informação de agravos de notificação em saúde (Sinan): desafios no desenvolvimento de um sistema de informação em saúde. Epidemiol. Serv. Saúde, Brasília , v. 13, n. 3, p. 135-146, 2004.
7. MAGALHAES, M. A. F. M.; MATOS, V. P.; MEDRONHO, R. A. Avaliação do dado sobre endereço no Sistema de Informação de Agravos de Notificação utilizando georreferenciamento em nível local de casos de tuberculose por dois métodos no município do Rio de Janeiro. Cad. saúde colet., Rio de Janeiro, v. 22, n. 2, p. 192-199, 2014.
8. DAVIS JR., C.A.; FONSECA, F.T. "Assessing the Certainty of Locations Produced by an Address Geocoding System." Geoinformatica 11(1): 103-129, 2007.
9. BALUSAMY, B. et al. Cloud Database Systems: NoSQL, NewSQL, Hybrid. In: Advancing Cloud Database Systems and Capacity Planning With Dynamic Applications. IGI Global, 2017. p. 225-245.
10. DALL'OGLIO, P. PHP Programando com Orientação a Objetos 3ª Edição. Novatec Editora, 2015.

11. BRASIL. Ministério da Saúde. Sistema de Informação de Agravos de Notifcação: normas e rotinas. Brasília: Ministério da Saúde; 2006.
12. Skaba, Daniel Albert. Metodologias de geocodificação dos dados da saúde. Diss. 2009.
13. Singh, Sushant K. "Evaluating two freely available geocoding tools for geographical inconsistencies and geocoding errors." Open Geospatial Data, Software and Standards 2.1 (2017): 11.
14. Jung, Chin-Te. "Geolocation Services." The International Encyclopedia of Geography (2017).
15. De Azevedo, Fernando Mendes, Lourdes Mattos Brasil, and Roberto Célio Limão de Oliveira. Redes neurais com aplicações em controle e em sistemas especialistas. Visual Books, 2000.

Author: Jorge Luis da Silva Lustosa
Institute: Universidade de Brasília, FGA Gama
Street: Área Especial de Indústria Projeção A
Neighborhood: Setor Leste
City: Brasília, Distrito Federal
Country: Brasil
Email: jorgelustosa@gmail.com
Phone Number: +55-61-99224-2507

SISTEMA ESPECIALISTA PARA DIRECIONAMENTO NA DIFERENCIAÇÃO ENTRE AS PATOLOGIAS CAUSADAS PELO AEDES AEGYPTI

R.R. Amaral*, L.C.A de Sousa*, J.L.S Lustosa*, H.M. Sanches*, L.M. Brasil*
* Universidade de Brasília, Faculdade Gama (UnB/FGA), Programa de Pós-Graduação em Engenharia Biomédica, Brasília
e-mail: rodrigo.com2@gmail.com

Resumo: Os Sistemas Especialistas (SE) estão diretamente relacionados com a Inteligência Artificial (IA) e são desenvolvidos para resolverem problemas que apenas o especialista da área tem domínio. Este projeto tem o objetivo de propor a implementação do uso de um SE no apoio ao diagnóstico do vírus da Dengue, Chikungunya e Zika. As três doenças possuem várias características em comum. Assim, o SE direcionará o usuário para uma hipótese diagnóstica através de perguntas e respostas univaloradas e multivaloradas, que permitem ao sistema chegar a um resultado final mostrando o possível diagnóstico de forma simples e rápida. Para alcançar esse objetivo foi necessário o auxílio de um especialista da área de saúde, que conhece os sintomas de cada doença, para coletar e organizar as informações disponibilizadas pelo Ministério da Saúde (MS). A ferramenta utilizada foi um software livre, denominado Shell Expert SINTA, que utiliza um modelo de representação do conhecimento baseado em regras de produção e probabilidades. O resultado foi um programa capaz de produzir uma hipótese diagnóstica somente com a anamnese e observações clínicas do paciente.
Palavras-chave: Sistema especialista, hipótese diagnóstico, dengue, Chikungunya, Zika.

Abstract: *Expert Systems (ES) are directly related to Artificial Intelligence (AI) and designed to solve problems that only the specialist area has domain.This project has the objective of propose the implementation of the use of an ES to support the diagnosis of Dengue, Chikungunya, and Zika virus.The three diseases have several characteristics in common. The ES will direct the user to a diagnostic hypothesis through questions and answers, with single or many values, that allow the system reach a final result directing to diagnose quickly and easily. To reach this objective it has been necessary the assistance of expert in the health area ,who knows the symptoms of each disease, to collect and organize the information provided by the Ministry of Health (MH).The tool used was a free software called Shell Expert SINTA, that uses model of knowledge representation based on production rules and probability. The result was a program to produce a diagnostic hypothesis just with history and clinical observation of the patient.*
Keywords: *Expert System, diagnostic hypothesis, Dengue, Chikungunya, Zika.*

Introdução

IA é uma ciência capaz de sintetizar e automatizar tarefas intelectuais, sendo que o seu principal objetivo é fazer o computador "pensar (raciocinar)" [1] [2]. Para alcançar esse objetivo, é necessário realizar aquisições, representações e manipulação do conhecimento. O processo de manipulação é capaz de deduzir ou inferir novos conhecimentos a partir de um conhecimento existente, que geralmente são utilizados para resolver problemas complexos [3]. Para isso, a partir da década de 1970, foram implantados os SE [2]. O SE é um sistema computacional que provê ao computador mecanismos e meios de forma que ele se equipare ao especialista humano [2] [4]. O SE deve conter, basicamente, duas perspectivas distintas: a do conhecimento, processável pelo homem, e a simbólica, processável pelo computador [5]. Vários passos são necessários para obter êxito na criação de um SE, que vão desde a estrutura física da máquina e o seu abastecimento com conhecimento, ao aprendizado do sistema propriamente dito (Figura 1). O engenheiro do conhecimento tem um papel fundamental na interpretação e transformação do conhecimento do especialista, de forma a viabilizar o seu armazenamento e construir a base de dados. Para que o sistema funcione é necessário que haja um motor de inferência, um elemento permanente, responsável por avaliar as regras da base de conhecimento para serem analisadas, direcionando o processo de tomada de decisão. Necessita também de um mecanismo que mostre ao usuário como se chegou à determinada resposta. A esse processo dá-se o nome de Interface de Explanação (Figura 1) [5] [4].

Visando uma maior viabilidade econômica na implementação de um SE foram criadas ferramentas, shells, aptas a realizar muito do trabalho necessário para transpor um SE para um computador [6]. O shell é o elo entre o usuário e o sistema, que permite ao criador do programa preocupar-se somente com a representação do conhecimento do especialista e dessa forma, simplificar ao máximo a implementação de um software [7]. O shell utilizado para realizar a diferenciação na tríplice viral neste trabalho, denominado Shell Expert SINTA, é um software livre, de código aberto e gratuito, ou seja, permite ao usuário a liberdade de executar para qualquer uso, de adaptá-lo às suas necessidades, redistribuir cópias e melhorar o programa [9].

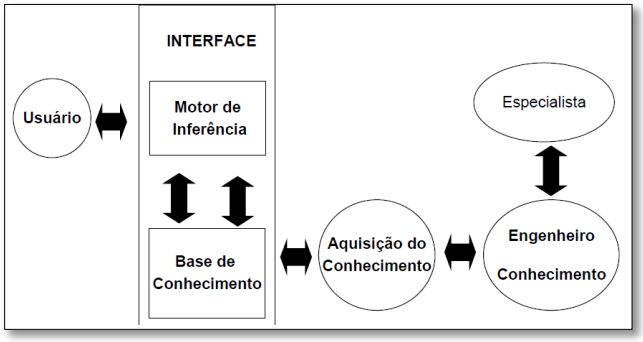


Figura 1: Estrutura dos sistemas especialistas

O Aedes aegypti é hoje um dos principais problemas em saúde pública, devido ao seu papel como vetor do vírus da febre amarela, e da tríplice viral: dengue, chikungunya e zika. A dengue foi a primeira da tríplice viral a reemergir no Brasil na década de 1980. Depois veio a Chikungunya, que emergiu no Brasil em 2014 e a Zika, a mais recente, que surgiu em 2015. Ambas possuem uma sintomatologia muito parecida e são facilmente confundidas [10] [11]. É importante destacar que quando ocorre uma epidemia, não é indicada a confirmação de todos os casos suspeitos por exame laboratorial. A maior parte dos casos será confirmada por critério clínico-epidemiológico [12] [13].
O objetivo deste artigo é propor um SE capaz de diferenciar a tríplice viral causada pelo mosquito do gênero Aedes que compromete a saúde pública brasileira e de outros países utilizando a ferramenta computacional, Shell Expert SINTA.

Materiais e métodos

A ferramenta computacional utilizada para se construir o SE foi o Shell Expert SINTA desenvolvido pelo Laboratório de Inteligência Artificial (LIA) que constitui um conjunto de laboratórios de pesquisa do Mestrado e Doutorado em Ciência da Computação da Universidade Federal do Ceará (UFC) [6].
A base de conhecimento foi criada a partir de informações e protocolos disponibilizados pelo MS e de uma especialista com conhecimento na área. O papel do especialista, foi coletar informações disponibilizadas pelo MS e apresentá-las de maneira clara e coerente ao programador. As publicações do MS datam desde 1987 até 2016 [11]-[14].

Resultados

A partir das informações obtidas para a implementação do SE, oriundos do MS, que datam desde 1987 até 2016, com esses dados foi elaborada a Tabela 1 para os sintomas mais frequentes, os quais estabeleceram as variáveis e os seus critérios de avaliação que seguem a Escala Análoga Visual (EAV), especificado pelo MS, segundo a circular normativa da Divisão de Doenças Genéticas, Crónicas e Geriátricas (Nº 09/DGCG de 2003) para dor (cefaléia, mialgia, artralgia), edema e náusea. Foram realizadas combinações entre os sintomas mais frequentes, em grupos de três, conforme protocolos disponibilizados pelo MS (2016), para classificação de risco e manejo clínico do paciente com suspeita de Dengue, *Zika* e *Chikungunya*, como demonstrados na Figura 2 [13]-[16].

Tabela 1: Variáveis e seus Valores

Variáveis	Valores
Febre	Ausente; < 38.5ºC; >38.5ºC; entre 39-40ºC
Exantema Maculopapular	Ausente, leve, moderado, intenso
Híperemia Conjuntival	Ausente, leve, moderado, intenso
Mialgia	Ausente, leve, moderado, intenso
Artralgia	Ausente, leve, moderado, intenso
Edema	Ausente, leve, moderado, intenso
Dor Retrorbital	Ausente, leve, moderado, intenso
Cefaléia	Ausente, leve, moderado, intenso
Náusea/vômito	Ausente, leve, moderado, intenso
Hemorragias	Ausente; presente
Prova do laço	Positiva; negativa
Hemograma	Leucopenia (sim/não); Trombocitopenia (sim/não); Hemoconcentração (sim/não)
Linfadenopatia	Ausente, leve, moderado, intenso
Resultado	Dengue/*Chikungunya*/*Zika*/Não se encaixa na tríplice viral

Para realizar o cruzamento dos sintomas foram seguidos os protocolos para classificação de risco e manejo do paciente que estabelece:
- Dengue: febre alta com duração máxima de 7 dias mais pelo menos dois sintomas (cefaleia, dor retrorbital, exantema, prostração, mialgia, artralgia) [15].
- Chicungunya: febre baixa a inexistente por até 7 dias acompanhada de artralgia intensa de início súbito associada a outros sintomas característicos (cefaleia, mialgia, exantema) [16].
- Zica: exantema maculopapular pruriginoso com início até 48h, após os primeiros sintomas (artralgia, mialgia, hiperemia conjuntival sem secreção, cefaleia, mal estar) acompanhados ou não de febre [13].

Figura 2: Possibilidades de cada sintoma.

As intensidades dos sintomas, tais como a dor, são muito subjetivos, por isso, apesar das escalas para aferição das mesmas, as queixas e as descrições do paciente tem sempre o maior peso levado em

consideração.

Para realizar a representação do conhecimento dos dados coletados foi utilizada a ferramenta Shell Expert SINTA que utiliza o modelo de representação do conhecimento baseado em regras de produção e probabilidades, o que o torna uma ferramenta simples (Figura 3).

A partir dos conhecimentos coletados do MS e das informações sobre o Shell em questão, foi criado um questionário que guia o usuário através de perguntas em que ao final apresenta um resultado apontando a melhor hipótese diagnóstica. Para chegar a este resultado foi necessário criar regras que analisassem cada uma das combinações de acordo com os sintomas (Figura 3). No desenvolvimento do questionário foi importante diminuir a quantidade de perguntas, para que o programa se tornasse mais simples e, dessa forma, foram estabelecidas questões para uma triagem inicial, onde o usuário marca apenas 3 sintomas pontuais dentre os 12 apresentados. Esta seleção agiliza o tempo de resposta de um resultado, como se pode ver na Figura 3A. No passo seguinte, o sistema pergunta sobre o hemograma, uma vez que o resultado do hemograma elimina uma série de outras probabilidades, e aumenta o grau de confiança de outras. Na Figura 3B e 3C pode-se visualizar estas outras interações. Após análise da triagem e do hemograma, questões com possibilidades mais detalhadas para escolha são apresentadas (Figura 3D). Na tela de resultado (Figura 3E) pode-se clicar na guia Histórico e visualizar quais regras foram utilizadas para construir a hipótese de diagnóstico e entender como o sistema chegou ao resultado apresentado (Figura3F).

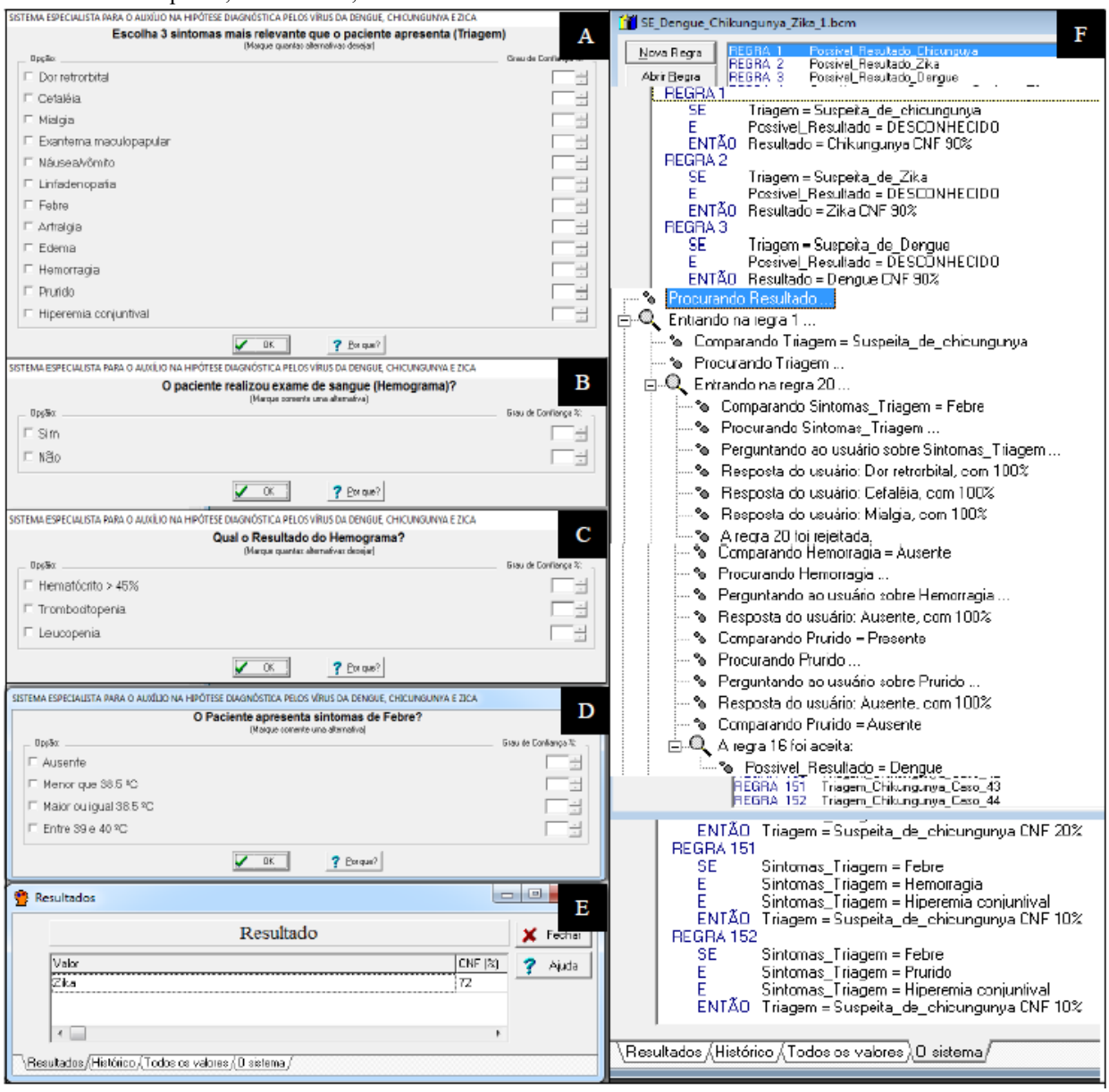


Figura 3: Estrutura do Shell Expert SINTA mostrando as telas do questionário e número de regras do SE com seus respectivos intervalos de confiança (CNF). (A) Triagem inicial. (B; C) telas do hemograma – realização e resultados. (D) continuação do questionário: sintomas de febre, (E) direcionamento para a hipótese diagnóstica com seu respectivo CNF. (F) demonstração do encadeamento de regras, bem como a árvore de decisão para se chegar a um resultado aceitável

Discussão

Ao analisar os protocolos de diagnósticos fornecidos pelo MS, foi possível observar a grande quantidade de probabilidades envolvidas na diferenciação do diagnóstico na tríplice viral. Essa grande quantidade de probabilidades podem, em muitos casos, gerar um diagnóstico equivocado. Sendo assim, a utilização de um SE se torna notoriamente importante para auxiliar os profissionais de saúde a otimizarem a triagem e minimizar possíveis erros na escolha de um tratamento.

O Shell Expert SINTA é uma ferramenta voltada para o ambiente acadêmico e o seu apelo é ser uma ferramenta simples que possibilite a qualquer profissional da área de saúde, criar um SE e usar no seu dia a dia. Devido à complexidade dos dados e as limitações do Shell Expert SINTA, outra linguagem, conhecida como 'C' Language Integrated Production System (CLIPS) [17], que é muito próxima ao Shell Expert SINTA foi cogitada. A mesma é atualizada constantemente e tem uma grande comunidade de adeptos, cuja a principal vantagem de admitir o uso de funções e encadeamento de regras que permitem simplificar a quantidade de regras a serem utilizadas.

Conclusão

O Shell Expert SINTA é uma ferramenta muito simples e demonstra como a IA é algo que pode estar ao alcance de todos. O SE deste projeto demostrou que se podem obter resultados satisfatórios com a implementação do software, auxiliando o profissional a desenvolver suas tarefas do cotidiano com mais precisão e agilidade. Em épocas de endemias e epidemias não se faz necessário a comprovação laboratorial de todos os casos suspeitos, tornando assim um SE para esse fim uma ótima ferramenta para o direcionamento da tomada de decisões médicas.

Conclusão

O Shell Expert SINTA é uma ferramenta muito simples e demonstra como a IA é algo que pode estar ao alcance de todos. O SE deste projeto demostrou que se podem obter resultados satisfatórios com a implementação do software, auxiliando o profissional a desenvolver suas tarefas do cotidiano com mais precisão e agilidade. Em épocas de endemias e epidemias não se faz necessário a comprovação laboratorial de todos os casos suspeitos, tornando assim um SE para esse fim uma ótima ferramenta para o direcionamento da tomada de decisões médicas.

Referências

[1] Pinto E. Sistema inteligente para especificação do aperto ideal em operações de parafusamento [Dissertação]. São Paulo Universidade de Taubaté, 2005.
[2] Russel S, Norvig P. Inteligência Artificial. 2ª. ed. Rio de Janeiro: Elsevier, 2004.
[3] NCE. Visão Geral Sobre Inteligência Artificial, 2016.Disponivel em: http://www.nce.ufrj.br/GINAPE/VIDA/ia.htm.
[4] Brasil LM. Informática em saúde. Brasília: Universa, 2008.
[5] Rezende SO. Sistemas Inteligentes: fundamentos e aplicações. Barueri, SP: Manole, 2003.
[6] Expert SINTA - Versão 1.1 - Manual do Usuário - Laboratório de Inteligência Artificial /LIA - UFC, CE, 1996.
[7] Pisco L. Disciplina: sistema operacional, 2016 [internet]. Abril de 2016. Disponivel em: http://www.simonsen.br/its/pdf/apostilas/base-tecnica/1/sistema-operacional-1-ano-de-informatica-completa.pdf.
[8] Spirlandell, lP. Sistemas Especialistas: um estudo de caso com o Expert SINTA. Revista eletrônica de sistemas da informação e gestão tecnológica, v. 1, n. 1, 2011.
[9] O'Brien JA. Sistemas de informação e as decisões gerenciais na era da Internet. 2ª. ed. São Paulo: Saraiva, 2004.
[10] Donalisio MR, Freitas ARR. Chikungunya no Brasil: um desafio emergente. Revista Brasileira de Epidemiologia. São Paulo, jan-mar, 2015.
[11] MS, Ministério da Saúde. Dengue: aspectos epidemiológicos, diagnóstico e tratamento. Ministério da Saúde, Fundação Nacional de Saúde. Brasília: Fundação Nacional de Saúde, 2002.
[12] Oliveira WKD. Zika Vírus - Informações Sobre a Doença e Investigação de Síndrome Exantemática no Nordeste. Brasília: Ministério da Saúde, 2005.
[13] MS. Ministério da Saúde. Manual de vigilância epidemiológica da febre amarela - Brasília: Ministério
[14] BVS, Biblioteca virtual em saúde. Como diferenciar Dengue, Chikungunya e Zika. Ministério da saúde. Brasília, 2016.
[15] MS, Ministério da Saúde. Direção-Geral da Saúde. Divisão de doenças genéticas, crônicas e geriátricas. Circular normativa n. 09 – DGCG, 2003.
[16] Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância das Doenças Transmissíveis. Dengue: diagnóstico e manejo clínico: adulto e criança [recurso eletrônico] / Ministério da Saúde, Secretaria de Vigilância em Saúde, Departamento de Vigilância das Doenças Transmissíveis. – 5. ed. – Brasília: Ministério da Saúde.
[17] Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância das Doenças Transmissíveis. Febre de chikungunya: manejo clínico / Ministério da Saúde, Secretaria de Vigilância em Saúde, Secretaria de Atenção Básica. – Brasília: Ministério da Saúde, 2015.
[18] MS, Ministério da Saúde. Febre de chikungunya: manejo clínico. Ministério da Saúde, Secretaria de Vigilância em Saúde, Secretaria de Atenção Básica. - Brasília: Ministério da Saúde, 2015.
[19] CLIPS [internet]. Abril de 2016. Disponível em: http://www.clipsrules.net/?q=AboutCLIPS, acessado em:02/05/2016