Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

# Business Intelligence em Vigilância Epidemiológica baseada em dados produzidos pelos Laboratórios de Saúde Pública

Ronaldo de Jesus

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Gladston Luiz da Silva

Brasília
2018

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Jb

Jesus, Ronaldo
Business Intelligence em Vigilância Epidemiológica baseada em dados produzidos pelos Laboratórios de Saúde Púbica / Ronaldo Jesus; orientador Gladston Luiz da Silva. -- Brasília, 2018.
140 p.

Dissertação (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2018.

1. Business Intelligence. 2. Séries Temporais. 3. Gráfico de Controle. 4. Vigilância em Saúde. 5. Laboratórios de Saúde Pública. I. Luiz da Silva, Gladston , orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

# Business Intelligence em Vigilância Epidemiológica baseada em dados produzidos pelos Laboratórios de Saúde Pública

Ronaldo de Jesus

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Gladston Luiz da Silva (Orientador)
CIC/UnB

Prof. Dr. Marcelo Ladeira
Universidade de Brasília

Prof. Dr. Alessandro Pecego Martins Romano
Ministério da Saúde

Prof.a Dr.a Aletéia Patrícia Favacho de Araújo
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 27 de dezembro de 2018

# Dedicatória

Se nos alegramos por aquilo que nos dá potência, então não haveriam outros a quem dedicar este trabalho.

Se amamos o que desejamos e se amamos também o que se faz presente no tempo presente, então não haveriam outros a quem dedicar este trabalho.

Se nos sacrificamos pelo que nos é sacro, então, também não haveriam outros a quem dedicar este trabalho.

Este trabalho é dedicado aos meus pais, minha amada esposa (ela é tudo para mim), meus filhos, meus irmãos e todos os que fazem parte do meu ciclo de vida. Que este trabalho possa despertá-los e fazê-los acreditar em um futuro melhor.

A todos, invoco a refletir que: ***"O caminho é mais importante que o destino!"***

# Agradecimentos

Primeiramente agradeço a Deus, por ter dado-me saúde, sabedoria e condições para concluir este trabalho. Ademais, são tantos aqueles que tornaram este trabalho possível que não conseguiria agradecer a todos nominalmente. Escolho nomear alguns, na esperança que a gratidão a estes aqui expressa transcenda aos demais. Agradeço a todos amigos e colegas do Ministério da Saúde que de alguma forma colaboraram com este trabalho, em especial aos colegas Rosa Maria da Silva, Silvano Barbosa, Mariana Verotti, André Luiz e Osnei Okumoto. Aos colegas de mestrado por todos momentos intensos, maravilhosos e inesquecíveis que vivemos juntos. À Universidade de Brasília e a todos seus professores, funcionários e alunos, em especial aos Professores: Dr. Gladston Luiz da Silva (meu orientador - pela paciência, apoio e direcionamento), Dr. Marcelo Ladeira (por me ouvir, ajudar e auxiliar tantas vezes), Dr. Edgard Costa Oliveira, Dra. Ligia Maria Cantarino e Dr. Alessandro Pecego Martins Romano. Não poderia deixar de agradecer, também, a Luíza Tuler por contribuir com este trabalho. Estes levarei comigo guardados no meu coração por toda a vida. A minha família, em especial a minha esposa, companheira, mulher e amada, Andreia Carla, pelo apoio incondicional, pelos incentivos nos momentos difíceis, pela paciência e amor demonstrados ao longo deste trabalho. Te amo e sempre te amarei; Aos meus filhos, Vitor Hugo, por aquilo que é inominável no meu aprendizado do dia a dia e que certamente está presente nessas linhas, Rafael e Pedro (meus gêmeos), que chegaram neste ano mudando completamente nossas vidas com mais amor e alegria; Ao caríssimo e amor dos meus pais pelo apoio, por dividir suas histórias de vida que me traz força e orgulho de tê-los como pais. A eles (neste momento me emociono ao trazer em minha memória todos os desafios que eles transpuseram para que eu estivesse aqui e nestas condições...) dou-lhes saudações especiais e o meu amor de filho, que é incondicional... que este trabalho possa refletir seus ensinamentos e perseveranças; E por fim, pelas projeções que em mim representa e que neste trabalho se fizeram presentes.

# Resumo

A globalização econômica e informacional proporcionou às pessoas maior interatividade e locomoção por diversos países, trazendo muitos benefícios mas, também, riscos. Neste sentido, pode-se citar, por exemplo, a circulação de novas doenças e agravos à saúde humana por todo o mundo [1] [2]. Essas interconexões entre os países vêm impactando a área da saúde e transformado a forma com que o mundo tem tratado as políticas e as ações em saúde, criando a valorização dos direitos humanos, da cadeia de valor, desafiando a política, os profissionais de saúde e reforçando as responsabilidades nacional e internacional dos países [2] [3]. Embora a saúde pública seja uma preocupação antiga, ela tem ganhado contornos diferentes a partir das experiências epidemiológicas construídas, elevando o conhecimento e transformado a prática clínica para todos [4]. Apesar do Brasil possuir muitas bases de dados com informações em saúde e muitas ações de vigilância epidemiológica serem realizadas, verifica-se que há outros mecanismos que podem ser implementados para melhorar, ainda mais, a vigilância, como por exemplo, a implementação de controles estatísticos do processo de produção de exames de saúde pública, que é um importante indicador de surtos e ameaças a saúde pública. Por outro lado, a computação, a estatística e o conhecimento de especialistas têm impulsionado a inteligência dos processos de vigilância, proporcionando componentes e procedimentos mais eficientes para as tomadas de decisões. Este trabalho propõe Indicador de vigilância epidemiológica a partir dos dados laboratoriais que são armazenados no Sistema Gerenciador de Ambiente Laboratorial (GAL), seguido de aplicações de modelos de Séries Temporais e Controle Estatístico de Processo (CEP), utilizando a linguagem de programação em **R** integrada a software de *Business Intelligence (BI)* para monitorar e detectar mudanças no comportamento da doença de Influenza no município de Curitiba-PR. Os resultados deste estudo foi a proposição de uma arquitetura de software, o que permitiu monitorar e controlar a doença de influenza, identificar eventos que indicaram situações fora dos limites de controles. Esse processo mostrou aplicável a outras doenças, contribuindo de forma relevante para a vigilância em saúde.

**Palavras-chave:** Vigilância em Saúde, Epidemiologia, Controle Estatístico de Processo

# Abstract

The globalization economic and informational to provide people with greater interactivity and mobility through different countries, bringing many benefits but also risks. In this sense, we can mention, for example, the circulation of diseases all over the world [1] [2]. These interconnections between countries have impacted in the area of health and transformed the way the world has handled health policies and actions, creating value for human rights, the value chain, challenging policy, health professionals and strengthening the national and international responsibilities of countries [2] [3]. Although public health is an old concern, it has gained different contours from constructed epidemiological experiences, raising awareness and transforming clinical practice for all [4]. Although Brazil has many databases with health information and many epidemiological surveillance actions are carried out, there are other mechanisms that can be implemented to further improve surveillance, such as the implementation of statistical controls of the process of producing public health tests, which is an important indicator of outbreaks and threats to public health. On the other hand, computing, statistics and expert knowledge have driven the intelligence of surveillance processes, providing more efficient components and procedures for decision making. This work proposes epidemiological surveillance indicator from the laboratory data that is stored in Sistema Gerenciador de Ambiente Laboratorial (GAL), followed by applications of Time Series models and Statistical Process Control (SPC), using the programming language in **R** *Business Intelligence (BI)* software to monitor and detect changes in the behavior of Influenza disease in the city of Curitiba-PR. The results of this study were the implementation of a solution in Information Technology (IT), which allowed to monitor and control influenza disease, identifying events that indicated situations outside the limits of controls. This process has shown to be applicable to other diseases, contributing significantly to health surveillance.

**Keywords:** Surveillance in Health, Epidemiology, Statistical Process Control

# Sumário

# Lista de Figuras

# Lista de Tabelas



,

# Lista de Abreviaturas e Siglas

**AIC** Critério de Informação de Akaike.

**AICc** Critério de Informação de Akaike Corrigido.

**AR** Auto-Regressivos.

**ARIMA** Auto Regressivos Integrados de Médias Móveis.

**ARMA** Auto Regressivos-médias Móveis.

**BI** Business Intelligence.

**BIC** Critério de Informação Bayesiano.

**CDC** Centro de Controle de Doenças.

**CEP** Controle Estatístico de Processo.

**CGDT** Coordenação Geral de Doenças Transmissíveis.

**CGLAB** Coordenação Geral de Laboratórios de Saúde Pública.

**CPF** Cadastro de Pessoa Física.

**CUSUM** Somas Acumuladas.

**DEGEVS** Departamento de Gestão da Vigilância em Saúde.

**DIS** Data Integration Server.

**ELK** Elasticsearch, Logstash e Kibana.

**ETL** Extração, Tratamento e Carga.

**FAC** Funções de Auto-correlação.

**FACP** Função de Auto-correlação Parcial.

**GAL** Sistema Gerenciador de Ambiente Laboratorial.

**GC** Gráfico de Controle.

**IBGE** Instituto Brasileiro de Geografia e Estatística.

**ID** Integração de Dados.

**JDBC** Java Database Connectivity.

**LACEN** Laboratórios Centrais de Saúde Pública.

**LC** Limite Central do Controle.

**LIC** Limite Inferior do Controle.

**LRN** Laboratório de Referência Nacional.

**LRR** Laboratórios de Referência Regional.

**LSC** Limite Superior do Controle.

**MMEP** Média Móvel Exponencial Ponderada.

**MS** Ministério da Saúde.

**OLAP** Processamento Analítico Online.

**OMS** Organização Mundial da Saúde.

**OPAS** Organização Pan-americana de Saúde.

**PDI** Pentaho Data Integration.

**SARIMA** Modelo Auto-regressivo Integrados de Médias Móveis Sazonal.

**SG** Síndrome Gripal.

**SGBD** Sistema Gerenciador de Banco de Dados.

**SINAN** Sistema de Informações de Notificações de Doenças e Agravos.

**SISLAB** Sistema Nacional de Laboratórios de Saúde.

**SQL** Structured Query Language.

**SRAG** Síndrome Respiratória Aguda Grave.

**SUS** Sistema Único de Saúde.

**SVG** Scalable Vector Graphics.

**SVS** Secretaria de Vigilância em Saúde.

**TI** Tecnologia da Informação.

**UnB** Universidade de Brasília.

**UTI** Unidade de Terapia Intensiva.

**VSP** Vigilância em Saúde Pública.

# Capítulo 1

# Introdução

A globalização econômica e informacional proporcionou às pessoas maior interatividade e locomoção por diversos países, trazendo muitos benefícios mas, também, riscos. Neste sentido, pode-se citar, por exemplo, a circulação de novas doenças por todo o mundo [1] [2].

Essas interconexões entre os países vêm impactando a área da saúde e tem transformado a forma com que o mundo trata as políticas e as ações em saúde, criando a valorização dos direitos humanos, da cadeia de valor, desafiando a política, os profissionais de saúde e reforçando as responsabilidades nacional e internacional dos países [2] [3].

Neste contexto, a Organização Mundial da Saúde (OMS) tem um importante papel, independente e mediador, no monitoramento, na formação de conhecimento e evidência em saúde, e, no direcionamento e auxílio aos líderes dos países na construção de políticas mais adequadas [3].

Historicamente, a vigilância em saúde pública surgiu no século XIX com o desenvolvimento da microbiologia, dos estudos e conceitos das doenças infecciosas, processamento de dados e combate ao crescimento de epidemias, evoluindo para a manutenção de alertas, monitoramento e controle de focos de doenças e agravos à saúde humana [5].

Atualmente, a vigilância em saúde é compreendida como um processo contínuo e sistemático de coleta, consolidação, disseminação de dados sobre eventos relacionados à saúde, que visa o planejamento e a implementação de medidas de saúde pública para a proteção da saúde da população, a prevenção e controle de riscos, agravos e doenças, bem como para a promoção da saúde [6].

Embora a saúde pública seja uma preocupação antiga, ela tem ganhado contornos diferentes a partir das experiências epidemiológicas construídas, elevando o conhecimento e transformado a prática clínica para todos [4].

Dentre as mais recentes práticas inovadoras, destaca-se a realizada pelo Centro de Controle de Doenças (CDC) dos Estados Unidos da América, que monitorou as redes sociais (Twitter, Facebook) e cruzou estas informações com os pontos geográficos dos celulares das pessoas que viajaram. Estas informações foram fornecidas pela maior empresa de telefonia móvel do Haiti. A partir desses dados, estimou-se a magnitude e a tendência de locomoção da população, comparado com a propagação da bactéria de *vibrio cholerae* (cólera). Fato este que comprovou o fluxo da bactéria entre vários países [7].

Constituem outras iniciativas inovadoras de vigilância em saúde:

1. **Google Flu Trends e Google Dengue Trends:** foi um sistema desenvolvido pela Google para estimar a magnitude da gripe e dengue, com base em padrões de busca. O sistema não está mais publicado [8].
2. **HealthMap:** é um sistema de monitoramento de doenças que coleta informações de diferentes fontes, realizando a categorização, priorização e disponibilização destas na Internet [7].

As técnicas exemplificadas demonstram que a vigilância em saúde tem sido realizada com o armazenamento, processamento e análise de dados. Isso não deixa dúvida de que os sistemas de informações e a estatística são fundamentais para a Saúde Pública [9].

No Brasil, essas ações e políticas públicas de saúde vêm sendo realizadas por meio do Ministério da Saúde (MS), Estados e Municípios. Os pilares básicos para prover a qualidade dos serviços e vigilância em saúde estão contidos e fundamentadas no Sistema Único de Saúde (SUS), Lei nº 8.080[1] de 17 de setembro de 1990, que são: universalidade, igualdade, equidade, integralidade, intersetorialidade, direito à informação, dentre outros [10].

O SUS inclui a saúde pública em todas as áreas, difundindo-a em vigilância sanitária, epidemiológica, saúde do trabalhador, alimentação e nutrição. Não obstante, o MS observa os cuidados coletivos e individuais, proporcionando as execuções de consultas, exames, internações, transportes, dentre outros. Por meio destas ações, têm-se construído bases de dados para subsidiar as tomadas de decisões.

Apesar de o Brasil possuir muitas bases de dados com informações de saúde e muitas ações a serem realizadas no quesito de vigilância epidemiológica, há muitos outros mecanismos que precisam ser implementados para melhorar, ainda mais, a vigilância, como por exemplo, a implementação de métodos estatísticos no processo de produção de resultados dos exames realizados pelos laboratórios de saúde pública. Esses são potenciais e importantes indicadores do processo de saúde.

[1]Lei nº 8.080/90: dispõe sobre as condições para a promoção, proteção e recuperação da saúde, a organização e o funcionamento dos serviços correspondentes e dá outras providências.

## 1.1 Apresentação

A estrutura de vigilância em saúde no Brasil é descentralizada entre os entes federativos (Federal, Estadual e Municipal), que compartilham informações em todo o sistema nacional de vigilância epidemiológica e adotam diferentes fontes de monitoramento e de indicações de surtos.

No âmbito Federal, o Ministério da Saúde adota como fonte principal de detecção de surtos e de monitoramento de doenças o Sistema de Informações de Notificações de Doenças e Agravos (SINAN). Ele foi desenvolvido no ano de 1990 para coletar informações de notificações de doenças e agravos à saúde, conforme relação publicada em portaria ministerial[2]. Sobre seu funcionamento, cabe ressaltar que a partir de uma notificação, a secretaria de vigilância inicia uma investigação do caso, buscando informações que possam comprová-lo ou descartá-lo [11].

Entretanto, a diversidade geográfica brasileira e as dificuldades tecnológicas de alguns municípios obrigou a construção do SINAN de forma descentralizadas e não online. Tal fato, contribui para a complexidade e dificuldade no armazenamento e fluxo das informações do sistema, o que produz diversos entraves e atrasos no processo de recebimento e consolidação dos dados de notificações ao MS.

Outro aspecto mais grave ainda, é a subnotificação de doença, ou seja, há uma irregularidade nos registros dos casos por parte dos Estados e Municípios, o que fragiliza a cadeia de informações do sistema de saúde, trazendo danos, prejuízos às ações de prevenção e controle, colocando em risco a saúde da população [12]. Neste sentido, não é incomum a imprensa fazer a divulgação de surto de uma doença antes mesmo de o MS ter recebido todas as informações de notificações, o que tem causado desconforto ao Órgão.

Por outro lado, outras fontes de informações têm se destacado durante o processo de investigação, dentre as quais encontram-se as informações clínicas e exames laboratoriais que são realizados pela rede de laboratórios de saúde pública [13].

Dada a importância dos exames laboratoriais para o desenvolvimento de ações em saúde pública, o Ministério da Saúde, por meio da Portaria nº 2.031/2004[3], criou o Sistema Nacional de Laboratórios de Saúde (SISLAB), que é um conjunto de redes nacionais de laboratórios (Laboratórios de Referências Nacionais, de Fronteiras, Estaduais e Municipais), organizado em subredes, por agravos ou programas, de forma hierarquizada por grau de complexidade (baixa, média e alta) das atividades relacionadas à vigilância em saúde.

[2]Portaria nº 204/2016 MS: define a lista nacional de notificação compulsória de doenças, agravos e eventos de saúde pública nos serviços de saúde públicos e privados em todo o território nacional, nos termos do anexo, e dá outras providências.

[3]Portaria nº 2.031/2004: dispõe sobre a organização do Sistema Nacional de Laboratórios de Saúde Pública.

A Coordenação Geral de Laboratórios de Saúde Pública (CGLAB), unidade do Ministério da Saúde, é a responsável pela avaliação e controle das Redes de Laboratórios de Saúde Pública, conforme critérios estabelecidos na Portaria nº 70/MS[4] de 23 de dezembro de 2004. A CGLAB tem como parte de suas políticas a promoção, coordenação, o apoio e fomento de ações, objetivando a melhoria contínua dos serviços prestados pelos Laboratórios de Vigilância em Saúde Ambiental e Epidemiológica.

Uma das iniciativas da CGLAB foi o desenvolvimento de um sistema de informações, chamado Sistema Gerenciador de Ambiente Laboratorial (GAL), para armazenar os dados produzidos pelo Sistema Nacional de Laboratórios de Saúde. Esses dados referem-se a todos os exames realizados pelos laboratórios, relacionados às doenças e agravos de notificações. Entretanto, ressalta-se que o GAL encontra-se em processo de implantação no Estado de São Paulo e no Distrito Federal, o que impacta na obtenção dos dados destes entes federativos [14].

Apesar das informações dos exames produzidos pelos laboratórios serem extremamente valiosas para a efetividade das ações em vigilância em saúde, até o presente momento esses dados tem sido pouco utilizados pela Secretaria de Vigilância em Saúde (SVS). Assim, a CGLAB tem buscado formas de explorar tais dados para fomentar a vigilância em saúde de modo que esses registros possam, também, ser usados no controle e monitoramento de surtos, epidemia, incidência e detecção de casos.

Desta forma, considerando a importância das informações laboratoriais, esta pesquisa pretende utilizar os dados armazenados no Sistema Gerenciador de Ambiente Laboratorial para criar mecanismos de controle e monitoramento de vigilância em saúde da doença de influenza, no município de Curitiba-PR.

A doença de influenza foi escolhida para compor este trabalho por possuir uma rede de vigilância muito bem estruturada. Esta é composta, além dos serviços de atendimentos de rotinas, pela vigilância sentinela de Síndrome Gripal (SG)[5], de Síndrome Respiratória Aguda Grave (SRAG)[6] em pacientes internados em Unidade de Terapia Intensiva (UTI) e pela vigilância universal de SRAG, sendo estas, distribuídas em todas as regiões geográficas do país. No total, conta com 252 unidades sentinelas da gripe ativas, destas 140 são sentinelas de SG, 112 sentinelas de SRAG e 17 unidades de saúde sentinelas tanto de SG quanto de SRAG [15] [16].

[4]Portaria nº 70/2004 MS: estabelece os critérios e a sistemática para habilitação de Laboratórios de Referência Nacional e Regional para as Redes Nacionais de Laboratórios de Vigilância Epidemiológica e Ambiental em Saúde

[5]SG: indivíduo com febre, mesmo que referida, acompanhada de tosse ou dor de garganta e início dos sintomas nos últimos 07 dias.

[6]SRAG: indivíduo hospitalizado com febre, mesmo que referida, acompanhada de tosse ou dor de garganta e que apresente dispneia. Também podem ser observados os seguintes sinais: saturação de O2 menor que 95% ou desconforto respiratório ou aumento da frequência respiratória.

As amostras coletadas pela rede de vigilância sentinela da Influenza são distribuídas entre os Laboratórios Centrais de Saúde Pública (LACEN), os Laboratórios de Referência Regional (LRR) e Laboratório de Referência Nacional (LRN). Estes dois últimos são responsáveis pelas análises complementares às realizadas pelos LACEN. Os LRN fazem parte da rede global da Influenza e enviam anualmente *isolados virais e amostras clínicas* para Centro de Controle de Doenças (CDC) dos Estados Unidos da América e para o Centro Colaborador da Organização Mundial da Saúde (OMS) das Américas, para subsidiar a seleção das estirpes virais para a composição da vacina anual pela OMS [15] [16].

Quanto a escolha do município, esta foi realizada com base no quantitativo de exames produzidos por esses entes federativos. Após análise, verificou-se que Curitiba-PR foi o município que apresentou o maior número de exames realizados nos últimos 9 anos.

## 1.2 Justificativa da Pesquisa

Esta pesquisa é motivada pela possibilidade de:

1. se construir um sistema de Vigilância em Saúde com base nos resultados dos exames de média e alta complexidade que são produzidos pela rede de laboratórios de saúde pública.
2. se criar mecanismos de controles e de monitoramento do comportamento epidemiológico de doenças e agravos, relativas ao campo laboratorial da Rede Nacional de Laboratórios de Vigilância Epidemiológica, que são estabelecidos na portaria nº 1.419/MS[7], de 08 de junho de 2017.

Acredita-se que este trabalho contribuirá para o fortalecimento das ações de vigilância em saúde.

## 1.3 Escopo

Considerando o universo das doenças e agravos à saúde que são monitoradas pelo MS, este trabalho pretende utilizar as técnicas estatísticas de Séries Temporais, Controle Estatístico de Processo e linguagem de programação **R**, integrada a à ferramenta de *Business Intelligence (BI)*, para monitorar o comportamento dos resultados dos exames da Influenza, no município de Curitiba-PR.

[7]Portaria nº 1.419/MS: aprova os regimentos internos e o quadro demonstrativo de cargos em comissão e das funções de confiança das unidades integrantes da estrutura regimental do Ministério da Saúde.

## 1.4 Objetivos Geral

O objetivo geral desta pesquisa é propor uma arquitetura de software que possa ser utilizada no processo de monitoramento e controle de doenças e agravos à saúde pública, a partir dos resultados dos exames de média e alta complexidade que são produzidos pelos Laboratórios Centrais de Saúde Pública.

Quanto aos objetivos específicos, busca-se:

1. propor um indicador de vigilância em saúde que auxilie no processo de monitoramento da doença de Influenza a partir dos dados laboratoriais.
2. selecionar e parametrizar um Gráfico de Controle Estatístico para o indicador de vigilância em saúde para a Influenza.
3. propor ferramenta de visualização de dados para monitorar o comportamento da doença de Influenza, no município de Curitiba-PR.

## 1.5 Estrutura da Dissertação

O restante deste trabalho está estruturado em quatro capítulos. O Capítulo 2 trata da metodologia utilizada. O Capítulo 3 aborda o referencial teórico relevante para a pesquisa. O Capítulo 4 apresenta o estudo de caso, com aplicação das técnicas estatísticas de Séries Temporais e Gráfico de Controle, além de um modelo para a visualização dos dados. Por fim, são apresentadas as conclusões e recomendações da pesquisa.

# Capítulo 2

# Metodologia

A metodologia utilizada para o desenvolvimento deste trabalho contempla as informações relativas à sua classificação, estruturação das atividades, técnicas estatísticas, coleta de dados, principais ferramentas e softwares utilizados.

## 2.1 Classificação da Pesquisa

De acordo com os objetivos deste trabalho e o conhecimento que se espera gerar a partir dele e, ainda, os direcionamentos do trabalho de Silva e Menezes [17], esta pesquisa classifica-se sob os ângulos de sua natureza, abordagem, objetivos e procedimentos.

Classificação quanto: [17]

1. **sua natureza:** classifica-se como ***Pesquisa Aplicada***, pois visa a geração de conhecimento para a aplicação prática, buscando a solução do problema definido neste trabalho e, ainda, o interesse da Coordenação Geral de Laboratórios de Saúde Pública na melhoria dos processos de monitoramento de doenças e agravos à saúde pública.
2. **a forma de abordagem do problema:** classifica-se em ***Pesquisa quantitativa***, porque serão utilizadas técnicas estatísticas como parte da solução do problema proposto na pesquisa.
3. **aos seus objetivos:** esta pesquisa enquadra-se como ***Pesquisa Exploratória***, pois nos processos de entendimento do problema, base e panorama geral das doenças e agravos foram utilizados pesquisa documental e levantamento bibliográfico, culminando com a proposição de soluções para o problema apresentado na pesquisa.
4. **aos procedimentos técnicos:** esta pesquisa classifica-se como ***Estudo de caso***, pois, concentra estudo intenso e aprofundado no problema e na busca de

soluções, de forma que atenda aos objetos e permita um extenso e detalhado conhecimento para aplicação real.

Dada as circunstancias apresentadas neste trabalho, sua classificação predominante é Estudo de Caso.

## 2.2 Estruturação da pesquisa

Esta pesquisa foi estruturada em:

1. **Referencial Teórico:** aborda o referencial teórico relevante para a pesquisa, no que tange à contextualização das doenças; indicadores de vigilância em saúde; aplicações de técnicas estatísticas de séries temporais; Controle Estatístico de Processo e ferramentas de visualização de dados.
2. **Indicador de Saúde:** consiste no processo de proposição de um indicador de saúde que comporá o Controle Estatístico de Processo da Influenza.
3. **Séries Temporais:** implementação da análise de Séries Temporais com base no indicador proposto e informações laboratoriais.
4. **Controle Estatístico de Processo:** implementação do Controle Estatístico de Processo para o indicador de vigilância em saúde da Influenza.
5. **Monitoramento:** proposição de ferramenta de visualização de dados, integrada com a linguagem de programação ***R***, com uso das técnicas de estatísticas de Séries Temporais e do Controle Estatístico de Processo para monitorar a doença/agravo de Influenza no município de Curitiba-PR.

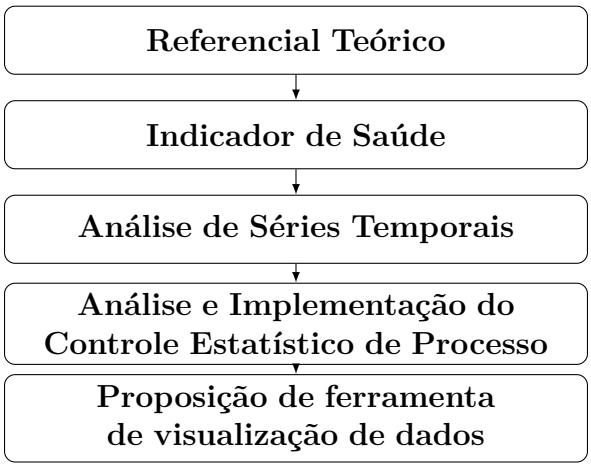



Figura 2.1: Fases da Pesquisa
(Fonte: [18])

## 2.3 Referencial Teórico

A revisão da literatura foi realizada por meio de consultas bibliográfica e eletrônica nas bases de periódicos científicos, como por exemplo, Biblioteca Virtual em Saúde (BVS), Scientific Electronic Library Online (SciELO), Web of Science. Ainda, pesquisas em sítios e documentos técnicos de organismos nacionais e internacionais de saúde, como a Organização Mundial da Saúde (OMS), Organização Mundial de Saúde Animal (OIE), Organização Pan-americana de Saúde (OPAS), Centers for Disease Control and Prevention (CDC) e Ministério da Saúde (MS).

Na pesquisa bibliográfica utilizou levantamento textual referentes à vigilância em saúde, indicadores epidemiológicos de saúde, análise de Séries Temporais (ARIMA, SARIMA), Controle Estatístico de Processo, mais especificamente sobre os gráficos de Shewhart, Soma Acumulada e Médias Móveis Exponenciais Ponderadas. Ainda, foi realizado descritores a respeito da vigilância laboratorial e epidemiológica das doenças e agravos de notificações, ferramentas, técnicas e inteligência em negócios *(Business Intelligence (BI))*

## 2.4 Indicador de saúde (Surto)

Na seleção e proposição do indicador de vigilância em saúde, foram utilizados a pesquisa documental e experiências bem-sucedidas da Secretaria de Vigilância em Saúde.

1. **Pesquisa documental:** foi realizada por meio de consultas eletrônicas nas bases de periódicos científicos, sítios, boletins e documentos técnicos de organismos nacionais e internacionais de saúde, como Ministério da Saúde, Organização Mundial da Saúde, Organização Pan-americana de Saúde e Centro de Controle de Doenças.
2. **Experiências bem-sucedidas:** foram realizadas reuniões com os grupos técnicos de vigilância em saúde da Coordenação Geral de Doenças Transmissíveis (CGDT) e da CGLAB, ambas pertencentes ao MS. As reuniões foram conduzidas e registradas pelo pesquisador visando o entendimento do negócio, as principais variáveis e o indicador de monitoramento de surto. Os Informes Epidemiológicos (Boletins), guia de vigilância em saúde publicado em 2017 pelo MS e o resumo dos dados estatísticos das informações que são armazenadas no Sistema Gerenciador de Ambiente Laboratorial foram as principais ferramentas de apoio às reuniões.

   Os Grupos Técnicos foram compostos por gestores e técnicos de vigilância em saúde do Ministério da Saúde.

## 2.5 Análise da Série Temporal do Indicador de Saúde

A Análise da Série Temporal para Indicador de Saúde foi realizada pela metodologia Box & Jenkins. O método foi utilizado na proposição e ajustes do modelo paramétrico da série (estacionária ou não estacionária) temporal observada a partir dos dados produzidos pelos Laboratórios Centrais de Saúde Pública e armazenados no Sistema Gerenciador de Ambiente Laboratorial, utilizando como base a data da coleta do material utilizado para a realização dos exames dos pacientes.

Foi utilizado o teste de *Dickey-Fuller* com o objetivo de verificar se a série era ou não estacionária.

O teste *Ljung-Box* foi utilizado para verificar a correlação dos dados (observações).

A estratégia adotada para o processo de proposição e ajustes do modelo foi realizada por ciclo iterativo, por meio das fases de identificação, estimação e diagnóstico, que compõem a metodologia *Box & Jenkins*.

1. **Identificação do modelo:** inicialmente, os dados foram distribuídos graficamente visando o entendimento destes, ou seja, perceber se eles eram estacionários ou não estacionários, se havia tendência e ou componente sazonal. A partir da análise inicial, aplicou-se as Funções de Auto-correlação (FAC) e Função de Auto-correlação Parcial (FACP) para identificar o padrão das informações da série, aplicando diferenciações nos dados até que este atingisse a estacionaridade. Observando-se o componente sazonal, utilizou-se o produto dos componentes não sazonais e sazonais ((p, q e d) e (P, Q e D)).
2. **Estimação dos parâmetros:** nesta fase foi utilizado o princípio do método de máxima verossimilhança, o que permitiu a estimação da população que mais se assemelha com a amostra observada.
3. **Validação do modelo:** a identificação e validação do modelo mais adequado foi realizada por meio dos critérios de Critério de Informação de Akaike Corrigido (AICc) e Critério de Informação Bayesiano (BIC), considerando o principio da parcimônia, ou seja, levou-se em consideração o modelo com menos parâmetros e graus de liberdade.
4. **Previsão:** foi calculada e está condicionada aos valores das observações passadas, tendo assim o valor esperado para o horizonte de um período de tempo projetado. Os valores são obtidos pela equação do modelo SARIMA.

**Obs.:** Estas etapas foram realizadas por meio do software ***R Studio***, linguagem de programação **R**. Ressalta-se, que as etapas de 1 a 3 foram desenvolvidas por *Veloso* du-

rante a pesquisa de conclusão da graduação em Estatística no Departamento de Estatística da UnB [19].

A Figura 2.2 visa melhorar o entendimento das fases descrita acima.

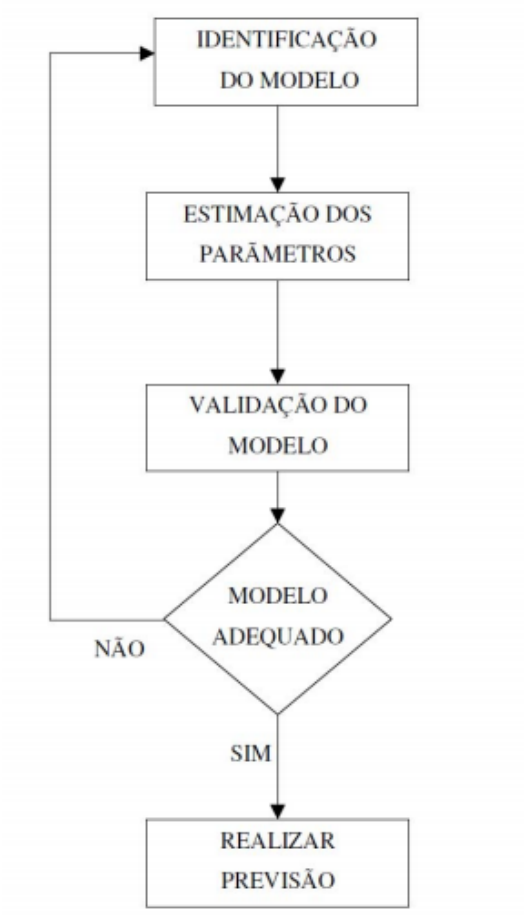



Figura 2.2: Metodologia de Box & Jenkins
(Fonte: [20]).

## 2.6 Gráficos de Controle Estatístico de Processo para o Indicador de Saúde

Com base nas informações obtidas pela análise da Série Temporal do Indicador de Saúde, foi proposto o gráfico de Controle Estatístico de Processo.

A partir da descoberta de tendência e de sazonalidade na série de dados, foi definida a informação (do indicador) que seria plotada no Gráfico de Controle, ou seja, se não houvesse tendência e/ou sazonalidade seria utilizado o indicador direto no Gráfico de Controle, caso contrário, e foi o que ocorreu, seria utilizado os resíduos do modelo ajustado.

A escolha do tipo de gráfico de controle foi baseada no referencial teórico levantado neste estudo e testes estáticos realizados na fase da análise da Serie Temporal do indicador de saúde (surto), o que levou ao entendimento de que o Gráfico de Controle de Média Móvel Exponencial Ponderada (MMEP) seria o mais indicado para o Controle Estatístico de Processo (CEP).

Outro critério utilizado na escolha do gráfico foi a necessidade de se monitorar pequenas alterações no processo com base nos dados históricos das amostras.

Com a definição do tipo de Gráfico de Controle, o próximo passo foi a definição dos parâmetros do Gráfico de Controle de Média Móvel Exponencial Ponderada. Os valores da média, variância, fração de não-conformidades, entre outros que tem influência direta no Limite Central do Controle (LC), são representados pelo valor MMEP da série de dados, o Limite Superior do Controle (LSC) e o Limite Inferior do Controle (LIC), ambos estimados conforme proposto por Montgomery [21].

Os limites de controles foram calculados a partir da análise do Comprimento Médio da Sequência (CMS).

A elaboração dos gráficos de controle foi precedida da metodologia padrão 3 $\sigma$, adotando as sugestões de Montgomery [21].

## 2.7 Visualização dos Dados

Nesta fase, utilizou-se a pesquisa documental, experiências bem-sucedidas da Secretaria de Vigilância em Saúde e técnicas computacionais.

1. **Pesquisa documental:** foi realizada por meio de consultas eletrônicas nas bases de periódicos científicos, sítios, boletins e documentos técnicos.
2. **Experiências bem-sucedidas:** foram realizadas buscas em sistema de controle de versões distribuídas - *GIT-Hub*, *RPubs - publishing from* ***R***, soluções tecnológicas em *Python/Django* com implementações *Chatjs, Jquery, Canvas, Highcharts*, técnicas de *Extração, Tratamento e Carga* e pesquisas nas comunidades de tecnologia da informação.
3. **Softwares e técnicas computacionais:** foram aplicados os software de *Pentaho Data Integration (PDI)* para a *ETL* dos dados, com conexões nos bancos de dados PostgreSQL e Oracle; RStudio para executar a programação em linguagem ***R***, das Séries Temporais e Gráfico de Controle; e o *Elasticsearch, Logstash e Kibana (ELK)* para carregamento, indexação e visualização do Gráfico de Controle do Controle Estatístico de Processo da Influenza, no município de Curitiba-PR, de forma automatizada e *online*.

4. **Processamento da Informação:** foi utilizada computador com processador Intel i7, 16GB de RAM e 1TB de capacidade de armazenamento de dados.
5. **Fonte de Dados:** os dados foram fornecidos pelo Ministério da Saúde e são armazenados no Sistema Gerenciador de Ambiente Laboratorial. Eles são compostos por todos os exames de média e alta complexidade que são realizados pelo Sistema Nacional de Laboratórios de Saúde.

# Capítulo 3

# Referencial Teórico

O referencial teórico desta pesquisa visa estabelecer os principais conceitos, técnicas e ferramentas adotadas no processo de vigilância em saúde, com a finalidade de reforçar e direcionar o desenvolvimento da pesquisa.

## 3.1 Vigilância em Saúde

Segundo Choi [22], os primeiros registros de vigilância ocorreram no Egito, com a primeira epidemia na história humana conhecida como uma grande pestilência durante o reinado do Faraó Mempses, na Primeira Dinastia.

Choi [22] apresenta um pequeno resumo histórico da formação da Vigilância em Saúde Pública:

> ...vigilância em saúde iniciou-se com o registro da primeira epidemia em 3180 A.C., no Egito; John Graunt (1620-1674) introduziu a análise de dados sistemática; Samuel Pepys (1633-1703) começou a investigação de campo epidêmica; William Farr (1807-1883) fundou o conceito moderno de vigilância; John Snow (1813-1858) ligou dados à intervenção e Alexander Langmuir (1910-1993) deu a primeira definição abrangente de vigilância...

Portanto, fazer vigilância em saúde é um ato de monitoramento, ações e políticas no sentido de neutralizar qualquer epidemia[1], ou seja, é coletar dados, analisar, interpretar, investigar, divulgá-los rotineiramente para que seja possível realizar planejamento, implementação e avaliação das ações de saúde pública [5].

A Vigilância em Saúde Pública desempenha um papel primordial para a saúde humana. Esta consiste na maior arma para evitar epidemias. O sistema de saúde pública possui

[1]Epidemia: doença de caráter transitório, que ataca simultaneamente grande número de indivíduos em uma determinada localidade

cinco papéis essenciais: avaliação da saúde, vigilância da saúde, promoção da saúde, prevenção de doenças e proteção da saúde [22] [4].

No Brasil, a vigilância em saúde é uma atribuição/competência do Sistema Único de Saúde. A saúde é considerada um bem social, oriundo das interconexões políticas, econômicas, culturais e ideológicos. Neste contexto, a vigilância tem acontecido como uma estratégia e modelo de atenção à saúde, proveniente de reformas sanitaristas brasileiras, visando a organização logística, serviços e as relações entre ofertas e demandas da população [4][5] [23]. A vigilância em saúde foca a coletividade, porém, sua formulação remete a duas fronteiras, sendo:

> ...*primeira*, de caráter amplo, reitera as ações para a contenção das doenças, mas requer visão ampliada da saúde-doença ao incorporar a interpretação dos determinantes, à luz da epidemiologia crítica. A *segunda* concepção, considerada mais restrita, compreende a vigilância da saúde como um conjunto de ações voltadas para o conhecimento, a previsão, a prevenção e o enfrentamento dos problemas de saúde selecionados e relativos aos fatores e condições de risco, significando, apenas, uma ampliação modesta da vigilância epidemiológica [24].

Por outro lado, a vigilância tem três vertentes no tocante ao foco de ação:

> A *primeira* equivale à análise de situações de saúde, que toma como objeto as situações de saúde dos distintos grupos populacionais em função de suas condições de vida. A *segunda* vertente coloca a possibilidade de integração institucional entre a Vigilância Epidemiológica e a Vigilância Sanitária. E a *terceira* refere-se a uma redefinição das práticas sanitárias, integrando duas dimensões: uma técnica, que resulta na concepção da Vigilância da Saúde como modelo assistencial alternativo conformado por um conjunto de práticas sanitárias que encerram combinações tecnológicas distintas e destinadas a controlar determinantes, riscos e danos e uma outra dimensão que enfatiza a gerência, caracterizando a Vigilância da Saúde como uma prática que organiza os processos de trabalho em saúde [24].

Historicamente, as três epidemias consideradas devastadoras que atingiram a humanidade foram: “a Praga de Justiniano” (541-591 dC), que durou 50 anos; “A Morte Negra” (1348-1351), que durou 4 anos, e “Influenza espanhola” (1918), que durou cinco meses[22].

A Tabela 3.1 traça um panorama das principais epidemias registradas na história:

Tabela 3.1: Principais epidemias registradas na história

| **Ano*** | **Lugar** | **Evento** |
|---|---|---|
| 3180 A.C. | Egito | Primeira epidemia registrada: “Uma grande pestilência” durante o reinado do Faraó Mempses na Primeira Dinastia, foi a primeira epidemia registrada na história humana |
| 1495 A.C. | Egito | “A Peste do Faraó”, possivelmente causada pela seca. |
| 1471 A.C. | Kadesh | A praga causando 14.700 mortes, possivelmente causada por terremoto. |
| 1190 A.C. | Grécia | A loimos (grego, que significa uma praga ou peste), agora acredita-se ser uma peste bubônica, possivelmente causada pela Guerra de Troiá (1194-1184 A.C.). |
| 1017 A.C. | Israel | A pestilência “3 dias", causando 70.000 mortes. |
| 431 A.D-427 A.C. | Etiópia, depois se espalhou para o Egito, o Império Persa e Atenas | “A Praga de Tucídides”, agora acredita-se ser tifo e sarampo, possivelmente causada pela Guerra do Peloponeso (432-411 A.C.). |
| A.D. 166 | Roma | Possivelmente varíola, espalhada pelos soldados que retornam da guerra Parthian (A.D. 161-166). |
| 541-549 | Constantinopla, posteriormente espalhou-se para o Egito e todo o mundo povoado | Primeira das três epidemias mais devastadoras para atingir a raça humana: "A Peste de Justiniano”. |
| 664-689 | Inglaterra | “A praga amarela”, agora acreditado ser a febre recidivante com icterícia, causando a morte de "uma grande multidão de homens". |

| **Ano*** | **Lugar** | **Evento** |
|---|---|---|
| 1348-1351 | Ásia Central, depois se espalhou para o leste para a China, sul para a Índia, oeste para Portugal, norte para a Inglaterra (1349), Noruega (1350), e Rússia (1351) | Segunda de três epidemias mais devastadoras para a raça humana foi “a morte negra”, hoje acredita ser a peste bubônica. Ela matou milhares de pessoas, possivelmente causada por navios contaminados seguindo as rotas comerciais. A quarentena foi utilizada para reter viajantes de áreas infectadas. |
| 1374-17º século | Alemanha (1374), espalhou-se então pela França (1518) e Itália (século XVII) | "Dancing Mania", possivelmente causada por transtorno psicogênico de massa e / ou a mordida de uma aranha. |
| 1665 | Londres | “A Grande Peste de Londres”, causada por condições sanitárias pobres, população densa, e habitação superlotada. A epidemia foi terminada por intervenções naturais, com geadas de inverno e o “Grande Fogo de Londres” em 1666 que destruiu e limpou os bairros. |
| 1817-1875 | Calcutá (1817), toda a Índia (1821), China (1820), Japão (1822), Rússia (1823), Inglaterra (1831), Canadá e Estados Unidos (1832), África (1837) e América do Sul (1875) | Quatro pandemias de cólera (1817-1823; 1826-1837; 1846-1863; 1863-1875), causadas por barcos a vapor e migração em massa durante a Revolução Industrial. Em 1849, John Snow mapeou casos de cólera em Londres e encontrou água contaminada da bomba da Broad Street. A neve removeu o cabo da bomba em 1854 e a epidemia diminuiu. |

| **Ano*** | **Lugar** | **Evento** |
|---|---|---|
| 1918 | França (Abril), Inglaterra (Junho), China (Julho) e EUA (Agosto) | Terceira das três epidemias mais devastadoras que atingiu a raça humana: "Influenza espanhola", causada por um vírus. A doença matou 22 milhões de pessoas, cerca do dobro das 10 milhões de mortes causadas pela Primeira Guerra Mundial (1914-1918). O vírus foi isolado em 1933, e sua vacina foi desenvolvida em 1972. Uso de máscaras e lavar as mãos têm sido sugeridos para prevenir a propagação da gripe. |
| 1940 - agora | No mundo todo | Epidemia de câncer de pulmão, causada pelo tabagismo. Na década 1990-1999, houve um total de 6,6 milhões de mortes atribuíveis em todo o mundo. |
| 1997 agora | No mundo todo | Epidemia de obesidade, causada por uma combinação de excesso de ingestão de alimentos, falta de atividade física e susceptibilidade genética. Antes do século XX, a obesidade era rara. Em 1997, a Organização Mundial da Saúde reconheceu formalmente a obesidade como uma epidemia global. Estima que 1,4 bilhões de adultos estão com sobrepeso ou obesos, e 2,8 milhões de adultos morrem todos os anos como resultado de excesso de peso ou obesidade. |

Fonte: Paper - The Past, Present, and Future of Public Health Surveillance [22].

A Tabela 3.1[2] mostra a gravidade e a letalidade de algumas doenças quando não são monitoradas corretamente. Entretanto, nos últimos anos tem-se observado inúmeras ações e tecnologias que vêm melhorando a vigilância em saúde.

Na área laboratorial, a tecnologia e a ciência estão transformando rapidamente os ambientes laboratoriais de saúde, trazendo novos métodos de diagnósticos, proporcionado a redução do tempo no diagnóstico. É possível ter resultados de forma ágil e em tempo oportuno, permitindo a identificação avançada de agentes patogênicos, vírus, cepas em circulações, novas doenças e agravos à saúde humana [5] [14].

[2]Ano*: a coluna "Ano" na Tabela 3.1, refere-se ao momento em que uma epidemia foi relatada pela primeira vez em um lugar. Ela pode se repetir em um local posterior ao ano citado.

Cabe ressaltar, ainda, que a Tecnologia da Informação (TI), a matemática e a estatística, quando bem utilizadas, fornecem suporte adequado para a coleta, processamento e divulgação das informações ao redor do mundo. Neste sentido, novos métodos estatísticos continuam sendo desenvolvidos, refinados e aperfeiçoados, enquanto a computação, por sua vez, realiza o trabalho de consolidação e produz resultados que possibilitam a vigilância intervir rapidamente em investigações complexas, como por exemplo, a influenza pandêmica [14] [22] [5] .

Segundo Choi [22], com a inclusão tecnológica, surgiram novas formas e áreas que buscam reforçar a vigilância, tais como:

> ...o processo sistemático de seleção de indicadores; metodologia para converter os resultados de diferentes inquéritos de saúde com diferentes definições de indicadores a um nível padrão e compatível; metodologia para aumentar as taxas de resposta aos inquéritos por subgrupos populacionais; incorporação de dados laboratoriais na vigilância da saúde da população de rotina; desenvolvimento de notificações automáticas, laboratoriais e eletrônicas de doenças... [22].

Nos últimos anos, o uso da estatística, como aplicações de métodos de séries temporais, permitiu analisar e interpretar com mais precisão os dados de vigilância. Pode-se citar, ainda, outras técnicas que contribuem significativamente para a vigilância de doenças, como por exemplo, métodos geográficos, espaciais e temporais. Entretanto, deve-se observar métodos de ponderação, visando ajustes dos dados, tais como: sexo, idade, raça, estado civil e outros [25].

Velasco et. al [25] conceituou a Vigilância em Saúde Pública em:

1. **Vigilância baseada em eventos:** baseia-se na captura das informações de forma organizada e rápida sobre os eventos que oferecem riscos a saúde.
2. **Vigilância baseada em indicadores:** que geralmente são baseados em métodos estatísticos para comparar casos de patógenos com taxas. O objetivo é encontrar números maiores de cluster, em um horário, local ou período específico, visando indicar uma ameaça. Para isso, são utilizados parâmetros epidemiológicos, sazonalidade, fatos de cada doença. Os algoritmos estatísticos são ajustados para aumentar o nível de sensibilidade e eficiência.

A diferença da vigilância baseada em eventos e a vigilância baseada em indicador, é que na primeira o processo de coleta das informações é realizado diretamente ou indiretamente das pessoas que testemunharam ou foram afetadas pelo evento. Essa coleta utiliza fontes da Internet como mídias sociais, aplicativos e canais de comunicações. A segunda tem seus pilares no processo de coleta por meios de sistemas de informações convencionais.

A vigilância baseada em eventos detecta casos mais rápido, possui menos dados estruturados e estes são coletados, filtrados por meios de técnicas de mineração de dados e por estruturas pré-estabelecidas pela vigilância. Contudo, há um prejuízo para os eventos em que este não foi mencionado nas fontes de coleta, ou seja, haverá pessoas que não tem acesso a celulares mais modernos, Internet ou hábito de comunicação [25].

Quanto aos sistemas baseados em indicadores, pode-se dizer que são concebidos, de forma estruturada, para coleta e análise de dados com base em procedimentos de vigilância para cada doença. Estes protocolos são definidos por profissionais da área de saúde e visam entender e detectar as tendências das doenças e agravos [25].

Por outro lado, há de se considerar o tempo no processo de coleta das amostras e das informações, o que pode ser um limitador para a descoberta de surtos, que poderão ser percebidos antes do processamento das informações de saúde [25].

Pode-se, também, combinar as duas técnicas para obter bons resultados de vigilância, essa combinação é conhecida como "inteligência epidêmica"[25].

## 3.2 Indicadores

Indicadores, segundo Minayo [26], são parâmetros quantitativos ou qualitativos que dão suporte ao processo de avaliação dos objetivos aspirados, indicando se o objeto monitorado está evoluindo dentro ou fora dos padrões e se foram atingidos ou não os objetivos propostos. Eles traduzem a ideia de sinalizadores dos fatos almejados, expressando em pluralidade do tempo, o sentido de medida e marcação do objetivo pretendido. Indicadores são utilizados para aferir a abrangência de objetivos, metas e resultados.

Dessa forma, eles constroem unidades para aferir ou tornar visíveis estruturas cobiçadas durante o planejamento, e podem ser considerados:

> ...instrumentos para mensurar a disponibilização de bens e atividades, assim como para conceber parâmetros de acesso de diferentes atores a um programa, a relevância que ele possui para a vida de cada um, sua intensidade e seu sentido... [26].

Para Trevisan e Van Bellen [27], indicador

> ...é derivado da palavra latina indicare, que significa tornar patente; demonstrar, revelar, denotar; expor... [27].

Entretanto, é necessário entendê-lo como parâmetro que provê informações a respeito de uma condição de estado. Sua aplicação está diretamente ligada ao processo tático e operacional, fonecendo discernimentos aos resultados a serem acompanhados (esperados) [27].

Portanto, ele pode ser visto como fonte de medição, que visa guiar os avaliadores no processo de intervenção, aprimoramento, monitoramente e obtenção do resultado esperado. Segundo Minayo [26] um indicador de qualidade deve seguir ou atender alguns requisitos:

> a) que estejam normalizados e que sua produção histórica (sua temporalidade) se atenha sempre à mesma especificação ou forma de medida, permitindo a comparabilidade;
> b) que sejam produzidos com regularidade, visando à formação de séries temporais e permitindo visualizar as tendências dos dados no tempo;
> c) que sejam pactuados por quem (grupos, instituições) os utiliza e quem pretende estabelecer comparabilidade no âmbito nacional e até internacional;
> d) que estejam disponíveis para um público amplo e de forma acessível, propiciando à opinião pública um formato simples de acompanhamento do desempenho de instituições e de políticas públicas ou que recebam financiamento público [26].

Tronchin et. al [28] entende que a elaboração de indicadores inicia-se pelo conceito, descrição de um estado, vislumbrando o processo de avaliação da disposição ou modificação de uma situação durante um intervalo de tempo, nas ações de saúde a serem realizadas.

Indicador é um instrumento que possibilita o alcance de conhecimentos a respeito de uma realidade. Este pode ser um fenômeno individual ou um aglomerado de informações, possuindo os seguintes atributos:

> ...simples de entender; quantificação estatística e lógica coerente; e comunicar eficientemente o estado do fenômeno observado [29].

. Para tanto, é necessário observar:

> ...o objetivo, a equação, a população ou amostra, o tipo, a fonte de informação, o método para coletar dados, a frequência e os fatores avaliativos da variação... ....quanto aos tipos de indicadores, a literatura considera diferentes classificações; na saúde, um dos mais adotados é o evento sentinela (vigilância): caracterizado pela seriedade do evento e pelo grau através do qual pode ser evitado um risco. Este indicador mede processos ou acontecimentos graves, indesejados e eventualmente evitáveis, tais como, monitoramento dos vírus respiratórios circulantes, com por exemplo, influenza H1N1, H7N9, zika, dengue... [28].

Para obter um indicador de qualidade deve-se levar em conta um leque de informações que visem melhorar sua aplicação, sendo:[28]

1. **Descrição do indicador:** deve descrever com clareza e fidedignidade todas as funções que serão apresentadas ou realizadas por ele;
2. **Termos do indicador:** os termos utilizados na definição do indicador devem assegurar que todos o utilizarão da mesma forma e para o que foi proposto;

3. **Identificação do tipo de indicador:** esse processo classifica o indicador de acordo com o evento que se espera medir, monitorar, como por exemplo, proporção, taxa, coeficiente etc.
4. **Princípio/razão:** deve-se descrever os motivos e a utilidade do indicador no processo de monitoramento e avaliação dos eventos;
5. **Descrição da população:** a partir dos atributos, técnicas ou fatos deve-se organizar a população de forma categórica para que seja possível medir ou avaliar as ocorrências desta;
6. **Lógica do indicador:** deve-se criar uma cronologia que reflita a linha de recuperação e associação dos dados de forma a facilitar a avaliação dos casos identificados;
7. **Fatores relevantes:** são os principais fatos e variantes que incidem diretamente sobre os dados, podem explicá-lo ou induzir soluções do problema.

Entretanto, no processo de validação de indicador, deve-se levar em consideração as seguintes evidências: [28]

1. **Acurácia:** é a qualidade da informação apresentada, aferida pelo indicador em relação aos resultados esperados; e
2. **Precisão:** estar pautado na capacidade de reprodução, credibilidade e coerência da avaliação.

Ainda, devem ser considerados os atributos de: [28]

1. **Validade:** o indicador alcança os objetivos para os quais ele foi proposto;
2. **Atribuível:** é a capacidade que o indicador possui para demonstrar as características do evento ao qual se relaciona;
3. **Credibilidade:** denota a importância e a aceitação do indicador;
4. **Sensibilidade:** é a capacidade do indicador em alcançar todas as possibilidades possíveis dos eventos;
5. **Especificidade:** é a condição em que o indicador identifica exclusivamente o caso para o qual foi proposto, não incluindo os demais;
6. **Acessível:** facilidade na obtenção dos recursos, dados imprescindíveis que compõem o cálculo do indicador;
7. **Comunicável:** de fácil explicação;
8. **Efetiva:** as funções são coerentes e desempenham o papel para o qual foi proposto;

9. **Exequível:** a capacidade de execução conforme foi projetado.

Na área da saúde, os indicadores são utilizados de forma abrangente, categorizados e com ênfase em cada situação. Eles assinalam avaliações quantitativas ou qualitativas, capazes de demonstrar informações que não se revelam por si só. O uso quantitativo prevalece sobre o qualitativo tais como a contagem, a incidência, os casos de doenças e agravos. Entretanto, números absolutos podem expressar valores agrupados, representando a coletividade (grupos específicos) [30].

Normalmente, um indicador é uma variável numérica sendo apresentado como absoluto ou uma relação entre dois fatos, sendo denominador e numerador. O primeiro reflete a população e o último explicita a medida do evento, devendo ser expresso de forma objetiva, clara e relevante para o serviço [28].

A medida do tipo proporção é uma medida matemática em que o subconjunto do numerador é um subconjunto do denominador. Eles possuem características de admissão, ou seja, o resultado é calculado com ausência (podendo ser "zero") ou ocorrência (podendo ser "um"). Entretanto, na epidemiologia, a proporção pode ser aplicada sobre o valor total do item. Ainda, podem-se usar proporções expressas em frações decimais multiplicados por 100, 1.000 ou outros múltiplos de 10 para representar os resultados. Pode-se, também, usá-la em favor da formação de grupos, como por exemplo, faixa etária, sexo [30].

Por outro lado, coeficiente, em epidemiologia, é a medida cuja proporção de um evento do numerador concebe certo risco de ocorrência em relação ao denominador. Podendo ser aferido de duas formas, sendo, a prevalência (quem está doente) e incidência (quem ficou doente). A proporção em ambos os casos, corresponde à fração dos indivíduos, podendo ser representado em percentual, em relação à frequência geral. Desta forma, pode-se eleger dois tipos de indicadores, sendo indicador de mensuração de prevalência e de incidência [30].

## 3.3 Indicadores de Vigilância em Saúde

Os indicadores de saúde podem ser medidas quantitativas e qualitativas, e são utilizados para expor e representar uma situação [31] [32].

Para aferir as medidas quantitativas são elaborados indicadores, como por exemplo: [33]

1. Indicadores epidemiológicos de: incidência, prevalência, mortalidade, letalidade, morbidade;
2. Indicadores de oportunidade;
3. Indicadores de representatividade;

4. Indicadores demográficos: natalidade, fecundidade, expectativa de vida.
5. Indicadores socioeconômicos: renda per capita e familiar, escolaridade, saneamento, renda, etc.

Eles podem ser compostos por: [33] [32].

1. **Coeficiente/Taxa:** é a representação do "risco" de um determinado evento ocorrer em uma população do país, estado, município, nascidos vivos, mulheres, etc. Resumidamente: é a relação entre o número de eventos reais e os que poderiam acontecer.
2. **Proporção:** é o resultado da relação entre a frequência atribuída de determinada ocorrência e o total geral. No numerador relaciona-se a frequência absoluta da ocorrência, que estabelece o subconjunto da ocorrência contida no denominador. Por exemplo: número de casos de dengue positivos em relação ao número total geral de dengue.
3. **Razão:** é a medida de frequência de um grupo de ocorrência relativa à frequência de outro grupo de ocorrência.

### Formulação Geral

**Coeficiente:** é uma razão entre a quantidade em que um evento foi registrado e o número total de vezes que o evento poderia ter sido registrado[33].

**Fórmula genérica:**

$$Fórmula\ genérica\ =\ \frac{Numerador}{Denominador}\ .\ 10^k \tag{3.1}$$

**Numerador:** número de vezes que um evento ocorreu durante um intervalo de tempo e para uma determinada área.

**Denominador:** população que correu o risco de obter a doença, na mesma área e intervalo de tempo, do evento mensionado no item Numerador.

**K:** é uma constante normal ($10^k$ normalmente representa o número de habitantes de uma determinada área, podendo ser: 10, 100, 1.000, 10.000, 100.000...).

**Conceituação:** quantitativo de eventos confirmados de uma doença, em uma população que reside em uma área específica, considerando o ano do evento.

### Definições Gerais

Para construir-se um indicador de saúde é importante conhecer as variáveis possíveis que o comporão [33] [32].

**Mortalidade:** variável caracteriza a medida de mortalidade em um dado intervalo de tempo.

**Coeficiente de mortalidade:** é a relação entre o número absoluto de óbitos e o quantitativos de pessoas que estão expostas ao risco de morrer.

**Letalidade:** é o grau de gravidade de uma doença em provocar a morte de indivíduos.

**Coeficiente de letalidade:** calcula-se a letalidade a partir do quantitativo de óbitos e número de pessoas que foram atingidas pela doença, representado, normalmente em percentual.

**Morbidade:** refere-se ao grupo de pessoas que contraíram doenças em um dado intervalo de tempo específico. Morbidade é o comportamento de uma doença em uma determinada população exposta.

**Coeficiente de morbidade:** calculado a partir do quantitativo de uma doença sobre o quantitativo de pessoas expostas ao risco de contrarir a doença.

$$Coeficiente\ Morbidade = \frac{Qtd.\ casos\ da\ doença}{População} \ .\ 10^n \tag{3.2}$$

$$Indicadores\ de\ investigação\ epidemiológica\ =\ \frac{Qtd\ de\ casos\ investigados}{Número\ de\ casos\ notificados} \ .\ 100 \tag{3.3}$$

**Prevalência:** quantitativo de caso da doença existente em uma determinada população e em um determinado tempo (momento).

**Coeficiente de prevalência:** tem por objetivo verificar a prevalência e resistência de uma doença na saúde pública. Calculado a partir da relação entre o número de casos identificados de uma doença e a população.

$$CMP\ =\ \frac{N^o\ de\ casos\ conhecidos\ de\ uma\ dada\ doença}{População} \ .\ 10^n \tag{3.4}$$

**Incidência:** na epidemiologia, infere-se a intensidade de ocorrências de uma doença, sendo o somatório do número de eventos ocorridos.

**Coeficiente de incidência:** é a razão entre o quantitativo dos novos casos de uma doença/agravo que incide sobre uma população, em determinado período de tempo. Ainda, considera-se a população sujeita a risco de contrair a doença/agravo durante o mesmo espaço de tempo, multiplicando pela potência de 10 o valor do resultado.

$$C.\ Incidência\ =\ \frac{n^o\ de\ nova\ doença,\ de\ um\ local,\ em\ período\ especifico}{n^o\ de\ pessoas\ expostas\ ao\ risco\ de\ uma\ doença\ em\ um\ período} \ .\ 10^n \tag{3.5}$$

## 3.4 Tecnologia em Business Intelligence (BI)

Desde os primórdios, a informação tem sido o maior aliado das organizações no processo de tomada de decisões. Atualmente, o volume de dados gerados e armazenados pelos sistemas de informação são enormes, com tendência a crescimento constante. Se por um lado gerar grandes volumes de dados é importante, por outro transformá-los em conhecimento é um desafio muito maior.

Neste sentido, inúmeras técnicas de *Business Intelligence (BI)* estão sendo construídas visando dar suporte aos processos de tomada de decisões. O uso dessa abordagem favorece o processo analítico dos dados e fortalece o modelo de tomada de decisões com base em dados históricos [34].

No entanto, *Business Intelligence* são mais que ferramentas ou *Software*. Desde a década de 90, o conceito de *Business Intelligence* tem sido lapidado e transformado. Para Hermanández-Julio et. al [34], a definição mais coerente é:

> ...são processos, tecnologias e ferramentas necessárias para transformar os dados em informações, informações em conhecimento e conhecimento em planos que impulsionam a ações empresariais... [34]

Magaireah at. al [35] destaca que BI são ferramentas tecnológicas e funcionais que inclui software, arquiteturas, bancos de dados, ferramentas de TI, processos analíticos e comerciais capazes de auxiliar na coleta, armazenar, distribuição e controle de diferentes fontes de dados, e por fim, transformando estes em informações e conhecimentos necessários para que as partes interessadas possam tomar decisões com base em conhecimentos extraídos de fontes seguras e confiáveis [35].

### 3.4.1 Extração, Transformação e Carga de Dados

As soluções de *BI* estão diretamente ligadas às estruturas e softwares computacionais, que por sua vez são capazes de Extrair, Transformar e Carregar (em inglês: *Extract, Transform, Load - ETL*). Os dados extraídos dos bancos de dados transacionais por meio das ferramentas de ETL são, comumente, armazenados em *Data Werehouse*, que tem a capacidade de estruturar e armazenar os dados de forma organizada, facilitando o acesso e a manipulação destes para a criação de painéis e *dashboard* a partir dos indicadores, permitindo suportar a decisões em todos os níveis das organizações [36].

Os componente principais do processo de BI são a Integração de Dados, Processamento Analítico Online (OLAP), *Extração, Tratamento e Carga (ETL)* e ferramentas de Visualização de Dados. Dentre todos estes componentes, os mais críticos e complexos são a ID e o processo ETL. A ID exige muito poder computacional para o processamento e o

armazenamento de dados, pois transforma e transfere dados entre esquemas diferentes (é a combinação de dados de diferentes fontes em uma fonte unificada), exigindo uma porcentagem alta de tempo. Por outro lado, o ETL tem sido considerado uma das melhores práticas para a extrair, transformar e carregar dados de diferentes fontes [37].

O ETL é, também, capaz de remover erros, corrigir dados ausentes, executar e trabalhar com funções sistêmicas, gerenciar, unificar e controlar a extração de dados de diferentes fontes simultaneamente, como por exemplo, SQLServer, Postgre, Oracle. A compreensão clara do formato e do propósito de ter dados selecionados é crucial para copiar apenas dados de interesse [37].

A transformação de dados faz parte da etapa de limpeza e confirmação dos dados recebidos. O objetivo deste processo é obter dados precisos que sejam corretos, completos, inequívocos e consistentes. O processo de carregamento é a fase em que os dados são carregados para a estrutura de *Data Warehouse*. Esse processo pode variar de acordo com os requisitos a que foi projetado, por exemplo, os dados podem serem extraídos e carregados a cada hora, diariamente, semanalmente ou mensalmente.

A Figura 3.1 mostra o processo de Integração de Dados, ETL, *Data Warehouse* e entrega das informações aos usuários. A Integração de Dados é executado em múltiplas fontes, seguida do processo de *Extração, Tratamento e Carga* dos dados no *Data Warehouse* e, por fim, estes dados são consumidos pelas ferramentas de Visualização de Dados (*dashbords* - painéis gráficos) [37].

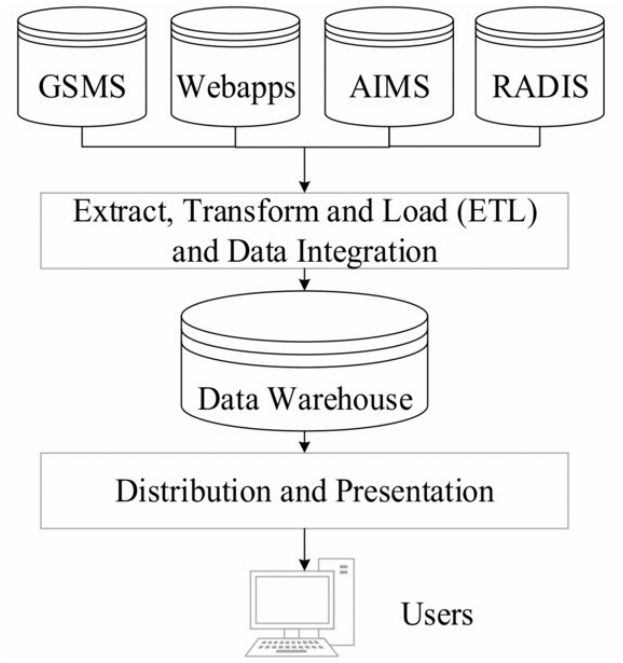



Figura 3.1: Processo de Extração, Tratamento e Carga do Dados (Fonte: [37]).

### 3.4.2 *Data Warehouse*

O uso de *Data Warehouse* tem aumentado consideravelmente nos últimos anos. Ele tem um papel importante no apoio ao processo de tomada de decisões, permitindo que as soluções de *Business Intelligence* acessem dados de forma mais rápida, flexível e melhor estruturada.

Entretanto, há várias formas de estrutura e carregar os dados em um *Data Warehouse*. Tudo dependerá da análise de negócios. Por exemplo, usando o conceito de empresa em tempo real, algumas áreas de negócios exigem processos de decisão cada vez mais rápidos e precisos, exigindo que o processo de carga dos dados seja com uma periodicidade muito curta, podendo acontecer mais de uma vez ao dia ou até mesmo que os dados transacionais e analíticos sejam sincronizados dinamicamente (tempo real). Isso torna o processo de *Extração, Tratamento e Carga* (sincronismos dos dados) um desafio e um ponto crítico do processo [38].

A escolha do tipo de sincronismos mais adequada para o *Data Warehouse* é importante, pois, permitirá o uso dos recursos tecnológicos e de manutenção de forma racional, com possibilidades de diminuir a complexidade das implementações. O alicerce desta escolha está no processo de análise dos domínios das aplicações transacionais e analíticas, bem como, as considerações do negócio [38].

#### Ambientes transacionais

Esses ambientes são projetados para sistemas transacionais, ou seja, transação por transação. Pode-se dizer que é um conjunto de funções que geram informações para que essas possam ser consumidas e armazenadas, mais tarde, por um *Data Warehouse*. Esses ambientes, normalmente, possuem subconjuntos de dados que devem ser considerados como uma unidade funcional dentro do ambiente transacional, onde cada uma dessas unidades pode ter comportamentos distintos relacionados ao seu sincronismo com o *Data Warehouse* [38].

Portanto, deve-se levar em considerar alguns fatores para a implementação do nível de sincronismo, sendo: [38].

1. A necessidade de manutenção:

   Podendo ser dividida em:

   (a) Estática - refere-se ao processo de carregamento tipicamente estático, sem necessidade de implementar atualizações dinâmicas;

(b) Dinâmica - possuem características de atualizações dinâmica a partir do ambiente transacional para o *Data Warehouse*, considerando o conjunto regras analisadas.

2. A capacidade de adaptação do ambiente transacional: avaliar a capacidade dos sistemas transacionas em relação a abordagem do carregamento dinâmico, se pode ser implementado ou não.
3. O intervalo de tempo desejável para o sincronismo: avalia-se as necessidades de sincronismos (atualização) dos ambientes, levando em consideração os recursos computacionais e o benefícios gerado a partir dos dados atualizados.

   O período de sincronismo pode ser:

   (a) Em tempo real: atualizações em tempo mínimo após a inclusão dos dados no ambiente transacional.
   (b) Intervalo determinado: os dados são carregados com período maior que os de tempo real, porém, com um tempo menor que os que são carregados de forma estática.
   (c) Sincronismo adiado: os dados são carregados regularmente, com por exemplo, diário, semanal, mensal.
4. Complexidade na propagação: deve-se levar em conta a complexidade do processo de propagação dos dados, pois quanto maior a complexidade, maior será o custo computacional e manuntenção.

### Ambientes analíticos

Os ambientes analíticos são caracterizados desta forma pela função a qual foram propostos, visando fornecer processos e estruturas que permitam aos usuários maiores capacidades de análise de dados.

As principais funções executadas neste tipo de ambiente (sobre o dado) estão relacionadas aos processos de agregação, soma, contagem, grupo e outras. Quanto maior o nível de agregação e abordagem, maior será a complexidade das consultas realizadas sobre os dados. Portanto, pode-se utilizar algumas estratégias para acelerar o tempo de resposta.

As técnicas mais utilizadas são: [38]

1. agregações dinâmicas calculadas durante as consultas.
2. desenvolvimento de visualizações com funções de agregação, que são executadas no momento das consultas.
3. armazenamento das visualizações em cache, com as funções de agregação calculadas.

4. implementação de tabelas de resumos, que visam facilitar e acelerar o processo de visualização dos dados.
5. armazenamento dos resultados das consultas no *Data Warehouse*, para uso futuro.
6. geração de dados agregados em cubos periodicamente.

### 3.4.3 Visualização de Dados

O *Big Data* tornou-se um tema importante em todas as áreas de conhecimentos, o que tem chamado a atenção das empresas para o armazenamento de dados que possam produzir riquezas e conhecimentos, hoje e/ou em um futuro próximo. Por outro lado, o principal desafio reside em capturar, armazenar, analisar e visualizar os dados. Quanto aos aspectos das análise dos dados, concentram-se em encontrar padrões interessantes em um enorme conjunto de dados.

Entretanto, na maioria das vezes o que se tem são números brutos, tornando a interpretação desses dados quase impossível pelo cérebro. Mas, a representação desses números em modo visual (graficamente) torna a compreensão mais fácil [39].

As ferramentas de visualização tradicionais alcançaram seus limites de capacidade quando se trata de conjuntos de dados muito grande, pois, eles exigem procedimentos e técnicas mais especializadas que forneçam velocidade e agilidade no processo de visualização dos dados [39].

Para Ali at. al [39] para reduzir o tempo de respostas de visualização dos dados, pode-se:

1. Utilizar dados pré-calculados.
2. Utilizar técnicas de paralelizar de processamento de dados e renderização.
3. Utilizar *middleware* preditivo

A maioria das ferramentas de visualizações de dados tem baixo desempenho em escalabilidade, funcionalidade e tempo de resposta. Uma boa ferramentas de visualização de dados deve ser capaz de lidar com dados estruturados, semi-estruturados e desestruturados de forma rápida e eficiente. As soluções programadas em linguagem ***R***, muitas vezes utilizadas em conjunto com o Hadoop[3], apresentam excelentes desempenho[39].

Várias ferramentas surgiram para auxiliar no processo de visualização de dados. Entretanto, o processo de escolha de uma ferramenta para apresentação dos dados deve se basear em melhores qualidades, agilidade, integração e interativo no processo de exibição dos dados. Mesmo com uma variedade de ferramentas de BI no mercado, as organizações

[3]Plataforma de software em Java de computação distribuída voltada para *clusters* e processamento de grandes volumes de dados, com atenção a tolerância a falhas

têm buscado maior flexibilidade para manipulação e apresentação dos dados, ou seja, utilizado o software BI que integra códigos, como por exemplo, linguagem R, Java, Python [40].

Exemplo de algumas ferramentas de visualização e suporte a *Business Intelligence*):

1. IBM Cognos Business Intelligence (IBM Watson Analytics)
2. Microsoft Power BI: interpreta códigos escritos em linguagem R
3. MicroStrategy BI e Data Analytics: Integra com software que interpreta códigos escrtos em linguagem R
4. Data Studio (Google)
5. Pentaho
6. Elasticsearch, Logstash e Kibana (ELK)
7. Django e Python
8. RHadoop [4]
9. SparkR [5]

### 3.4.4 Software de Visualização de Dados ELK

O *Elastic* é um conjunto de software formado por *Elasticsearch, Logstash e Kibana (ELK).* Juntos, estes proporcionam a coleta, processamento e visualização dos dados, além de contar com mais de 200 *plugins* que sofisticam as análises.

#### Elasticsearch

> **O Elasticsearch** é um software *open source*, que provê uma interface RESTful de pesquisa e análise de dados capaz de solucionar um número crescente de casos de uso. ...é um mecanismo de análise e pesquisa de texto completo de código aberto altamente escalável. Ele permite que armazene, pesquise e analise grandes volumes de dados rapidamente e em tempo quase real. É geralmente usado como o mecanismo/tecnologia subjacente que alimenta aplicativos que possuem recursos e requisitos de pesquisa complexos [41].

#### Logstash

O ***Logstash*** é um instrumento de coleta de dados de código aberto que permite a coleta em tempo real.

[4]RHadoop: O RHadoop é uma coleção de cinco (rhdfs, Rhase, plyrmr, rmr2 e Ravro) pacotes R que permitem aos usuários gerenciar e analisar dados com o Hadoop

[5]SparkR: é um pacote R que fornece uma interface leve para usar o Apache Spark no R. Ele fornece uma interface a dados distribuído e suporta operações como seleção, filtragem, agregação etc.

**O Logstash** é o motor central de fluxo de dados do *Elastic Stack* (Figura 3.2) para coletar, enriquecer e unificar todo tipo de dados, independentemente do formato ou esquema, de diversas fontes. O processamento em tempo real é especialmente eficiente quando associado ao Elasticsearch, ao Kibana e ao Beats [41].

Ele proporciona a unificação dos dados de forma dinâmica:

1. *de distintas fontes (processo de* ***Input***) com suporte a arquivos CSV, TCP/UDP, HTTP, API, *Json*, Banco de Dados etc;
2. *transformando-os com a aplicação de* ***Filter***, que podem ser realizadas por operadores condicionais, mutação (transformação) de dados de um campo de *string* para outro formato (por exemplo), acrescentar textos, cálculos e etc;
3. enviando-os (dados) a diferentes destinos, processo de ***Output***, como por exemplo: ***Análise:*** *Elasticsearch, MongoDB e Riak;* ***Monitoramento:*** *Nagios, Ganglia, Zabbix, Graphite, Datadog, CloudWatch;* ***Alerta:*** *Watcher with Elasticsearch, Email, Pagerduty, IRC, SNS.* Tudo isso é realizado de forma desnormalizada e limpa.
4. Ainda, conta com uma variedade de *plugins*, que impulsionam os processos de *input, filter e output* e muitos *codecs* nativos, que permite a simplificação dos processos de gestão e acelera os *insights* ao aproveitar um volume e uma variedade maiores de dados.

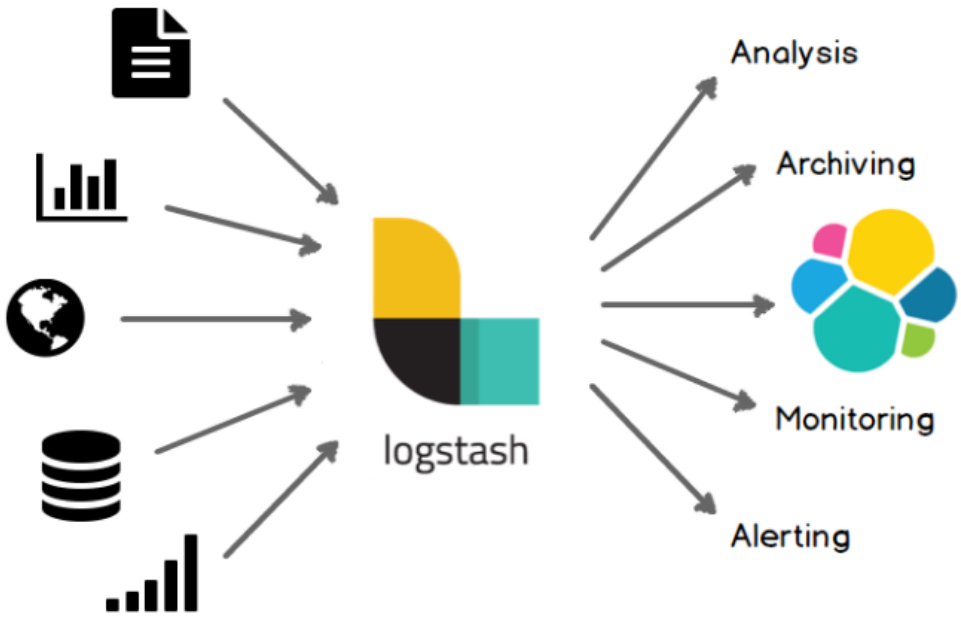



Figura 3.2: Processo do *Logstash* (Fonte: [42]).

O ***Logstash*** inicia, por padrão, o processo de coleta de dados a partir de um arquivo de configuração, que é programado com as instruções de *input:* concentra informações da origem e forma de coleta dos dados; *filter:* responsável pela filtragem dos dados; e, por último, o *output:* encaminha os dados ao destino indicado.

A estrutura básica do código da configuração arquivo de *coleta de dados (**Logstash**)* é formada por:

```
---------------------------
            input {
              ...
            }

            filter {
              ...
            }

            output {
              ...
            }
---------------------------
```

### Kibana

> **O Kibana** é uma janela dentro do Elastic Stack. Permite a exploração visual e analisar em tempo real dos dados no Elasticsearch. ...permite exploração de dados, visualização e criação de *dashboards* em questão de minutos. ...o núcleo do Kibana é fornecido com os clássicos: histogramas, gráficos de linha, gráficos de pizza, *sunbursts*... além disso, pode-se usar a linguagem ***Vega*** para criar visualizações personalizadas. Conta ainda com: séries temporais, visualização de dados geoespaciais, analise das relações com grafos, aprendizagem de máquina etc [41].

### Arquitetura do ELK

A arquitetura do *Elasticsearch, Logstash e Kibana* (Figura 3.3) é composta pelas interações entre os processos do ELK, onde o *Logstash* tem a função de coletar e processar os dados de acordo com os arquivos de configurações das fontes de dados, assim como enviá-los ao destino configurado no *Output* (neste caso o *Elasticsearch*).

O *Elasticsearch* recebe os dados, realiza a indexação e tem, como função principal, realizar buscas com alta performance.

O *Kibana* fornece uma interface para a criação de gráficos, dashboard e gerenciamento das informações indexadas pelo *Elasticsearch.*

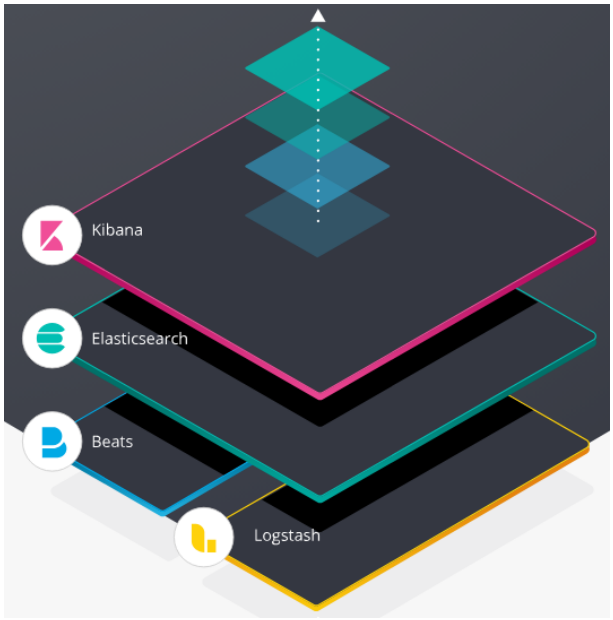



Figura 3.3: Arquitetura do ELK
(Fonte: [42]).

## 3.5 Séries Temporais

Segundo Antunes e Cardoso [43], Séries Temporais são formas de organizar informações quantitativas no tempo, ou seja, estas são deliberadas como um encadeamento de dados quantitativos referentes a períodos específicos e avaliados de acordo com sua disposição no tempo, podendo ser observações discretas ou contínuas, ter aplicações de formas diversificadas e em diversas áreas da ciência e do conhecimento [43].

Seja, então, $\{Z_t, t \in T\}$ uma Série Temporal em que $T$ representa os tempos $t$ em que a série foi observada [44] [19].

Para Latorre et. al [45] Séries Temporais são uma sequência lógica de dados arranjada em um período particular de tempo e o processo de análise destes, concentram-se em desenhar o objeto avaliado, delinear o comportamento da série, realizar aferições iniciais e investigar quais fatos podem influenciar no comportamento da série, procurando correlacionar causas e efeitos entre mais séries. As Séries Temporais podem ser: [45]

1. **Estacionária:** é considerada estacionária quando suas observações ocorrem aleatoriamente ao redor de uma média constante, apresentando certo equilíbrio, ou seja, não há tendência e nem sazonalidade.

2. **Não estacionária:** o processo é não estacionário quando possui tendência e/ou sazonalidade.

Dessa forma, as análises de Séries Temporais buscam encontrar os mecanismos que gerem Séries que preveem valores futuros, encontrem comportamentos e obtenham a periodicidade relevante dos dados. Por outro lado, as Séries são utilizadas para calibrar e mensurar os dados que compõem o Controle Estatístico de Processo [45][44].

Séries temporais $\{Z_t, t \in T\}$ são ditas fracamente estacionária quando possuem as seguintes características [19]:

$$E(Z_t) = \mu_t = \mu,\ constante\ , \forall t \in T; \tag{3.6}$$

$$E(Z_t^2) < \infty, \forall t \in T; \tag{3.7}$$

$$\gamma(t_1, t_2) = Cov(Z_{t_1}, Z_{t_2}), \text{é } uma\ fun\text{ção de } |t_1\ e\ t_2|. \tag{3.8}$$

Quanto às Séries não estacionárias, estas podem possuir tendências e/ou sazonalidades. Portanto, um processo $X(t)$ pode ser representado por [19]:

$$X_t = T_t + S_t + a_t, \tag{3.9}$$

onde $a_t$ é um Ruído Branco, ou seja, é parte aleatória que dispõe da distribuição $RB \sim (0, \sigma_a^2)$, $T_t$ é a parte da tendência e $S_t$ é a parte da sazonalidade.

Contudo, $T_t$ e $S_t$ possuem uma correlação forte e os procedimentos de estimação de $S_t$ podem ser impactados se não forem tratadas a tendência e a especificação de $S_t$ dependente da especificação de $T_t$.

Entretanto, a partir da transformação dos dados pode-se modificar a Série para estacionária. Esta mudança pode estabilizar a variância e o efeito sazonal torna-se complementar. A modificação mais utilizada é de diferenças sucessivas da série original da seguinte maneira:

$$\Delta X_t = X_t - X_{t-1}, \tag{3.10}$$

É a primeira diferença. O próximo cálculo é:

$$\Delta^2 X_t = \Delta[\Delta X_t] = \Delta[X_t - X_{t-1}], \tag{3.11}$$

Resultando na expressão:

$$\Delta^2 X_t = X_t - 2X_{t-1} + X_{t-2}, \tag{3.12}$$

E realizando "n" diferenças sucessivas, obtém-se a formúla geral:

$$\Delta^n = \Delta[\Delta^{n-1} X_t], \tag{3.13}$$

Entretanto, algumas vezes, antes de fazer a transformação das Séries, é necessário aplicar formas não lineares e geralmente usa-se a transformação de Box-Cox, sendo:

$$X_t^{\lambda} = \left\{ \begin{array}{l} \dfrac{X_t^{\lambda} - c}{\lambda}, se\lambda \neq 0 \\ \log X_t, se\lambda = 0' \end{array} \right\} \tag{3.14}$$

onde "c" e $\lambda$ são parâmetros a serem calculados. A mudança é considerada adequada no momento que o desvio-padrão da série for proporcional à média.

> ...a) quando durante um período os pontos oscilam ao redor de uma média e, depois, mudam de patamar (neste caso basta tomar uma diferença da série); e b) quando a série é não estacionária em relação à tendência (geralmente, para torná-las estacionárias é necessário tomar a segunda diferença)... [45]

A associação entre Séries Temporais pode refletir um processo mais complexo na identificação das variações nas séries. Este processo pode apresentar variações que sejam representativas para mérito de monitoramento da saúde, pois, podem proporcionar repetições estabelecidas no tempo que enfatizem determinados acontecimentos que não seriam observados se não fossem utilizadas Séries Temporais. Este fato é bastante usual na epidemiologia, podendo citar as variações sazonais, temperatura e cíclicas que dificultam a avaliação de diversas doenças e agravos à saúde humana [43].

Para Latorre et. al [45], durante o processo de análise dos dados, se não houver tendência ou sazonalidade pode-se utilizar os modelos Auto-Regressivos (AR) ou Auto Regressivos-médias Móveis (ARMA). Caso seja observado desempenho de tendência, pode-se valer-se do modelo Auto Regressivos Integrados de Médias Móveis (ARIMA) e em caso de sazonalidade, aplicando-se o modelo Modelo Auto-regressivo Integrados de Médias Móveis Sazonal (SARIMA).

### 3.5.1 Funções de Autocorrelação

A autocorrelação é determinada pela necessidade de um grau de correlação com seus próprios valores em momentos anteriores, ou seja, provê uma medida do grau de dependência entre os valores de uma série temporal em diferentes épocas. Portanto, ela é definida como a razão entre a autocovariância e a variância de uma sequência de dados, dada por:[44] [45]

$$P_{t,s} = (Z_t, Z_s) = \frac{Y_{t,s}}{\sqrt{Y_{t,t}} Y_{s,s}} \tag{3.15}$$

onde $Y_{t,s} = Cov(Z_t, Z_s), Y_{t,t} = \mathrm{Var}(Z_t)$ e $Y_{s,s} = \mathrm{Var}(Z_s)$.

Identificar o modelo do estudo é o objetivo primoridal da Função de Autocorrelação. Pois, quando é gerado o correlograma do modelo o número de *Lag*[6] será à ordem do tipo Média Móvel [46] [44] [45].

### 3.5.2 Função de Autocorrelação Parcial

A Função de Autocorrelação Parcial (FACP) é uma sequência de correlações entre ($z_t$ e $Z_{t-1}$), ($z_t$ e $Z_{t-2}$),..., desde que os efeitos das diferenças **"k"** anteriores mantenham-se fiéis. Tornando-a importante ao decidir a Ordem **"p"**[7] de um processo autorregressivo de um modelo ARIMA [44] [45].

Portanto, a FACP é dada por $\varphi_{kk}$ onde $\varphi_{kk}$ é valor do coeficiente das equações de Yule-Walker [44].

$$\varphi_{kk} = \frac{\begin{vmatrix} 1 & p_1 & \cdots & p_{k-2} \\ p_1 & 1 & \cdots & p_{k-3} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k-1} & p_{k-2} & \cdots & p_1 \end{vmatrix}}{\begin{vmatrix} 1 & p_1 & \cdots & p_{k-2} \\ p_1 & 1 & \cdots & p_{k-3} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k-1} & p_{k-2} & \cdots & p_1 \end{vmatrix}} = \frac{P_k^*}{P_k} \tag{3.16}$$

Em que,

$P_k$ representa a matriz de autocorrelações para um *lag* k;

$P_k^*$ equivale à matriz $P_k$ com a última coluna substituída pelo vetor de correlações.

### 3.5.3 Modelos Autoregressivos

Seja $\{X(t), t \in T\}$ uma série temporal, um processo autoregressivo de ordem $p$. O processo é descrito como $X_t \sim AR(p)$ e possui a seguinte representação: [44]

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + a_t \tag{3.17}$$

onde $\phi_0, \phi, \phi_2, ..., \phi_p$ são os parâmetros do modelo e $a_t$ é o ruído branco associado. Desta forma, pode-se definir o operador autoregressivo $\phi$(B) como:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 + ... + \phi_p B^p \tag{3.18}$$

[6] *Lag*: representa uma atraso de um intervalo de tempo na série.

[7] Ordem "p": é parte da especificação, que é chamada quase sempre de “identificação do modelo”

ou ainda

$$\phi(B)\bar{X}_t = a_t \tag{3.19}$$

Uma aspecto de um processo AR(p) é que a função de autorrelação é infinita em extensão que decai sob o formato de senóides amortecidas e/ ou exponecialmente. A função de autocorrelação parcial apresenta um corte no *Lag p* [44].

### 3.5.4 Modelos Média Móveis

Seja $\{X(t), t \in T\}$ um processo de média móveis de ordem $q$ é denotado por $X_t \sim MA(q)$ possui a seguinte lei de formação:

$$X_t = \mu + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ... - \theta_q a_{t-q} \tag{3.20}$$

onde $\mu$ e $\theta_1, \theta_2, ..., \theta_p$ são constantes. Escrevendo em função do operador de médias móveis de ordem $q$ obtém-se:

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_p B^p \tag{3.21}$$

ou também

$$\bar{X}_t = \theta(B)a_t \tag{3.22}$$

A Função de Auto-correlação Parcial (FACP) MA$q$ é finita, exibindo um corte *lag q*. A FACP trabalha de forma similar a Função de Autocorrelaçao de um processo AR($p$), baixando exponencial e/ou senóide amortecida.

### 3.5.5 Modelos Autoregressivos e de Médias Móveis

Esses modelos normalmente são utilizados para modelagem com poucos parâmetros. Sendo este, a junção dos modelos Autoregressivos e Médias Móveis, representado da seguinte forma:

$$X_t = \phi_0 + \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + a_t - \theta_1 a_{t-1} - ... - \theta_q a_{t-q} \tag{3.23}$$

Portanto, ele é retratado por ARMA(p, q) e contendo o operador:

$$\phi(B)\bar{X}_t = \theta(B)a_t \tag{3.24}$$

A Função Autocorrelação desse modelo é infinita, baixando exponencialmente e/ou em forma de senóide demostrando um corte no *Lag q e p*. A Função de Auto-correlação Parcial revela um comportamento como o de um processo MA *(*q).

### 3.5.6 Auto Regressivos Integrados de Médias Móveis

O modelo Auto Regressivos Integrados de Médias Móveis (ARIMA) é equivalente ao modelo ARMA*(p, q)*, ou seja, agrega o processo AR($p$) ao MA($p$). Todavia este modelo é apropriado para séries temporais não estacionárias, contendo a ordem **"d"** que relaciona-se à quantidade de diferenciações feitas até que a série seja estacionária. Portanto, a lei de formação do operador é representada por:

$$\phi(B)\Delta^d \bar{X}_t = \theta(B)a_t \tag{3.25}$$

é derivado por ARIMA *(p,d,q)*, onde $p$ é a ordem de $\phi(B)$ *e* $q$ *de* $\theta(B)$.

Segundo Sato [47], em 1970, George Box e Gwilym Jenkins desenvolveram o modelo ARIMA buscando apresentar modificações na Série Temporal, empregando abordagem matemática. O modelo foi fundamentado nos ajustes de valores observados. O objetivo foi diminuir a diferença para o mais próximo de zero entre o modelo e o valor observado.

As Séries Temporais permitem realizar previsões sobre o comportamento da série **"h"** passos a frente (previsões). Entretanto, três condições devem ser atendidas:[44] [47]

- Ter dados passados sobre a série.
- Habilidade de explicação das previsões apresentadas.
- Conjecturar que o padrão da série seja o mesmo.

Os modelos que se destacam no quesito de Séries Temporais são os de suavizações exponenciais (Simples, Holt e Holt-Winterns), os Box-Jenkins (ARIMA) e os de decomposição clássica. Dentre estes, os modelos ARIMA ajustam modelos autorregressivos agregados de média móveis a amostra dos moldes ARIMA (p,d,q), onde **"p"** é a ordem do modelo autorregressivo, **"d"** é o número de diferenciações e **"q"** é a ordem do modelo de média móveis [47][44] .

Apesar de haver multiplicidade do modelo ARIMA (p, d, q), de modo geral, sua forma é sem sazonalidade, sendo: [47]

1. **AR:** (p: grau da parte autoregressiva);
2. **I:** (d: grau da primeira diferença envolvida) e

3. **MA:** (q: grau da parte de média móvel).

Figura 3.4 mostra um diagrama esquemático do processo de estimação do modelo ARIMA

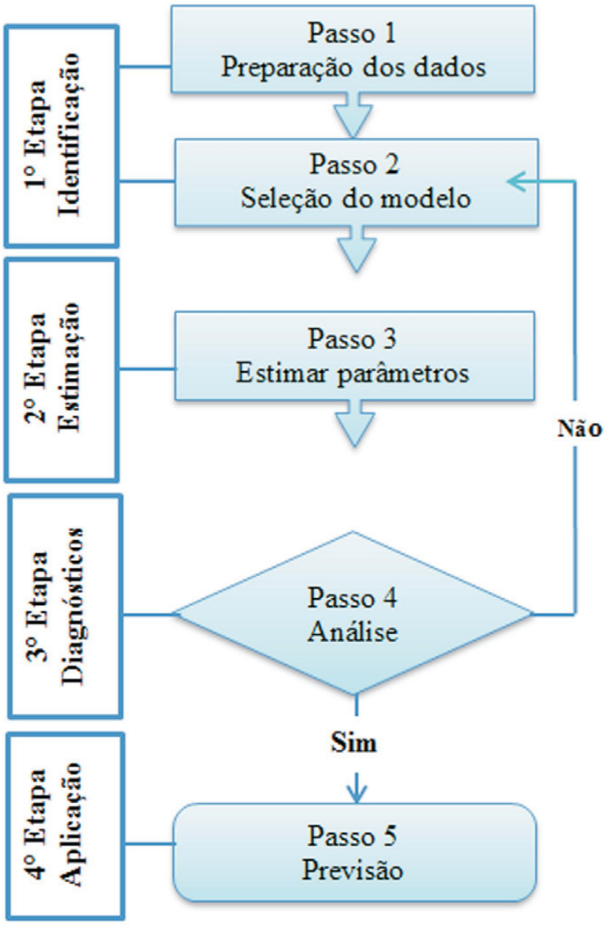



Figura 3.4: Esquema do processo de estimação do modelo ARIMA (Fonte: [47]).

Basicamente, o modelo divide-se em:

1ª Etapa: identificação - concentra-se na preparação dos dados (redução de variáveis e estacionariedade da série) e na seleção do modelo (avalia a autocorrelação e a autocorrelação parcial dos dados e destinge quais os padrões são aleatórios).

2ª Etapa: estimação - é o processo de parametrização do modelo identificado. As estimativas preliminares encontradas na fase de identificação são utilizadas como valores iniciais neste procedimento.

3ª Etapa: verificação do modelo (análise) - é o processo de ajustes do modelo através da análise de resíduos, para saber se este está adequado.

4ª Etapa: é a aplicação do modelo.

Buscando encontrar o modelo ARIMA mais conveniente, deve-se utilizar os seguintes passos: [47]

- Escolher o modelo geral para análise.
- Detectar o modelo mais intrínseco, sempre contemplando as autocorrelações, autocorrelações parciais e outras demais informações.
- Calcular os parâmetros do modelo.
- Identificar o modelo ajustado.

Entretanto, deve-se repetir este ciclo até que se tenha ajustado o modelo corretamente.

### 3.5.7 Modelos SARIMA

No modelo ARIMA, a autocorrelação dos valores de uma série é observada em momentos sucessivos. Entretanto, quando as observações estão distribuídas em períodos de um ano, a série pode apresentar autocorrelação a outra estação, denominada sazonalidade **"s"** [48]

Desta forma, os modelos que correspondem a esta modalidade (autocorreção Sazonal) são conhecidos como SARIMA.

A Sazonalidade é um fenômeno que ocorre periodicamente em período fixo em uma série de dados, sua avaliação é difícil e necessita de aplicações estatísticas [45].

O modelo SARIMA, normalmente, possui uma parte não sazonal, representada pelos parâmetros (p,d,q), e uma outra parte sazonal, representada pelos parâmetros (P,D,Q)s [48]

O modelo genérico é abordado pela equação: [48]

$$\begin{aligned}(1-\phi_1 L-...-\phi_p L^p)\ (1-\ \Phi_1 L^S\ -\ ...\ -\ \Phi_p L^{Ps})\\ (1-L)^d\ (1-L^s)^D Z_t\ =\ (1-\theta_1 L\ -\ ...\ -\theta_q L^q)\\ (1-\Theta_1 L^s\ -\ ...\ -\ \Theta_Q L^{Qs})\epsilon_t\end{aligned} \tag{3.26}$$

Portanto, temos:

A parte autorregressiva não sazonal de ordem **p**, corresponde a:

$$(1-\phi_1 L-...-\phi_p L^p)$$

A parte autorregressiva Sazonal de ordem **P** e parte Sazonal **s**, corresponde a:

$$(1 - \ \Phi_1 L^S \ - \ ... \ - \ \Phi_p L^{Ps}$$

A parte de integração não Sazonal de ordem **d**, corresponde a:

$$(1 - L)^d$$

A parte de integração Sazonal de ordem **s**, corresponde a:

$$(1 - L^s)^D$$

A parte não Sazonal de Médias Móveis de ordem **q**, corresponde a:

$$(1 - \theta_1 L \ - \ ... \ - \theta_q L^q)$$

A parte Sazonal de Médias Móvies de ordem **Q** e estação Sazonal **s**, corresponde a:

$$(1 - \Theta_1 L^s \ - \ ... \ - \ \Theta_Q L^{Qs})$$

### 3.5.8 Escolha do Modelo

Os modelos de previsão gerados são teóricos, entretanto estes são gerados por aproximação do que se espera. Assim, pode haver muitos modelos para o mesmo objeto monitorado. Por outro lado, com a eleição de um dos modelos, busca-se maior similaridade possível do cenário esperado. Entretanto, pode-se usar o Critério de Informação de Akaike (AIC), o Critério de Informação de Akaike Corrigido (AICc) e o Critério de Informação Bayesiano (BIC) para auxiliar neste processo. A escolha do modelo deve ser a partir do critério de informação que contenha os menores valores [49] [50].

### 3.5.9 Critério de Informação de Akaike

O Critério de Informação de Akaike (AIC) foi proposto em 1974, por Akaike, como instrumento para eleger modelos entre os modelos linear e não-linear. O AIC atribui ao modelo uma pontuação baseando-se na adequação aos dados e na ordem do modelo. Ele é representado por:

$$AIC = -2 \ Log \ L(\hat{\theta}) + 2(p) \tag{3.27}$$

onde $p$ é o número de parâmetros estimados no modelo.

Contudo, o AIC não é indicado para amostras pequenas, o que levou ao desenvolvimento do modelo Critério de Informação de Akaike corrigido (AICc) [49].

### 3.5.10 Critério de Informação de Akaike Corrigido

A partir dos estudos de Akaike, o Critério de Informação de Akaike foi readequado, criando-se Critério de Informação de Akaike corrigido (AICc), melhorando o desempenho para uso com pequenas amostras. Ele é representado pela fórmula:[51] [44]

$$AICc = -2\ Log\ L(\hat{\theta}) + 2(p) + \frac{p(p+1)}{n-p-1} \tag{3.28}$$

### 3.5.11 Critério de Informação Bayesiano

O Critério de Informação Bayesiano (BIC) foi criado, também, por Akaike em 1978, sendo mais uma opção ao AIC e AICc. Entretanto, ele foi desenvolvido sob a perspectiva Bayesiana, onde utiliza-se a probabilidade a *posteriori*. Diferentemente do AIC, este método parte do princípio que o modelo real tem dimensão infinita. Seu ajuste é dado por: [51] [44]

$$BIC = -2\ Log\ f(x_n|(\hat{\theta}) + p\ log\ n \tag{3.29}$$

## 3.6 Controle Estatístico de Processo

O Controle Estatístico de Processo (CEP) foi desenvolvido pelo físico *Walter Shewhart* (na década de 1920), com objetivo de melhorar o processo de fabricação industrial. Em seguida, foi aplicado em processos de ambientes laboratoriais e, posteriormente, houve aplicações diretas no atendimento ao paciente [52].

Thor et. al [52] destacam em seu artigo que:

> ...Controle Estatístico de Processo (CEP) é uma filosofia, uma estratégia e um conjunto de métodos para melhoria contínua de sistemas, processos e resultados. A abordagem CEP baseia-se na aprendizagem através dos dados e tem o seu fundamento na teoria da variação (entendendo causas comuns e especiais). A estratégia CEP incorpora os conceitos de um estudo analítico, processo de pensamento, prevenção, estratificação, estabilidade, capacidade e previsão. O CEP incorpora medidas, métodos de coleta de dados e experimentação planejada. Os métodos gráficos, como os gráficos *Shewhart* (mais comumente chamados de "Gráfico de Controle (GC)"), gráficos de execução, gráficos de frequência, histogramas, análise de pareto, diagramas de dispersão e diagramas de fluxo são as principais ferramentas usadas no CEP...[52]

Thor et. al [52] realizaram vasta análise a respeito da utilização do Controle Estatístico de Processo na área da saúde e concluíram que:

> ...Controle Estatístico de Processo é uma ferramenta versátil que pode ajudar diversas partes interessadas a gerenciar a mudança nos cuidados de saúde e melhorar a saúde dos pacientes...[52]

O Gráfico de Controle Shewhart foi proposto para ser utilizado no monitoramento de epidemias em 1946. No Brasil, não se tem registros do início de seu uso, mas acredita-se que a partir de 1976 seu uso foi expandindo-se [53].

O Gráfico de Controle é composto por um Limite Central do Controle (LC) que é a média do processo e os Limite Superior do Controle (LSC) e Limite Inferior do Controle (LIC), que são dados em relação a $\sigma$. Um processo é classificado dentro do controle se atender os atributos para este definido, caso contrário, este estará fora de controle. Esses atributos são delimitados por regras que os mantém dentro dos limites superior e inferior estabelecidos [54].

Por outro lado, os Gráfico de Controle são empregados para acompanhar se o processo está sob controle estatístico ou não, ou seja, se está dentro do esperado ou há variações inesperadas, incomum ou aleatórias. Quando o processo está sob controle, segue uma distribuição $(\mu, \sigma)$, geralmente com parâmetro desconhecido e que devem ser investigados [54].

Sendo $\mu$ a média e o $\sigma$ o desvio padrão, LC, LSC e LIC são dados por:

$$LSC = \mu + L\sigma \tag{3.30}$$

$$LIC = \mu - L\sigma \tag{3.31}$$

$$LC = \mu \tag{3.32}$$

Portanto, o processo é considerado sob controle se os pontos monitorados encontram-se dentro desses limites, caso contrário estará fora do controle.

Na área da Saúde, o processo de controle e acompanhamento das doenças e agravos pode ser comparado, de forma análoga, ao processo de produção industrial, ou seja, há uma variabilidade das unidades amostrais das doenças em um espaço-temporal, observando um limite máximo, intermediário e mínimo de controle [53].

O processo contínuo de vigilância no tempo visa monitorar determinados atributos das doenças e agravos à saúde, correlacionando suas ocorrências. Desta forma, pode-se aplicar ferramentas estatísticas (Gráfico de Controle) para controlar a incidência em um espaço de tempo. Enquanto estas taxas estiverem dentro de um padrão esperado, entende-se que a doença está sob controle [53].

Entretanto, é importante o monitoramento das alterações no processo amostral, que possam modificar-se para um processo fora do controle estatístico. Assim, diferentes métodos estatísticos podem ser utilizados na análise de Vigilância em Saúde Pública

(VSP), dentre eles os Gráfico de Controle de Somas Acumuladas (CUSUM), Shewhart, Média Móvel Exponencial Ponderada (MMEP) [53].

### 3.6.1 Gráficos de Shewhart

O Gráfico de Controle Shewhart é uma ferramenta muito útil no controle de processos onde há multivariantes. Ele é comumente implementado para monitorar o vetor médio do processo, ou seja, as características da qualidade do conjunto proposto para o gráfico de controle Shewhart multivariante. Supõe-se que a função de densidade de probabilidade das características de qualidade relacionada é a função de densidade para uma distribuição normal $p$ - *variável* e que existem amostras de subgrupos de tamanho $n$ disponível a partir do 'processo [55]

Existem duas diferentes fases no gerenciamento de controle e que cada uma delas possui especificações de limite de controle únicas. A primeira fase é utilizada para testar retrospectivamente se o processo estava sob controle, se as amostras e a estatística foram computadas. Essa fase visa estabelecer o limite de controle e o monitoramento futuro. Posteriormente, esses dados são utilizados para testagem do processo e saber se o processo está sob controle [55].

Segundo Aslam et. al [54], as pequenas alterações no processo não são possíveis de serem capturadas pelos gráficos de controle Shewhart.

Segundo Nidsunkid et. al [55] existem duas variantes do quadro de controle Shewhart multivariante: [55]

> "...uma para observações individuais ( N = 1 ), e outra para dados de amostra subgrupos (N>1)..." [55].

No gráfico de controle de Shewhart é indispensável que os valores dos componentes analisados (monitorados) sejam autônomos. Consequentemente, almeja-se que as variáveis não tenham correlação e que sejam estacionárias, visando a obtenção da máxima eficiência do controle [56].

Determinados métodos são utilizados para localizar padrões de não aleatoriedade, que podem sugerir ausência de controle, destacando-se: [56]

> ...**Pontos fora dos limites de controle:** esta é a indicação mais evidente de falta de controle de um processo, exigindo investigação imediata da causa de variação assinalável responsável pela sua ocorrência. Estes podem vir de resultados de erros de registro dos dados, de cálculos ou de medição ou, ainda, de algum instrumento descalibrado, de um erro do operador ou de defeitos nos equipamentos.
>
> Padrões cíclicos ou de periodicidade: acontecem quando os pontos, repetidamente, apresentam uma tendência para cima e para baixo, em intervalos de tempo que têm, aproximadamente, a mesma amplitude.

Sequência ou deslocamento de nível do processo: é uma configuração em que vários pontos consecutivos do gráfico de controle aparecem em apenas um dos lados da linha média. As sequências consideradas anormais são: sete ou mais pontos consecutivos; uma sequência com menos de sete pontos consecutivos, em que pelo menos dez de onze pontos consecutivos aparecem do mesmo lado da linha média; pelo menos doze de quatorze pontos consecutivos aparecem em um mesmo lado da linha média; e pelo menos dezesseis de vinte pontos consecutivos aparecem em um mesmo lado da linha média.

Tendência: é constituída por um movimento contínuo dos pontos do gráfico de controle em uma direção ascendente ou descendente.

Mistura ou aproximação dos limites de controle: é quando os pontos tendem a cair próximo ou levemente fora dos limites de controle, com relativamente poucos pontos próximos da linha média. Neste caso, podem existir duas distribuições sobrepostas, por exemplo, duas máquinas trabalhando de maneira diferente.

Estratificação ou aproximação da linha média: nesse caso, a maioria dos pontos está próximo da linha média, apresentando uma variabilidade menor do que a esperada. Pode ter ocorrido erro nos cálculos dos limites de controle ou que os subgrupos racionais (amostras) foram formados de maneira inadequada. Portanto, a aproximação da linha média não significa estar sob controle, mas, sim, a mistura de dados provenientes de populações distintas...[56]

Omar [57] demonstrou as regras para processo de produção em massa quando este é considerado fora de controle (Gráfico Shewhart). Para tal, utiliza-se a Figura 3.5 como referência:

1. Qualquer ponto que está fora do limite de controle de $3\sigma$ (ponto 3).
2. Pelo menos oito pontos consecutivos estão em um lado do gráfico (pontos 4 - 13).
3. Dois ou três pontos consecutivos estão fora do limite de advertência de $2\sigma$, mas dentro do limite de controle (pontos 20 e 21).
4. Quatro ou cinco pontos consecutivos estão além dos limites de $1\sigma$ (pontos 14 - 17).
5. Seis pontos em uma sequência sempre crescente ou decrescente (pontos 14-20).
6. Quinze pontos em sequência (tanto acima quanto abaixo da linha central).
7. Quatorze pontos em sequência alternadamente para cima e para baixo.
8. Oito pontos em sequência de ambos os lados da linha central.
9. Um padrão incomum ou não aleatório ocorre nos dados, como um padrão cíclico (pontos 14 a 25 que formam um padrão quadrático).
10. Um ou mais pontos estão perto de um limite de aviso ou controle (pontos 18 e 19).

Entretanto, apesar de haver regras claras, o desafio maior é entender e interpretar se o processo está fora de controle.

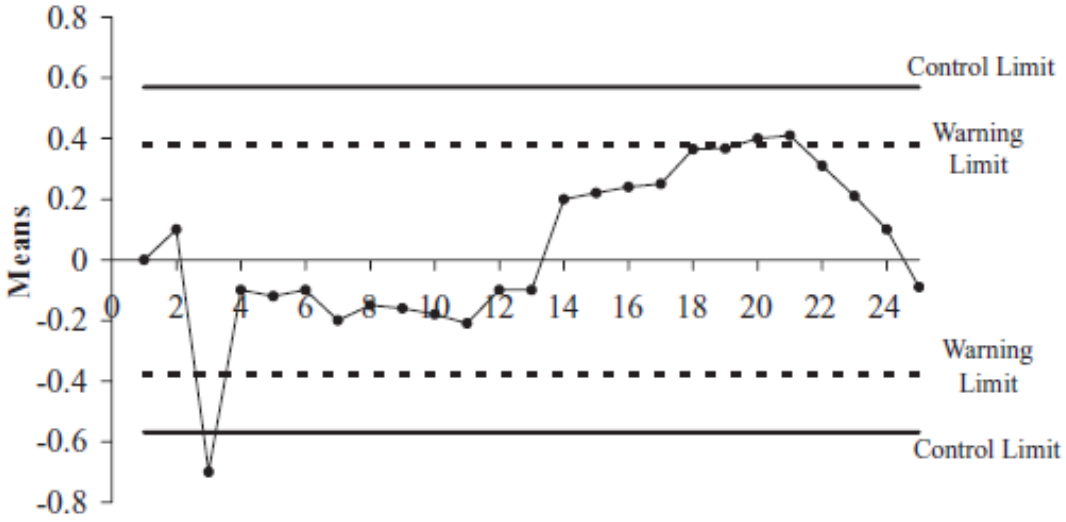



Figura 3.5: Controle Estatístico do Processo
(Fonte: [57]).

### 3.6.2 Gráficos de Controle de Soma Acumulativa

O Gráfico de Controle de Somas Acumuladas, embora seja mais sensível que o gráfico Shewhart para pequenas modificações na média, tem sido menos utilizado e não é tão conhecido na indústria brasileira. Este fato, ao que tudo indica, ocorre em virtude do grau de dificuldade de implementação e entendimento dos resultados apresentados por ele [58].

Os Gráficos de Controle de Soma Acumulativa (CUSUM) são técnicas de monitoramento sequenciais. Assim, a decisão sobre o que o gráfico deve ou não indicar irá depender da soma das observações de múltiplos períodos de tempo mais atuais. Em vez de um único período de tempo, como acontece no Gráfico de Shewhart [59].

Para o monitoramento de emergência (utilizando o Gráfico de CUSUM) do número de doenças semelhantes, como por exemplo, a influenza, a contagem das observações pode ser definida com menos de sete dias e os limites configurados heuristicamente com base nos ajustes visualizados nos dados históricos [59].

Desta forma, entende-se que o enfoque é eficaz para detectar alterações de curto prazo no processo de monitoramento. Entretanto, no Gráfico de Controle de Soma Acumulativa (CUSUM) a interpretação não é tão intuitiva [59].

Existem dois tipos de Gráfico de Controle de Soma Acumulada, o Tabular e Máscara "V". O primeiro é mais utilizado. Portanto, a CUSUM ($Ci$) depois de $n$ amostras é dada por: [60].

$$C_i = \sum_{j=1}^{i} (\bar{X}_j - \mu_0 \tag{3.33}$$

onde o $\mu_0$ é a média do processo e $\bar{X}_j$ a média da j-ésima amostra. $C_i$ acumula a informação dos valores anteriores.

Entretanto, é preciso descobrir se o processo está sob controle, assim, usa-se:

$$C_i^+ = máx[0, \bar{X}_t - (\mu_0 + k) + C_{i-1}^+] \tag{3.34}$$

$$C_i^- = máx[0, (\mu_0 - k) - \bar{X}_t + C_{i-1}^-] \tag{3.35}$$

onde $C_i^+$ e $C_i^-$ são estatísticas de CUSUM superior e inferior.

K é o valor de referencia e é a metade do valor encontrado da média, sendo representado por:

$$K = \frac{|\mu_1 - \mu_o|}{2} \tag{3.36}$$

Em seguida, deve-se criar o parâmetro que atua como regra decisiva para saber se o processo está ou não sob controle. o $H$ representa o intervalo de decisão [60].

$$LSC : H = +h\sigma \tag{3.37}$$

$$LIC : H = -h\sigma \tag{3.38}$$

O processo é considerado fora de controle se exceder o intervalo de $H$

Por outro lado, pode-se realizar a padronização do Gráfico de Controle de Soma Acumulada onde a nova variável possui a seguinte definição: [60].

$$Z_i = \frac{\bar{Z}_1 - \mu_o}{\sigma} \tag{3.39}$$

Com $Z_i$ sendo uma Normal.

Os limites de controle para o Gráfico de Controle de Soma Acumulada, são:

$$C_i^+ = máx[0, Z_i - k + C_{i-1}^+] \tag{3.40}$$

$$C_i^- = máx[0, -k - Z_i + C_{i-1}^-] \tag{3.41}$$

### 3.6.3 Médias Móveis Exponencialmente Ponderadas

O Gráfico de Controle de Média Móvel Exponencial Ponderada (MMEP) é um instrumento extremamente importante para o monitoramento do controle dos processos, sendo utilizado para a detecção e deslocamento constantes dos processos, tendo como proveito o aparecimento veloz de pequenos e moderados deslocamentos [61].

O Gráfico de Controle de Média Móvel Exponencial Ponderada (MMEP) (sigla em inglês EWMA) foi proposto para superar a fragilidade do gráfico de controle Shewhart em detectar pequenas alterações no processo pois, com sua aplicação, é possível utilizar informações atuais, e também, informações anteriores para definir se o processo está sob controle. Ainda, muitos autores criaram tabelas de controle usando a estatística EWMA [61].

Entretanto, ele tem melhores resultados para dados que rejeitam a hipótese de normalidade, ou seja, quando não segue uma distribuição normal, o MMEP é o tipo de gráfico mais indicado [60].

Primeiro é preciso transformar os dados, sendo:

$$Z_i = \lambda x_i + (1 - \lambda) Z_{i-1} \tag{3.42}$$

Onde $\lambda$ possui valor entre 0 e 1.

Usando o desvio padrão, referente às observações, pode-se calcular a variânca de $Z_i$, sendo: [60].

$$\sigma_{zi}^2 = \sigma^2 \left( \frac{\lambda}{2 - \lambda} \right) [1 - (1 - \lambda)^{2i}] \tag{3.43}$$

Sendo a próxima etapa a análise dos limites de controle:

$$LSC = \mu_0 + L\sigma \sqrt{\frac{\lambda}{2 - \lambda} [1 - (1 - \lambda^{2i})]} \tag{3.44}$$

$$LC = \mu_0 \tag{3.45}$$

$$LIC = \mu_0 - L\sigma \sqrt{\frac{\lambda}{2 - \lambda} [1 - (1 - \lambda^{2i})]} \tag{3.46}$$

Sendo 0,05 $\leq \lambda$ 0,25 e L=3 são as formas mais comum utilizadas, pois estas detectam pequenas perturbaçoes no processo.

# Capítulo 4

# Solução Proposta

No decorrer da análise da solução foi realizado de uma exploração aprofundada e detalhada do estudo de caso, com a proposição de uma arquitetura de software que apoie no processo de identificações de eventos anormais, com possíveis indicações de surtos ou emergências de saúde.

## 4.1 Entendimento de Negócio

O Entendimento de Negócio estabelece uma visão geral sobre as funções, processos e estrutura da Vigilância em Saúde, focando no objeto estudado e subsidiando o entendimento das atividades e estrutura da vigilância em saúde.

### 4.1.1 Sistema Único de Saúde

O contexto da saúde pública no Brasil é formado pelo compartilhamento de responsabilidades entre os governos municipais, estaduais e federal. Suas bases são reguladas pela Constituição Federal[1] de 1988 (CF/88), Lei nº 8.080[2] de 19 de setembro de 1990, normas e portarias.

A Constituição Federal de 1988 (CF/88) define que:

> ...a saúde é direito de todos e dever do Estado, garantido mediante políticas sociais e econômicas que visem à redução do risco de doença e de outros agravos e ao acesso universal e igualitário às ações e serviços para sua promoção, proteção e recuperação.

[1] CF/88: Constituição da República Federativa do Brasil de 1988 é a lei fundamental e suprema do Brasil.

[2] Lei nº 8.080: Dispõe sobre as condições para a promoção, proteção e recuperação da saúde, a organização e o funcionamento dos serviços correspondentes e dá outras providências.

> ....as ações e serviços públicos de saúde integram uma rede regionalizada e hierarquizada e constituem um sistema único, organizado de acordo com as seguintes diretrizes:
>
> I - descentralização, com direção única em cada esfera de governo;
>
> II - atendimento integral, com prioridade para as atividades preventivas, sem prejuízo dos serviços assistenciais...

A Lei nº 8.080/90 define que:

> ...o conjunto de ações e serviços de saúde, prestados por órgãos e instituições públicas federais, estaduais e municipais, da administração direta e indireta e das fundações mantidas pelo poder público, constitui o Sistema Único de Saúde (SUS).

Dentro deste conjunto de ações e serviços de saúde, encontra-se a vigilância em saúde que é estruturada por serviços e, também, descentralizada entre as três esferas: municipal, estadual e federal.

### 4.1.2 Ministério da Saúde

Apesar dos registros históricos mostrarem que a Saúde Pública Brasileira iniciou-se em 1808, o Ministério da Saúde (MS) só foi instituído em 1953, com a Lei nº 1.920/53. Ao longo destes mais de sessenta e cinco anos de existência, o Ministério passou por diversas reformas estruturais, o que corroborou com a definição de sua missão.

> Promover a saúde da população mediante a integração e a construção de parcerias com os órgãos federais, as unidades da federação, os municípios, a iniciativa privada e a sociedade, contribuindo para a melhoria da qualidade de vida e para o exercício da cidadania [62].

Ministério da Saúde (MS) é

> ...órgão que compõe o poder executivo federal, responsável pela organização e elaboração de planos e políticas públicas voltados para a promoção, prevenção e assistência à saúde dos brasileiros.
>
> Sua função é dispor de condições para a proteção e recuperação da saúde da população, reduzindo as enfermidades, controlando as doenças endêmicas e parasitárias e melhorando a vigilância à saúde, dando, assim, mais qualidade de vida ao brasileiro... [62]

Seu contexto interno é formado por uma estrutura verticalizada, hierárquica, e, também, política. Tendo como hierarquia o Gabinete do Ministro, Secretarias, Departamentos, Coordenações Gerais e, por fim, as Coordenações Auxiliares [62].

Em relação aos recursos humanos, observa-se uma diversidade de especialistas, o que torna seu quadro profissional qualificado e com *experts* nos processos políticos, administrativos e técnicos. Os recursos materiais e tecnológicos, são usados, primordialmente para atendimento de objetivos políticos, estratégicos e organizacional, visando o melhor atendimento à saúde da população [62].

### 4.1.3 Secretaria de Vigilância em Saúde

A Secretaria de Vigilância em Saúde do Ministério da Saúde (SVS/MS) está dividida em cinco departamentos de vigilância (Doenças Transmissíveis; Doenças e Agravos não Transmissíveis e Promoção a Saúde; Gestão da Vigilância em Saúde; Vigilância, Prevenção e Controle das Infecções Sexualmente Transmissíveis; Saúde Ambiental e do Trabalhador) e ainda conta com o Instituto Evandro Chagas, que é composto por vários laboratórios especializados em diagnósticos de doenças e agravos a saúde [62] .

A Portaria 1.378 de 2013[3], do Ministério da Saúde, traz em seus Artigos 2º e 4º algumas definições e atributos importantes para a vigilância em saúde.

> ...Art. 2º A Vigilância em Saúde constitui um processo contínuo e sistemático de coleta, consolidação, análise e disseminação de dados sobre eventos relacionados à saúde, visando o planejamento e a implementação de medidas de saúde pública para a proteção da saúde da população, a prevenção e controle de riscos, agravos e doenças, bem como para a promoção da saúde.
>
> Art. 4º As ações de Vigilância em Saúde abrangem toda a população brasileira e envolvem práticas e processos de trabalho voltados para:
>
> 1. a vigilância da situação de saúde da população, com a produção de análises que subsidiem o planejamento, estabelecimento de prioridades e estratégias, monitoramento e avaliação das ações de saúde pública;
> 2. a detecção oportuna e adoção de medidas adequadas para a resposta às emergências de saúde pública;
> 3. a vigilância, prevenção e controle das doenças transmissíveis;
> 4. a vigilância das doenças crônicas não transmissíveis, dos acidentes e violências;
> 5. a vigilância de populações expostas a riscos ambientais em saúde;
> 6. a vigilância da saúde do trabalhador;
> 7. vigilância sanitária dos riscos decorrentes da produção e do uso de produtos, serviços e tecnologias de interesse a saúde; e
> 8. outras ações de vigilância que, de maneira rotineira e sistemática, podem ser desenvolvidas em serviços de saúde públicos e privados nos vários níveis de atenção, laboratórios, ambientes de estudo e trabalho e na própria comunidade...

Nos Estados e Municípios, a organização dos serviços de vigilância possui variações, entretanto, todos eles possuem Secretaria de Saúde. Algumas são divididas em departamentos, superintendências e são estruturadas por diversos subsistemas, como hospitalar, laboratorial, farmácia, programas de saúde e de vigilância e etc. A partir dessa estrutura é realizada a vigilância em saúde.

[3]Portaria nº 1.378/2013: Regulamenta as responsabilidades e define diretrizes para execução e financiamento das ações de Vigilância em Saúde pela União, Estados, Distrito Federal e Municípios, relativos ao Sistema Nacional de Vigilância em Saúde e Sistema Nacional de Vigilância Sanitária.

### 4.1.4 Coordenação Geral de Laboratórios - CGLAB

A Coordenação Geral de Laboratórios de Saúde Pública (CGLAB) está ligada ao Departamento de Gestão da Vigilância em Saúde (DEGEVS), que por sua vez está ligado à Secretaria de Vigilância do Ministério da Saúde (SVS/MS). Entretanto, suas características e atributos permitem que a CGLAB forneça apoio, informação e fomento aos Laboratórios de Saúde Pública e as Vigilâncias em Saúde das três esferas de governo. [62].

A Portaria nº 1.419/MS[4], de 8 de Junho de 2017, que aprovou o regimento interno do Ministério da Saúde, traz em seu contexto duas áreas de atuação da CGLAB, que é Coordenação de Normatização de Laboratórios de Saúde Pública e Coordenação de Vigilância Laboratorial.

> ...Art. 17 - À Coordenação de Normatização de Laboratórios de Saúde Pública compete:
>
> 1. acompanhar a implementação e/ou implantação de normas técnicas e operacionais para a Rede Nacional de Laboratórios de Vigilância Epidemiológica e em Saúde Ambiental;
> 2. monitorar e avaliar o cumprimento das normas referentes aos sistemas de informação laboratorial em Vigilância Epidemiológica e Ambiental em Saúde;
> 3. monitorar e avaliar a conformidade das especificações dos equipamentos e produtos para saúde em atendimento ao diagnóstico laboratorial no âmbito da Rede Nacional de Laboratórios de Vigilância Epidemiológica e em Saúde Ambiental; e
> 4. habilitar, conforme critérios pré-estabelecidos, os Laboratórios de Referência Nacional e Regional para a Rede Nacional de Laboratórios de Vigilância em Saúde...
>
> ...Art. 18 - À Coordenação de Vigilância Laboratorial compete:
>
> 1. monitorar, avaliar e manter atualizados os sistemas de informação laboratorial em vigilância epidemiológica, vigilância em saúde ambiental;
> 2. monitorar o comportamento epidemiológico de doenças e agravos objeto de controle no campo laboratorial; e
> 3. colaborar tecnicamente e acompanhar a implantação do Sistema de Gestão da Qualidade e Biossegurança nas redes de laboratórios de Vigilância em Saúde, junto às demais unidades competentes...

### 4.1.5 Laboratórios de Saúde Pública

A Portaria 2.031/MS[5] de 23 de setembro de 2004, que dispõe sobre a organização do Sistema Nacional de Laboratórios de Saúde Pública, define-o como:

[4]Portaria nº 1.419/MS: aprova os regimentos internos e o quadro demonstrativo de cargos em comissão e das funções de confiança das unidades integrantes da estrutura regimental do Ministério da Saúde.

[5]Portaria 2.031/MS: dispõe sobre a organização do Sistema Nacional de Laboratórios de Saúde Pública.

> ...Sistema Nacional de Laboratórios de Saúde Pública – SISLAB é um conjunto de redes nacionais de laboratórios, organizadas em sub-redes, por agravos ou programas, de forma hierarquizada por grau de complexidade das atividades relacionadas à vigilância em saúde - compreendendo a vigilância epidemiológica e vigilância em saúde ambiental, vigilância sanitária e assistência médica.

As principais divisões da Rede SISLAB são:[63]

1. *Nível Local:* laboratórios públicos, em âmbito municipal, que integram a rede local de serviços para realização de exames básicos e essenciais.
2. *Nível Regional:* laboratórios públicos que realizam exames de complexidade intermediária e que visam atender as demandas dos níveis locais.
3. *Nível Estadual:* Laboratórios Centrais dos Estados (LACENs), que são os responsáveis pela gestão e desenvolvimento das redes de laboratórios nos Estados. Ainda, realizam exames de média e alta complexidade e que não são possíveis de serem realizados pelos Laboratórios de níveis locais e regionais.
4. *Nível Nacional:* laboratórios de alto nível de qualidade e excelência, habilitados para a realização de exames de média e alta complexidade que requerem infraestrutura adequada para diagnosticar doenças mais graves e contagiosas. Além de serem responsáveis pela pesquisa e produção de novas tecnologias de diagnósticos.

### 4.1.6 Sistema Gerenciador de Ambiente Laboratorial

A Coordenação Geral de Laboratórios de Saúde Pública (CGLAB) desenvolveu, em 2008, o Sistema Gerenciador de Ambiente Laboratorial (GAL) com o objetivo de informatizar, armazenar e estrutura todas as informações laboratoriais das amostras de origem humana, animal e ambiental produzidas pela Rede Nacional de Laboratórios de Saúde Pública (SISLAB); disponibilizar dados para as vigilâncias epidemiológicas e ambientais nos âmbitos municipal, estadual e federal; e, também, monitorar o comportamento epidemiológico de doenças e agravos com base nos exames. [14].

O GAL foi construído em linguagem de programação *Hypertext Preprocessor (PHP)*, com banco de dados PostGree e Oracle. Sua arquitetura é distribuída, robusta, flexível e todas as aplicações trocam informações entre si, por meio de WebService [14].

O software GAL concentra exames referentes as doenças e agravos de média e alta complexidade, como, por exemplo, sarampo, malária, febre amarela, poliomielite, influenza, tuberculose etc. Seu contexto destaca-se por:

1. possuir 30 instalações distribuídas nas 27 unidades federadas;

2. ser utilizado por mais de 4.644 municípios e 8.211 Unidades Básicas de Saúde (UBS);
3. possuir quase 30 mil usuários ativos;
4. possuir aproximadamente 9 milhões de pacientes registrados com algum tipo de doença/agravo à saúde;
5. liberar mais de 22 milhões de diagnósticos;
6. contar com mais de 700 tipos de exames e métodos cadastrados;
7. possuir mais de 1.200 laboratórios habilitados; e
8. possuir mais de 7100 centros de coletas ou cadastros de exames.

**Fluxos de informações**

O GAL encontra-se instalada em 27 Unidades Federativas (UF), duas nos Laboratórios de Referências Nacionais e uma no Ministério da Saúde. Dessa forma, todas as informações dos laboratórios são distribuídas nestas aplicações de acordo com suas UF. Por sua vez, a aplicação GAL Nacional (Ministério da Saúde) armazena todas as informações que são produzidas pelo país [14].

A partir de um pedido de exames (que pode conter vários exames de média e alta complexidade), o sistema gerencia e controla todos os fluxos de dados do pedido, podendo distribuí-los entre vários laboratórios (aplicações de UFs diferentes), quer seja este municipal, estadual e/ou de referência nacional [14].

Após a conclusão dos processos e liberações dos resultados dos exames laboratoriais, todas as informações da aplicação estadual são enviadas para uma base única do Ministério da Saúde (GAL Nacional). Entretanto, a vigilância em saúde pode acompanhar os resultados e os pedidos de exames a partir do momento em que este foi inserido pela Unidade Básicas de Saúde (UBS) [14].

Todas as interfaces de comunicações utilizadas são fornecidas por WebService, com base nos padrões definidas na Portaria nº 2.073/MS[6], de 31 de agosto de 2011 [14].

### 4.1.7 Monitoramento do comportamento epidemiológico de doenças e agravos no campo laboratorial

A CGLAB tem, dentre as suas atribuições, a realização do monitoramento do comportamento epidemiológico de doenças e agravos no campo laboratorial, e mesmo possuindo

[6]Portaria nº 2.073/MS: regulamenta o uso de padrões de interoperabilidade e informação em saúde para sistemas de informação em saúde no âmbito do Sistema Único de Saúde, nos níveis Municipal, Distrital, Estadual e Federal, e para os sistemas privados e do setor de saúde suplementar.

todas as informações laboratoriais concentradas no Ministério da Saúde, tem encontrado dificuldade no cumprimento desta atribuição.

Portanto, o atendimento dos objetivos desse trabalho auxiliará a Coordenação no processo de implementação do monitoramento do comportamento das doenças a partir de painéis, *dashboards* e de relatórios.

## 4.2 Entendendo os Dados

O entendimento dos dados é uma etapa importante para moldar as possíveis formas de monitoramento do objeto desejado, os formatos dos atributos, os dados faltantes, sua consistência e etc. Baseado nesta ideia, foram realizados testes nos dados que foram fornecidos pelo Ministério da Saúde, por meio de solicitação direta.

Os dados foram submetidos:

1. Análise superficial para identificações dos tipos de variáveis (atributos)
2. Teste de dados faltantes: verificação dos dados faltantes nos campos (total dos dados preenchidos subtraído pelo total geral de tuplas existentes)
3. Descrição inicial dos dados
4. Verificação da qualidade dos dados

Após os testes iniciais, foi verificado que os dados considerados essenciais para este trabalho possuem excelente preenchimento, pois aproximadamente 100% destes são de preenchimento obrigatório. Quanto à qualidade dos dados, observou-se que ela é boa, pois os atributos são de preenchimento obrigatório e 70% dos itens são pré-definidos (definidos por padrões) pelo sistema.

## 4.3 Pré-processamento

A etapa do pré-processamento é considerada a base para as demais etapas, pois, se essa não for bem definida poderá impactar negativamente nos processos e, ainda, não obter o conhecimento esperado.

Ações realizadas na fase de pré-processamento:

1. foi utilizada a ficha de requisição de exames para subsidiar a análise de todas as entradas de dados do sistema Sistema Gerenciador de Ambiente Laboratorial (GAL);

2. verificado o quantitativo de produção de exames (relatório de produção de exames do GAL);
3. avaliada a representatividade de mais 380.000 (trezentos e oitenta mil) amostras;
4. renomeados os atributos dos dados;
5. criada tabela de controle de descrição das variáveis, apelido e atributos;
6. efetuada operações preliminares de limpeza das informações.

   **Obs.:** A etapa de pré-processamento foi detalhada na Seção 4.4.6.

### 4.3.1 Entendendo as variáveis do Banco de Dados

O banco de dados disponibilizado para a pesquisa possui mais de 100 variáveis, com um conjunto de atributos que envolvem localização, espaço geográfico, qualificações dos tipos e resultados de exames, classificações de doenças, espaço amostral de tempo, características dos indivíduos, centros de saúde, materiais, kits e amostras biológicas.

De modo particular e por ser tratar de uma análise de pré-processamento, destacaram-se como variáveis essenciais os atributos: UF de residência; Município de residência; código do IBGE; data de nascimento; tipo de idade; sexo; ano; data da coleta; data de liberação; e resultado dos exames. Foi necessário, ainda, realizar a transformação de atributos em grupos: semana epidemiológica, ano, mês, região etc. Foram considerados atributos regulares as doença e agravos do GAL; finalidade; CID; nacionalidade; raça/cor; etnia; exame; material biológico e método. Os demais atributos foram considerados pouco relevantes para a análise em questão.

Os critérios utilizados na seleção das variáveis foram os de: localidade do paciente, data que mais se aproxima da doença acometida pelo paciente, exames e métodos que apresente o melhor resultado para diagnóstico.

Entretanto, considerando os objetivos de negócios e da pesquisa e, também, a análise preliminar dos dados, pode-se afirmar que pelo menos 6 variáveis são essenciais para o estudo e outras serão geradas a partir dos dados existentes.

Variáveis elencadas (importantes) para compor o estudo:

1. Chave (hash informações do paciente);
2. Região geográfica;
3. UF Residência do Paciente;
4. Município de Residência do Paciente;
5. Exames e Métodos;

6. Resultados dos Exames;
7. Data dos Sintomas;
8. Data da Coleta;
9. Data de Liberação do Exame;
10. Semana Epidemiológica;
11. Dia (da coleta da amostra);
12. Mês (da coleta da amostra);
13. Ano (da coleta da amostra).

### 4.3.2 Análise dos Dados

A análise dos dados visa estabelecer uma visão geral das informações em que se pretende trabalhar. Neste trabalho foram utilizados dados históricos dos exames de Biologia Médica, que foram produzidos pela Rede SISLAB, trabalhando-os de forma mais agrupada possível e com representações gráficas que pudesse fortalecer a pesquisa, a aplicação de Série Temporal e Gráfico de Controle. Desta maneira, pode-se identificar quais eram os principais bolsões de dados, levando em consideração a importância e evolução de cada agravo/doença.

Na Figura 4.1 é possível verificar quais foram os Estados que mais realizam exames (em destaque: CE, RN e PR), a Figura 4.2 apresenta a evolução dos dados ao longo dos anos, fato este que se deve ao processo de adesão ao sistema GAL.

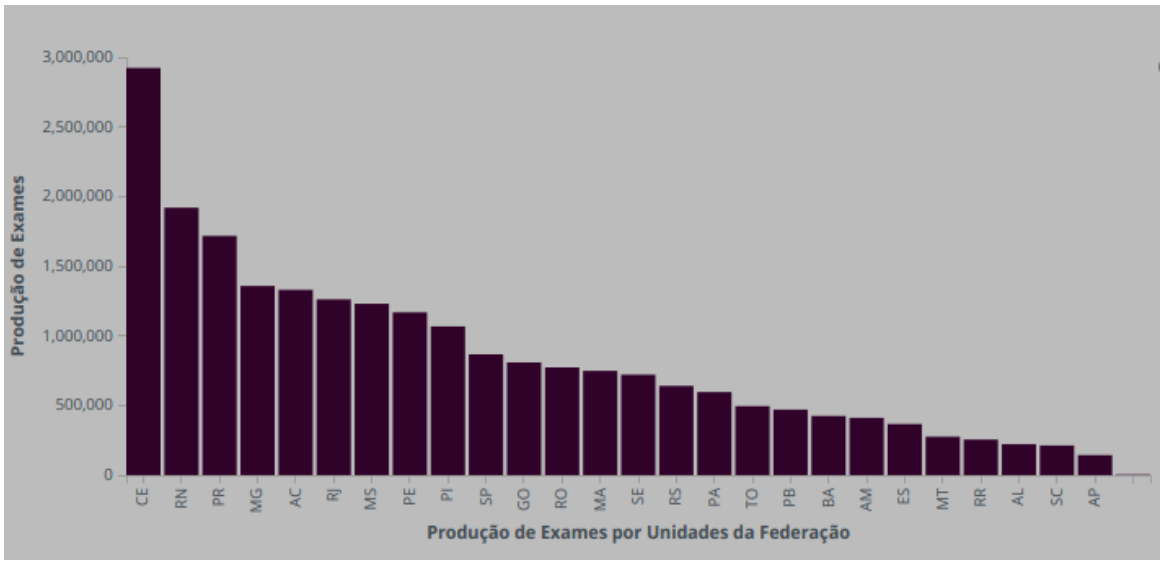



Figura 4.1: Produção de Exames por UF
(Fonte: [18]).

Nos anos de 2015, 2016 e 2017 (Figura 4.2) percebe-se certo equilíbrio nas informações, o que remete à estabilidade da curva de dados, entretanto, o sistema GAL, ainda, está em processo de implantação. O ano de 2018 é composto pelas informações de janeiro a maio desse ano, o que justifica a queda na evolução dos dados.

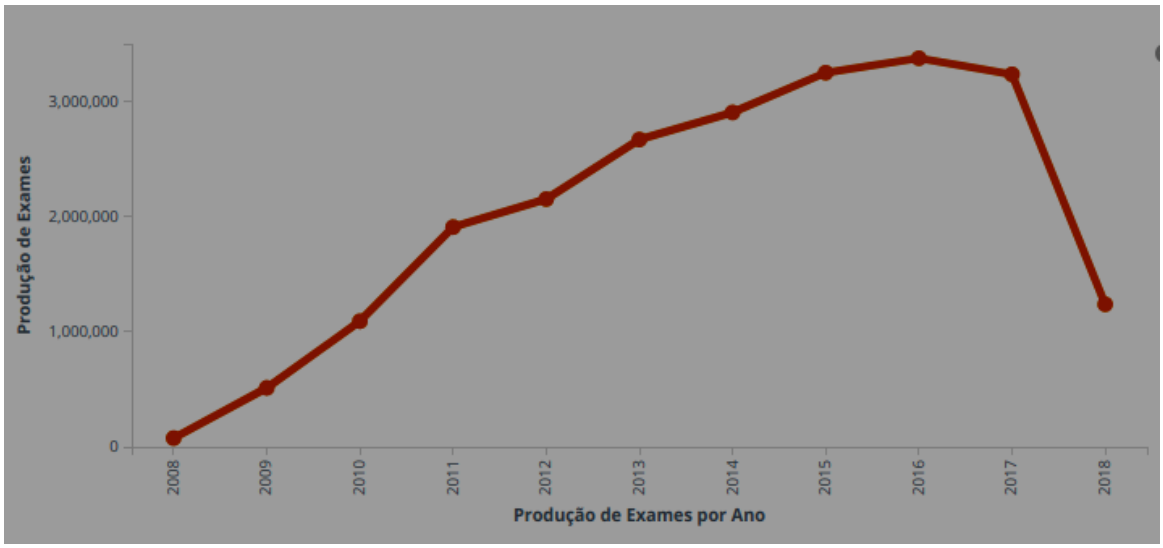



Figura 4.2: Produção de Exames por Ano
(Fonte: [18]).

Na Figura 4.3 percebe-se que mais de 50% da produção de exames laboratoriais está concentrada nas doenças de Hepatite, HIV, Dengue, Tuberculose. Entretanto, deve-se levar em consideração que os pacientes de Hepatite, HIV e Tuberculose realizam vários exames durante todo o tratamento, o que contribui significativamente para esse percentual.

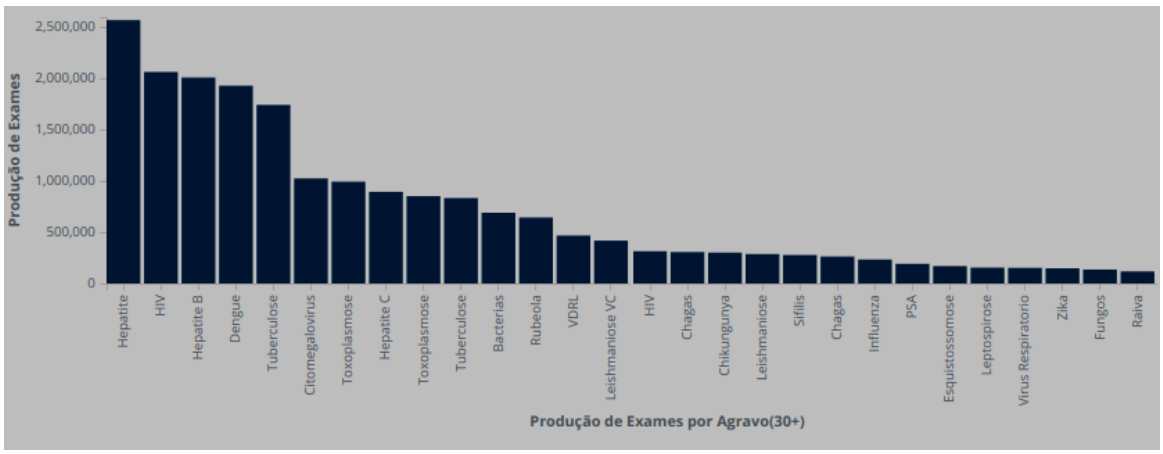



Figura 4.3: Produção de Exames por Agravos
(Fonte: [18]).

No total, 24 doenças são responsáveis pela representatividade de 95.3% das informações que são registradas. Destaca-se, ainda, os resultados referentes à circulação de novas doenças, tendo os primeiros registros de Chikungunya em 2014 e Zika em 2015.

As metodologias Enzimaimunoensaio, Imunoensaio e PCR representam 69.8% de todos os exames que foram realizados no Sistema Nacional de Laboratórios de Saúde Pública.

A doença de influenza foi escolhida para a aplicação do Controle Estatístico de Processo neste trabalho por possuir uma rede de laboratório bem estruturada e por fazer parte do monitoramento internacional (Organização Mundial da Saúde). Ressalta-se que poderia ser escolhida qualquer uma das doenças que possui informações cadastradas no sistema GAL para aplicação do CEP.

As Figuras 4.4 a 4.7 apresentam os resultados (Positivo: Vermelho e Negativo: Verde) dos exames de influenza que foram realizados e registrados no GAL.

A influenza, que ainda mata muitos pacientes, tem maior intensidade nas regiões Sul e Sudeste (Figura 4.4 e Figura 4.5) pelas características de clima mais frio. Ela é representada pelas barras *Influenza e Vírus Respiratórios*. O Paraná é o estado que realizou aproximadamente 40% de todos os exames de influenza produzidos no país, tendo registros desde o ano de 2008. Portanto, este Estado e doença tornam-se potenciais para a análise de Séries Temporais e aplicação do Controle Estatístico de Processo.

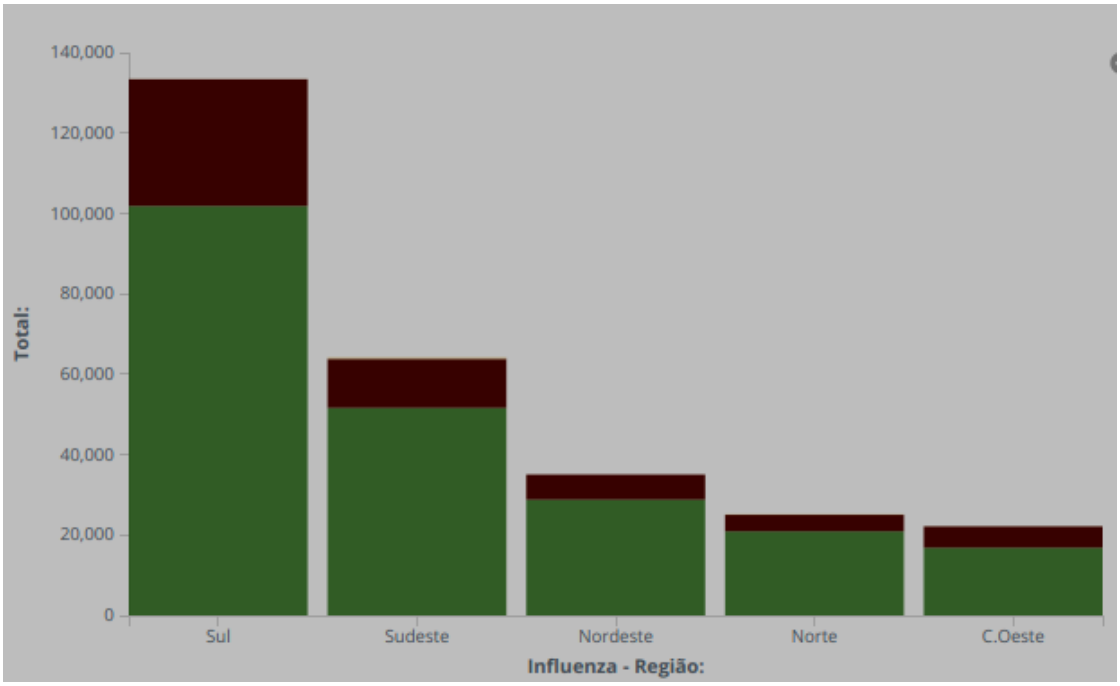



Figura 4.4: Positividade da Influenza, por Regiao
(Fonte: [18]).

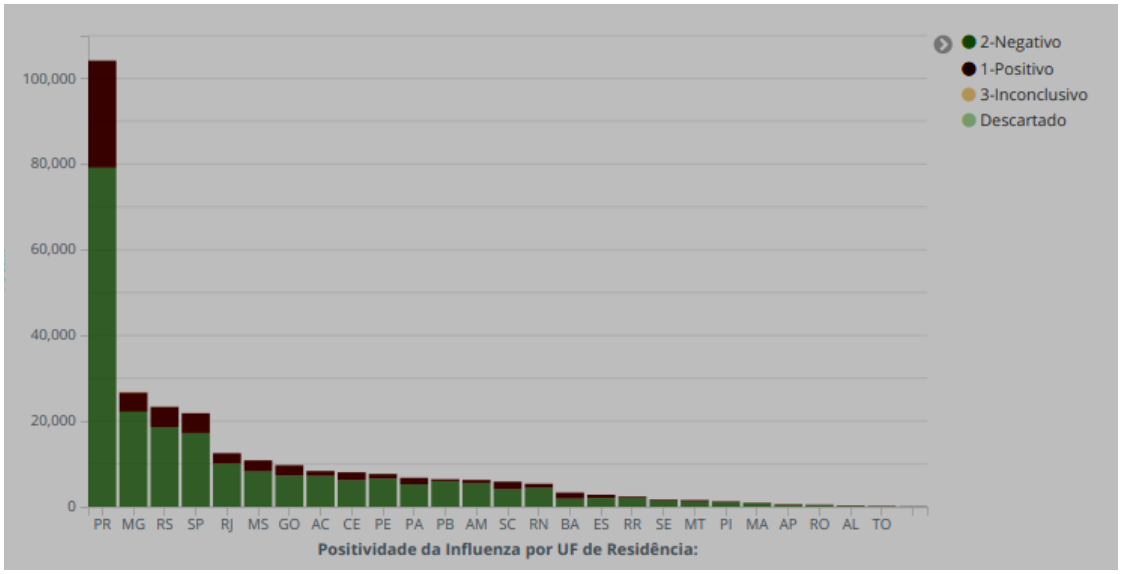



Figura 4.5: Positividade da Influenza, por UF
(Fonte: [18]).

Na mesma medida, percebe-se uma similaridade na curva dos dados quando se compara a Figura 4.2 com a Figura 4.6. Ainda, mesmo havendo um aumento no número de exames nos últimos anos, a proporção de positivos em relação aos negativos não sofreu muitas alterações.

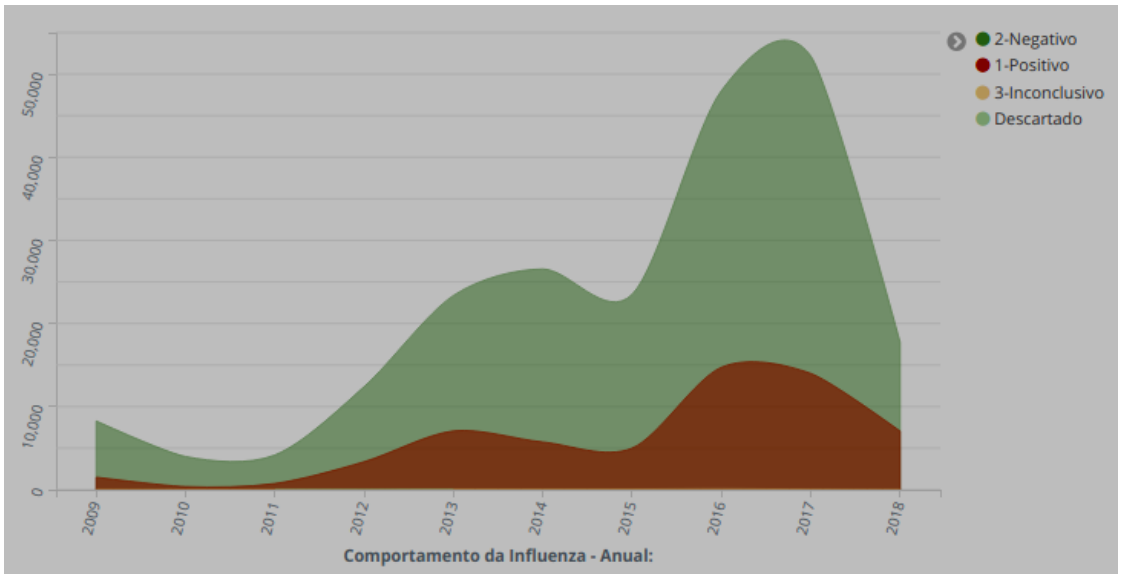



Figura 4.6: Positividade da Influenza, por Ano
(Fonte: [18]).

Na Figura 4.7 fica visível o efeito da sazonalidade da influenza, tendo início no final do mês de fevereiro, com aumento acentuado até o mês de maio. Seu declínio final acontece

no mês de agosto, retornando à normalidade em sequência.

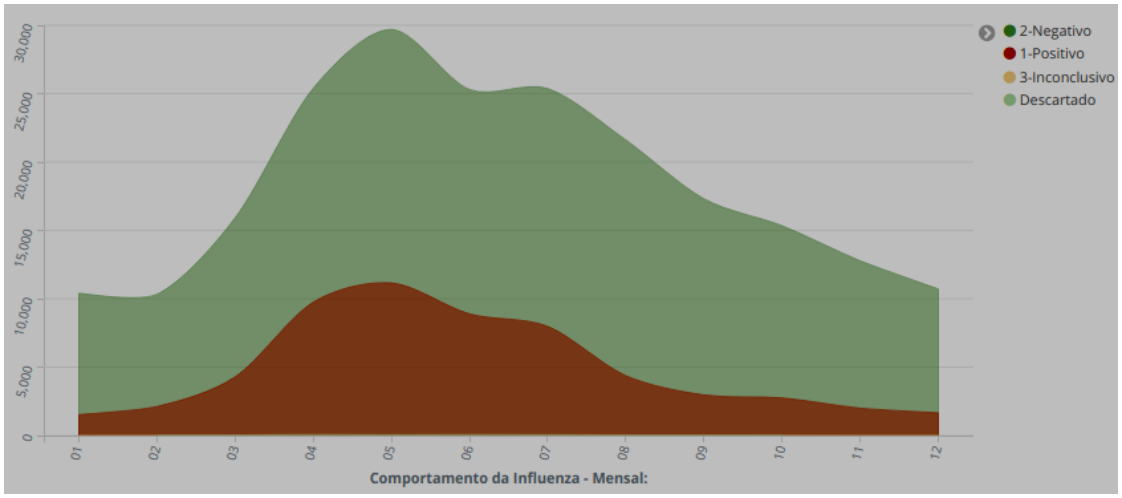



Figura 4.7: Positividade da Influenza, por Mês
(Fonte: [18]).

A Figura 4.8 exibe de forma ampliada o horizonte dos municípios que mais realizaram exames de influenza nos últimos anos (2009 a 2018). Respectivamente, os municípios que mais realizaram exames foram: Curitiba-PR, Belo Horizonte-MG, Cascavel-PR, Maringá-PR, Rio de Janeiro-RJ, São Paulo-SP, Porto Alegre-RS, Manaus-AM etc.

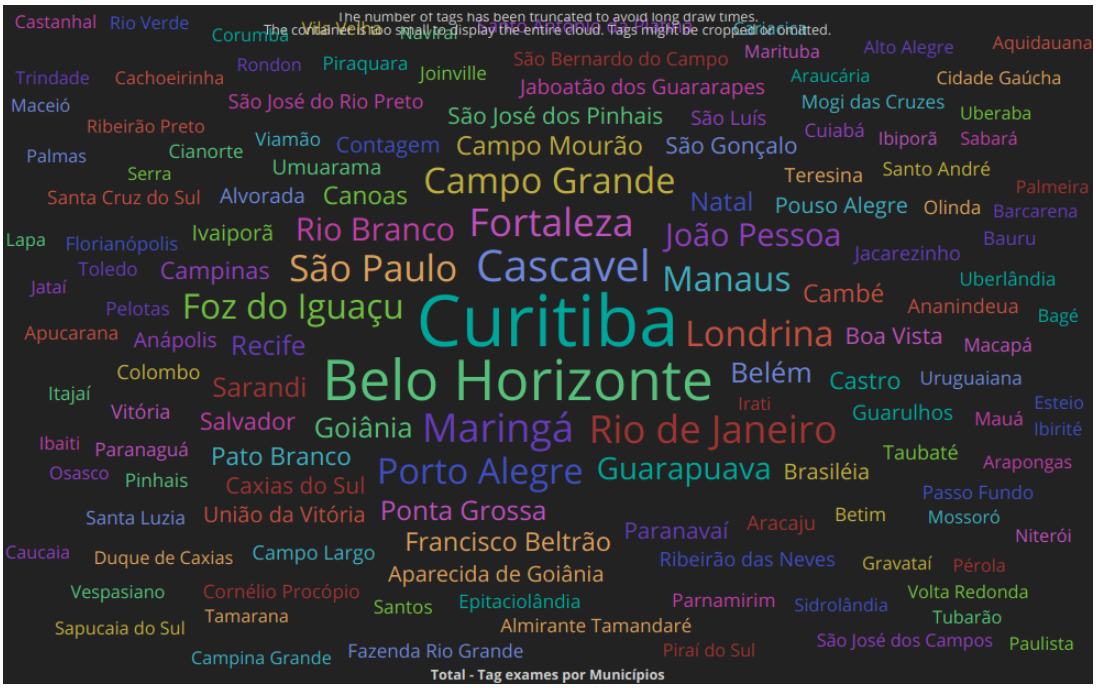



Figura 4.8: Produção de exames de Influenza, agrupados por municípios
(Fonte: [18]).

Enquanto a *TAG Município* (Figura 4.8) permite encontrar os municípios com maior incidência de influenza, o Mapa de calor, exibido na Figura 4.9, permite visualizar, em forma geográfica, a intensidade dos bolsões (concentração) da doença no país. O mapa confirma a concentração da influenza nas regiões Sul e Sudeste, destacando de cor azul mais claro as regiões com menor incidência e nas cores verde e vermelho o impacto da doença, sendo este último o mais preocupante.

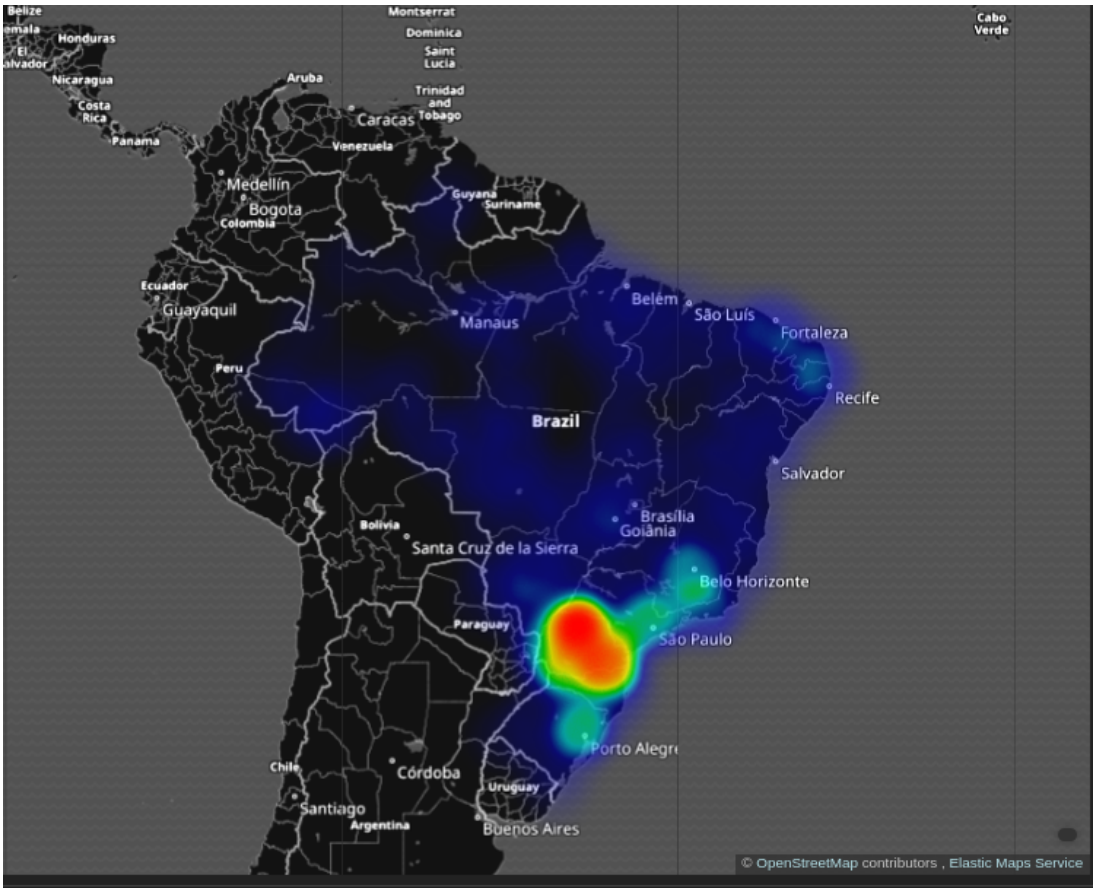



Figura 4.9: Mapa de Calor da intensidade da Influenza
(Fonte: [18]).

A análise conjunta das Figura 4.8 e Figura 4.9 possibilita maior entendimento da situação da influenza. Entretanto, não é possível afirmar que a doença está controlada ou não, o que contribui para a aplicação de Série Temporal e Gráfico de controle, buscando o Controle Estatístico de Processo da doença.

Por fim, são exportas na Tabela 4.2 as principais medidas descritivas dos dados armazenados no Sistema Gerenciador de Ambiente Laboratorial. Essas medidas auxiliam no entendimento da distribuição das informações ao longo dos anos (2009 a 2018).

Tabela 4.1: Medidas Descritivas dos Dados Laboratoriais - Brasil

| **Seq.** | **Ano** | **Mínimo** | **Máximo** | **Desvio Padrão** | **Media** | **Total** |
|---|---|---|---|---|---|---|
| 1 | 2009 | 1 | 151 | 2.520873 | 1.737361 | 511750 |
| 2 | 2010 | 1 | 84 | 2.014596 | 1.666071 | 1084732 |
| 3 | 2011 | 1 | 127 | 1.953630 | 1.664528 | 1896706 |
| 4 | 2012 | 1 | 137 | 2.222721 | 1.683882 | 2131626 |
| 5 | 2013 | 1 | 155 | 2.226937 | 1.737195 | 2649662 |
| 6 | 2014 | 1 | 135 | 2.292149 | 1.764815 | 2890661 |
| 7 | 2015 | 1 | 121 | 2.155114 | 1.695762 | 3225686 |
| 8 | 2016 | 1 | 100 | 1.968370 | 1.619788 | 3357830 |
| 9 | 2017 | 1 | 188 | 2.055209 | 1.606147 | 3181935 |
| 10 | 2018 | 1 | 238 | 2.041766 | 1.563921 | 1021444 |

Fonte: Tabela desenvolvida pelo autor

## 4.4 Solução de Business Intelligence (BI)

Considerando o arcabouço bibliográfico, as principais variáveis, os conceitos, o indicador de saúde e a necessidade do processo de monitoramento dos dados por meio de *Business Intelligence (BI)*, utilizou-se as seguintes ferramentas e técnicas para o desenvolvimento da solução Business Intelligence (BI):

1. Sistema Gerenciador de Banco de Dados (SGBD) PostGreSQL para armazenamento dos dados;
2. Pentaho Data Integration (PDI) para Extração, Tratamento e Carga dos Dados;
3. Software R, com interface R Studio, para interpretação de código em linguagem de programação *R*;
4. *Elasticsearch, Logstash e Kibana (ELK)* para processar, armazenar e visualizar os resultados dos dados.

### 4.4.1 Sistema Transacional

O sistema transacional (destacado na raia *sistema transacional* da Figura 4.10) é formado por aplicações do Sistema Gerenciador de Ambiente Laboratorial (GAL) que são distribuídas nos 27 Estados.

### 4.4.2 Sistema Dimensional

O Sistema intitulado como **Dimensional** foi desenvolvido por meio dos softwares: Sistema Gerenciador de Banco de Dados (SGBD) **PostgreSQL** (responsável pelo armazenamento das informações utilizadas na pesquisa); ***Pentaho Data Integration*** (realiza a Extração, Tratamento e Carga dos dados) e o **Software R** (responsável pela geração das informações da **Série Temporal e Gráfico de Controle**). Este processo é realizado por meio das "chamadas"do *Pentaho Data Integration* ao software **R** (*RScript*).

A Figura 4.10, raia Sistema Dimensional, ilustra o processo de geração de dados. Este processo foi detalhado na na Seção 4.4.6.

### 4.4.3 Visualização dos Dados

A visualização dos dados, destacada na última raia da Figura 4.10 é composta pelas soluções:

1. **Elasticsearch:** é um software *open source*, que fornece uma interface RESTfull para pesquisa e análise de dados.
2. **Logastash:** é o coração central do fluxo de dados do Elastic, tendo com principal função a de coletar, enriquecer e unificar todos os dados, independente do formato.
3. **Kibana:** é uma solução que possibilita e facilita a análise e a exploração de dados, e, ainda, a visualização e criação de *dashboards* de forma simples, rápida e objetiva.

Esta fase está melhor detalhada na Seção 4.4.8.

A Figura 4.10 demonstra sucintamente a solução proposta, sendo dividida em três raias principais, ou seja, sistema transacional, sistema dimensional e visualização de dados.

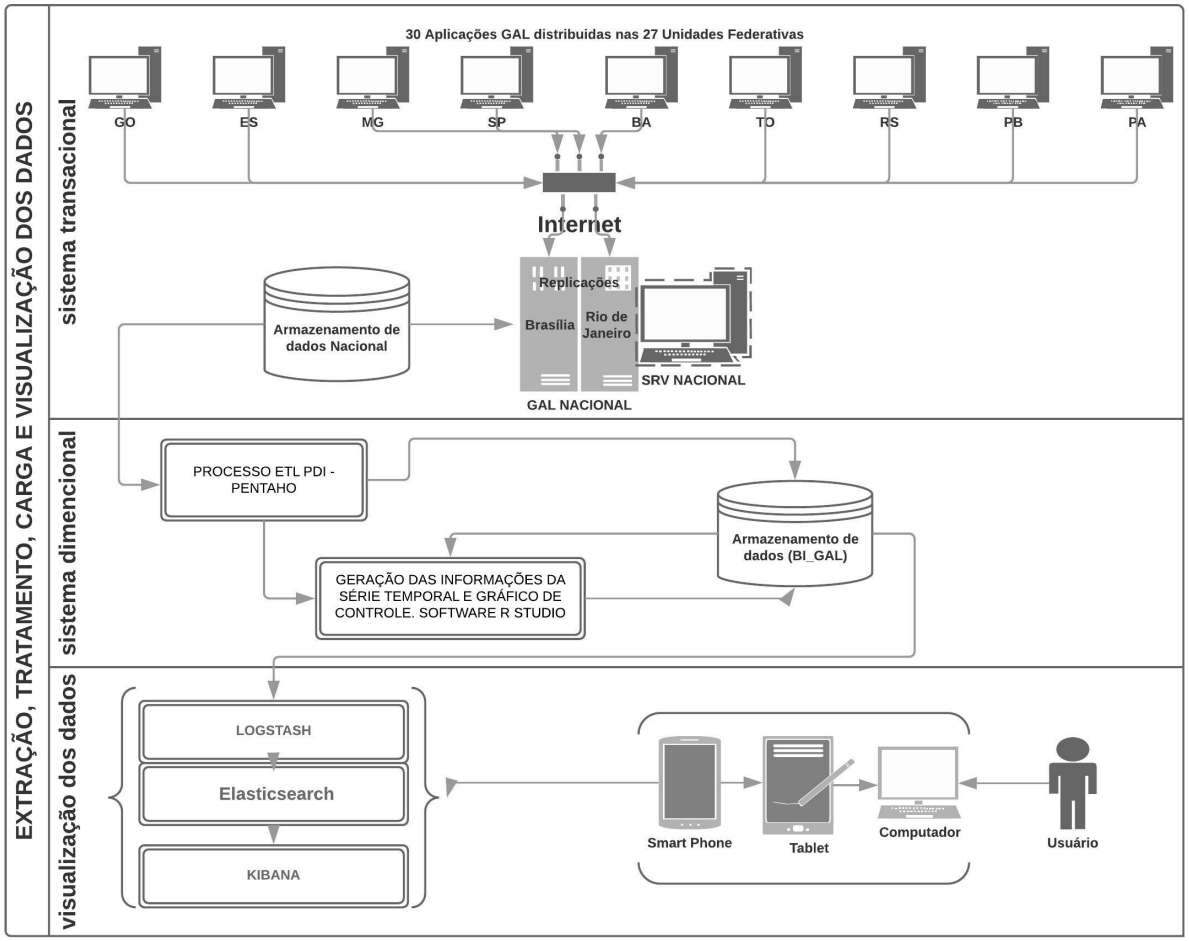



Figura 4.10: Extração, Tratamento, Carga e Visualização dos Dados (Fonte: [18]).

### 4.4.4 Indicador de Surto

A proposição do indicador foi realizada a partir do entendimento do negócio, concatenados com a necessidade de aplicar a Série Temporal, utilizando o modelo SARIMA (Sazonalidade), para a previsão e monitoramento da influenza município de Curitiba-PR, por meio do Gráfico de Controle, visando o Controle Estatístico de Processo.

**Indicador de Surtos proposto:** é soma do quantitativo de exames positivos de uma doença, somados por semana epidemiológica e ano da Data da Coleta.

$$Indicador\ de\ Surto = \sum_{i=1}^{n} S_i \tag{4.1}$$

$S_i$**:** Total de exames positivos na *i-ésima* Semana Epidemiológica[7].

$n$**:** número de Semana Epidemiológica.

### 4.4.5 Aplicação da Série Temporal e Gráfico de Controle

Nas análises de Série Temporal e Gráfico de Controle foram utilizados softwares **R**, Pentaho Data Integration e *Elasticsearch, Logstash e Kibana* para a visualização dos resultados.

O *RScript* realiza os cálculos de suavização do modelo, eliminação de tendência, sazonalidade e gera os dados (resíduos) do modelo.

Quanto ao processo de parametrização e ajustes da Série Temporal e Gráfico de Controle, foram utilizados como base os algoritmos, métodos e modelos propostos no trabalho de conclusão da graduação em Estatística, pelo Departamento de Estatística da Universidade de Brasília (UnB), de **Veloso** [19].

### 4.4.6 Extração, Tratamento e Carga dos Dados

Foram realizados estudos sobre os processos de *Extração, Tratamento e Carga (ETL)* dos dados, buscando a melhor solução e que atendesse aos requisitos do negócio. Além dos estudos, foram realizadas reuniões técnicas com os especialistas em banco de dados e técnicos de vigilância, visando a obtenção de informações relevantes a respeito dos dados.

Os processos Extração, Tratamento e Carga (ETL) dos dados foram divididos em etapas *(Jobs)* e utilizado o software *Pentaho Data Integration (PDI)*. Esta ferramenta foi escolhida por atender aos requisitos do negócio e ser de uso livre e sem custo.

[7]Semana Epidemiológica: é a distribuição temporal adotada para acontecimentos em Saúde Pública. Ela é definida por convenção internacional e é contada de domingo a sábado. o período contemplado pelos anos de 2010 a 2017, os anos foram compostos por 52 semanas epidemiológicas, salvo o ano de 2014 que possuiu 53 semanas epidemiológicas.

A *Extração, Tratamento e Carga* dos exames foram realizadas de forma a desnormalizar o Banco de Dados do Sistema Gerenciador de Ambiente Laboratorial (GAL), ou seja, concentrando as principais informações laboratoriais (paciente, localização, exames e resultados) em poucas tabelas e sem vínculo.

A desnormalização pode ser um problema se não for controlada. Por outro lado, o software *Elasticsearch* utiliza dados desnormalizados, o que levou a este processo.

**Jobs ETLs:** Nesta etapa foram concentrados esforços para capturar e desnormalizar todos os dados produzidos pela Rede SISLAB e que são armazenados no Sistema Gerenciador de Ambiente Laboratorial.

a) **1º passo - ETL dos Dados Laboratoriais:** responsável pela consulta e concentração dos principais códigos e chaves referenciais das informações laboratoriais elencadas para compor as tabelas desnormalizadas.

b) **2º passo - ETL da Produção de Exames Laboratoriais:** processo de Extração, Tratamento e Carga dos Dados responsável por agrupar e quantificar a produção de todos os exames laboratoriais do país.

c) **3º passo - ETL dos Dados Laboratoriais de Influenza:** ETL responsável por realizar a seleção de todos os exames de influenza realizados no país. Estes dados são gravados na tabela de Vigilância de influenza no Brasil.

d) **4º passo - ETL preliminares dos Dados:** realiza a transformação dos registros dos pacientes, como por exemplo, criação de novos campos, resultados (positivos/negativos), tipo de vírus, subtipagem dos vírus, região etc.

e) **5º passo - ETL RScript:** processo responsável por realizar as chamadas da aplicação ***R***, que por sua vez executa as instruções programada de modelagem, calibração e geração de dados da Série Temporal e do Gráfico de Controle.

f) **6º passo - Tabelas de Dados Laboratoriais:** foram construídas sete tabelas essenciais para armazenar, de forma desnormalizadas, os dados de suporte a esta pesquisa. Sendo estas:

i) Tabela ***ve-exames-brasil:*** construída para armazenar todos os códigos e chaves principais referentes aos dados laboratoriais produzidos no país;

ii) Tabela ***ve-producao-brasil:*** construída para armazenar todos os dados laboratoriais produzidos no país de forma agrupada e quantificada;

iii) Tabela ***ve-influenza-brasil:*** construída para armazenar todos os exames de influenza do país;

iv) Tabela ***ve-influenza-curitiba:*** construída para armazenar definitivamente todos os exames, sem duplicidades e com o *HASH* dos campos: paciente, município, data de nascimento, do município Curitiba-PR.

v) Tabela ***ve-influenza-curitiba-qcc:*** construída para armazenar os resultados das informações que são geradas pelo software ***R***, durante a execução do ***Script R***, que realiza a modelagem, suavização dos dados de influenza pelos algorítimos de Série Temporal e Gráfico de Controle.

vi) Tabela ***ve-influenza-curitiba-monitoramento:*** utilizada para armazenar todas as informações relacionadas as previsões e o monitoramento da influenza em tempo real, ou seja, os resultados positivos identificados na semana atual.

vii) Tabela ***ve-influenza-curitiba-violacao:*** utilizada para armazenar as informações de violação dos Limites de Controles do Gráfico de Controle, ou seja, armazena informações dos municípios e das semanas em que pode estar acontecendo surto de influenza ou que não esteja com os limites ou padrões de controle estipulado como normais. Estas informações são geradas pelo software ***R*** de forma automática, por meio de chamadas via software *Pentaho Data Integration* da Pentaho.

### Job de Extração, Tratamento e Carga (ETL) de Dados

A Figura 4.11 ilustra de forma global as chamadas aos processos de Extração, Tratamento e Carga (ETL) dos Dados. Cabendo ressaltar que, os processos são gerados de forma encadeada, ou seja, cada processo é executado de acordo com suas dependências, como por exemplo, os processos Extração, Tratamento e Carga (ETL) só são gerados após o primeiro passo (ETL - Rede SISLAB-Brasil - R25) ser finalizado.

Este ***Job*** está dividido em três fases, sendo:

a) ***Start:*** responsável pelo início do processamento e geração das informações, de acordo com os parâmetros do agendamento configurado para sua execução, podendo ser diário ou semanal. Por outro lado, é importante ressaltar que apesar do *Data Integration Server (DIS) - Pentaho* fornecer uma opção mais robusta para agendar a execução de tarefas e transformações, este trabalho foi executado como um processo do *daemon*.

b) ***Passos:*** no centro do ***Job*** encontra-se todos os ***Passos*** responsáveis pela geração das informações. O ***Job*** é responsável pelas chamadas a todos os ***Passos*** de acordo com a ordem programada. Neste caso, todos são executados de forma encadeada.

c) ***Sucess e Mail:*** sinaliza a finalização das execuções de todos os processos ***(Passos)***, seguido do envio de e-mail com informações *(logs)* do processo de execução do ***Job***.

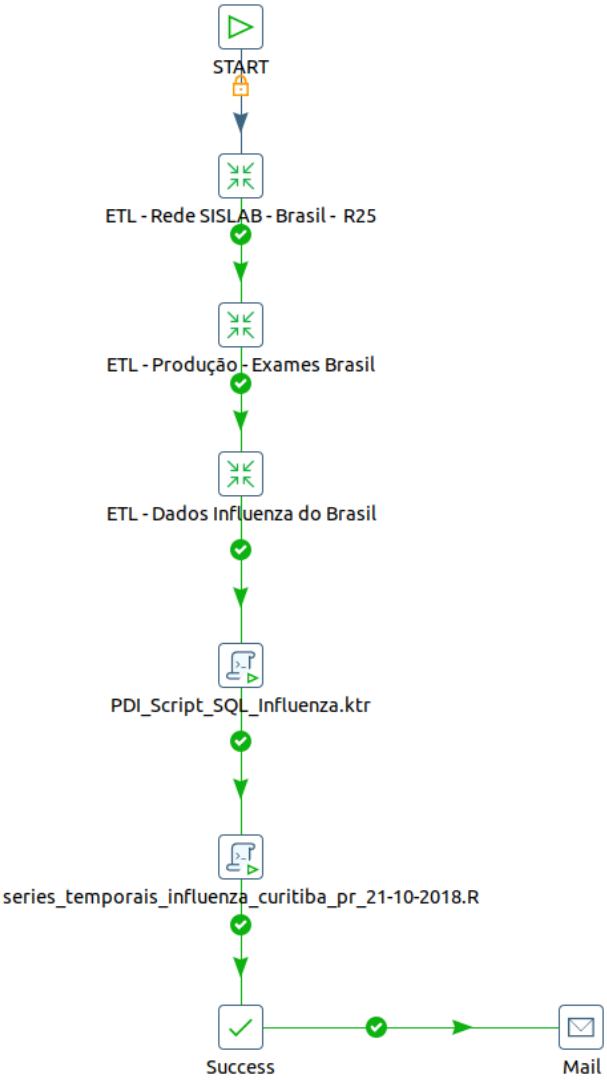



Figura 4.11: Chamadas aos Processos de Gerações dos ETL (Fonte: [18]).

### Passo de Extração, Tratamento e Carga - Informações Laboratoriais do Brasil

O Passo Extração, Tratamento e Carga das Informações Laboratoriais do Brasil, ilustrado na Figura 4.12, é composto por um processo de consulta a uma *View* que reúne informações (códigos e chaves) de várias tabelas, formando registros de dados que representem as

informações dos pacientes, exames, métodos, amostras, localidades e resultados em uma única tabela de dados.

Essa tabela torna-se um instrumento importante pois a partir dela é possível ter todos os resultados, dos mais de 850 (oitocentos e cinquenta) tipos de exames cadastrados no sistema e com diferente tipos de entradas de dados de resultados (mascaras de dados), de forma mais rápida.

Resumidamente, esta *View* permite concentrar todos os tipos de resultados dos exames em uma única tabela com 25 (vinte e cinco) campos de resultados. Este processo diminui consideravelmente o processamento de informações e acelera a retirada dos dados quando estes são urgentes.

O resultado desse processo de ETL é uma tabela sem vínculos, com os campos e informações importantes para a Vigilância em Saúde Pública (VSP), tornando-se fonte principal para os demais *Passos*.

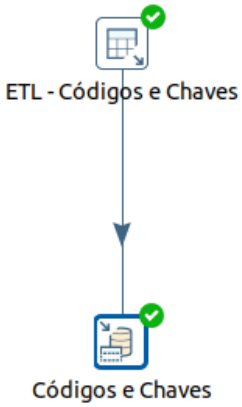



Figura 4.12: Passos - ETL das Informações Laboratoriais do Brasil (Fonte: [18]).

### Passo de Extração, Tratamento e Carga da Produção de Exames do Brasil

O Passo Extração, Tratamento e Carga dos dados referentes a produção de exames Laboratoriais do Brasil, ilustrado na Figura 4.13, tem por finalidade realizar o agrupamento das informações por Localidade (Região, UF, Municípios), Ano, Mês, Doenças/Agravos, Métodos, Status, Sexo e quantificá-las de acordo com os grupos.

Este Passo é composto por dez tabelas, sendo a principal delas a ***ve-exames-brasil*** (resultado do 1º Passo), que concentra os principais códigos relacionais com outras tabelas e as informações laboratorial do Brasil.

Esta é composta pelos principais campos do Sistema Gerenciador de Ambiente Laboratorial e a partir dela foram construídas junções entre as tabelas: *Status dos Exames,*

*Informações dos Pacientes, Informações de Localidade (Região, Municípios e UF), Classificações dos Exames, Métodos, Amostras e Resultados Laboratoriais.*

O resultado deste processo é a redução do volume de informações e a geração da tabela ***ve-producao-brasil***, que é utilizada pelo ***ElasctiSearch, Logstash e Kibana*** no processo de visualização dessas informações.

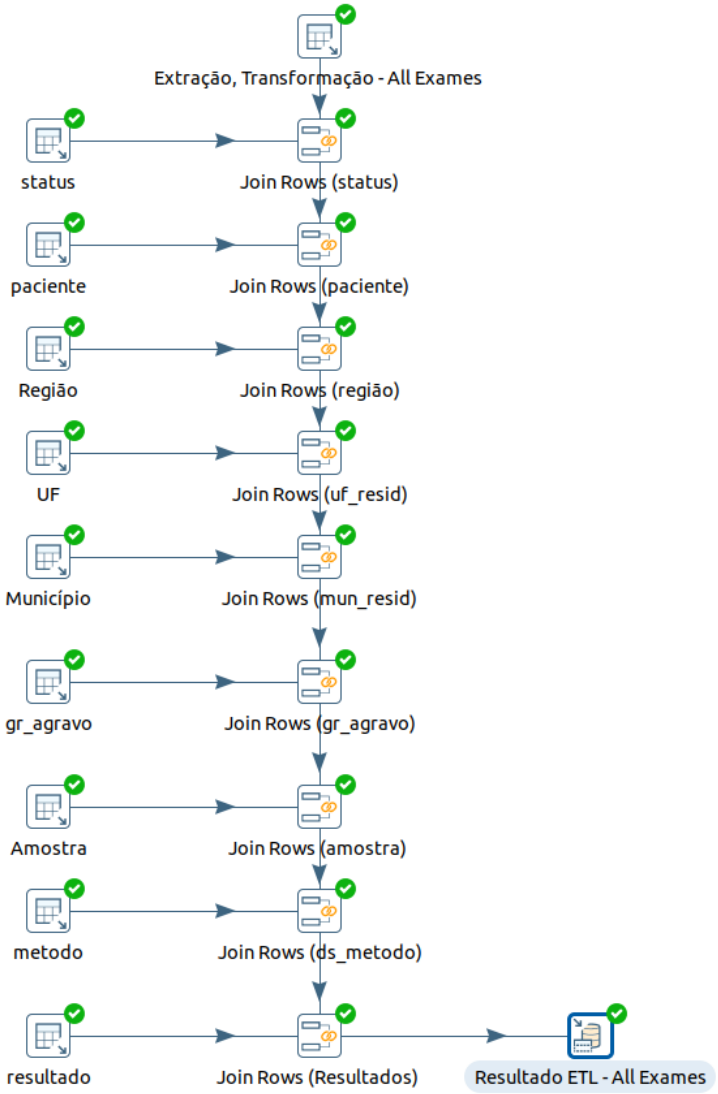



Figura 4.13: Passo - ETL da Produção de Exames do Brasil (Fonte: [18]).

### Passo: ETL dos dados de Influenza do Brasil

O processo de geração e seleção das informações do doença de Influenza inicia-se, também, pela tabela ***ve-exames-brasil***, que por sua vez, realiza as junções de outras tabelas (Região, UF, Municípios, Status, Semana Epidemiológica).

Este Passo tem por finalidade separar as informações de influenza de todo o país. Neste processo são incluídas informações de georreferenciamentos, o que permite a plotagem dos dados em mapas. As referências geográficas são incluídas a partir da tabela de municípios.

Quanto ao processo de seleção das informações de influenza é importante ressaltar que todos os métodos e tipos de resultados dos exames, que são realizados no processo de detecção da doença, são coletados por este Passo.

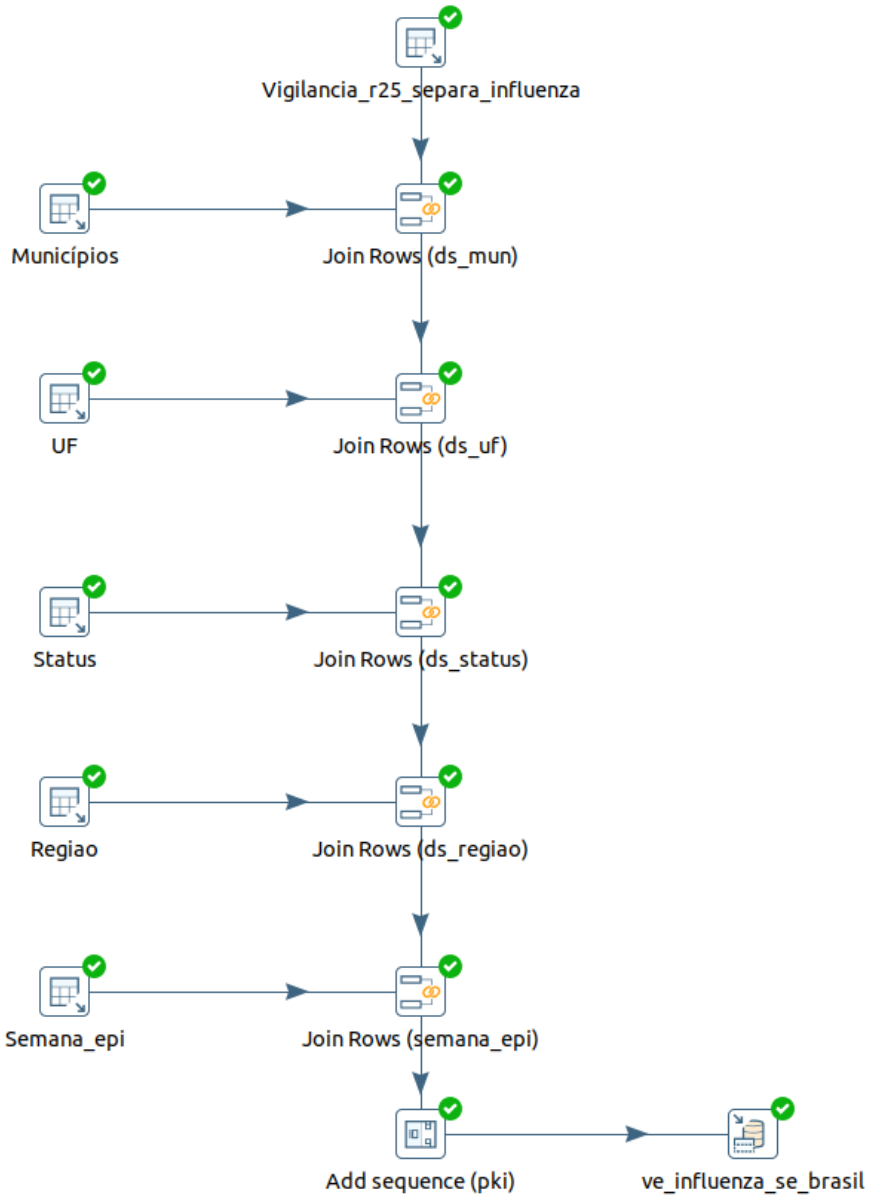



Figura 4.14: Passo - ETL dos dados de Influenza do Brasil (Fonte: [18]).

### Passo: ETL dos Resultados dos Exames de Influenza do Brasil

O ***Passo 4, ETL dos Resultados dos Exames de Influenza do Brasil***, realiza a transformação e padronização das informações dos dados de influenza.

Essas padronizações são necessárias pois há vários métodos de diagnósticos cada um com seus padrões de resultados, mas que representam a mesma informação.

Por exemplo, o método de diagnóstico ***"PCR em Tempo Real"*** adota como resultados os padrões ***"Detectável, Não detectável, Indeterminado"***, enquanto que, os métodos de diagnóstico de ***"Imunofluorescência Indireta e Indireta"*** adotam os padrões ***"Reagente, Não Reagente e Inconclusivo"***.

Esses tipos de resultados são transformados para ***Positivos, Negativos e Inconclusivos***, e ainda, há padronizações dos ***tipos de vírus em Influenza "A"ou Influenza "B"***, seguido da ***Subtipagem dos Vírus***, que dependem dos tipos de Vírus.

Parte do código deste processo pode ser visualizada no Anexo III.

### Descaracterização do Paciente

Os bancos de dados são fontes extremamente valiosas no processo de formação de novas políticas, processos e procedimentos. Entretanto, a diversidade geográfica do país, a cultura (por exemplo, a indígena que não permite pronunciar o nome da mãe após a morte) e dificuldades tecnológicas de algumas regiões brasileiras, impactam diretamente o processo de formação de bases de dados com maior qualidade nos seus registros. A criação de uma chave única, como por exemplo o Cadastro de Pessoa Física (CPF), poderia identificar corretamente e sem duplicidade os registros.

Durante o processo de análise e manipulação das informações cadastradas na base de dados do Sistema Gerenciador de Ambiente Laboratorial (GAL) percebeu-se que não há um código único obrigatório que possa identificar os pacientes de forma unívoca, o que levou a adotar como chave de identificação do paciente as informações de *Nome do Paciente, Data de Nascimento e Município de Residência.*

Por outro lado, a Lei nº 12.527/2011 que regulamentou o direito constitucional de acesso às informações públicas, previsto no inciso XXXIII do Art. 5º, no inciso II do § 3º do Art. 37 e no § 2º do Art. 216 da Constituição Federal, traz em seu escopo restrições e garantias de sigilos.

Portanto, foram adotados medidas para que as informações dos pacientes fossem ocultadas durante os processos de manipulação dos dados, ou seja, foi utilizado função de ***HASH***, algorítmo SHA256, para ocultar as informações dos pacientes, conforme descrito abaixo.

Parte do código deste processo pode ser visualizada no Anexo III.

**Passo: ETL via Script - Série Temporal e CEP**

A geração das informações da Série Temporal e Gráfico de Controle é um processo muito importante para o monitoramento da Doença de Influenza. Por outro lado, exige que os resultados dos processamentos ETL anteriores sejam executados de forma coerente, sistematizada e com alto grau de qualidade nas informações, as quais serão oferecidas como insumos de Extração, Tratamento e Carga ao Passo: *ETL via RScript - Série Temporal e Gráfico de Controle*.

Este Passo, Figura 4.15, tem seu funcionamento acoplado ao software ***R***, pois é por meio das chamadas via *Shell RScript, do Pentaho Data Integration,* que é acionada a execução do *RScript*, que por sua vez, executa todas as instruções programadas em *Linguagem R*, gravando seus resultados no Banco de Dados *PostgreSQL*. E por fim, é enviado um e-mail com o *log* do processo de execução para os endereçados como monitores e gestores das informações.

O ***ScriptR***, descrito no Anexo I desta pesquisa, é a programação completa do processo de geração dos dados (modelo) da Série Temporal e Gráfico de Controle. Parte dele é resultado do trabalho de **Veloso** [19].

O ***ETL RScript*** pode ser dividido em algumas fases, sendo:

a) **Leitura dos dados:** os dados são carregados para a memória do computador por meio das instruções *SQL* que são executadas pelo *R*. Estas instruções automatizam o processo de leitura dos dados, selecionam e padronizam as informações, e, ainda, auxiliam no processo de eliminação da duplicidade dos dados;

b) **Preparação dos Dados:** inicia-se com as instruções *SQL*, seguida dos ajustes por meio das instruções e pacotes do *R*, que transformam os dados de acordo com os padrões exigidos para o processamento e chamada à função que calcula o ***Indicador de Surto***, que é utilizado pela Série Temporal;

c) **Processamento da Série Temporal:** atua sobre os dados que há indicação de um processo de tendência ou sazonal do *Indicador* (este último é o caso dos dados de Influenza), levando à aplicação de diferenciações sazonais para ajustar o modelo de Série Temporal, ou seja, transformar a série de dados em estacionária e sem sazonalidade.

   A seleção do modelo é realizada com base nas autocorrelações e autocorrelações parciais estimadas, que são calculadas pelas Funções de Auto-correlação (FAC) e Função de Auto-correlação Parcial (FACP), que faz parte do código programado em *R* do *RScript*.

Para tornar possível o ajuste do modelo da Série Temporal do tipo SARIMA, foi necessário a definição do período em que ocorrem semelhanças no padrão da Serie Temporal.

Portanto, foi utilizado o pacote *Periodogram do R* para realizar análise espectral, afim de verificar o cosseno com amplitudes e frequências variadas.

Com auxílio do *Periodograma*, encontrou-se que essas similaridades acontecem, prioritariamente, a cada 53 semanas (s = 53). Dessa forma, como tentativa de amenizar a sazonalidade presente, foi realizada diferenciação da série em *lags sazonais de tamanho 53*.

Com apenas uma diferenciação nos dados a série de atenuou a sazonalidade presente, adequando a série de dados.

Definido o período do modelo e diferenciada a série, foi definida a ordem do modelo. Todavia, como sugerido Morettin e Toloi, citado por Montgomery [21], foram considerados alguns modelos de baixa ordem e modelos mais específicos que parecem adequados segundo a FAC e FACP. Entretanto, o gráfico da FAC e FACP não aparentaram sugestão nitidamente a um modelo.

Seguindo a orientação de Box e Jenkins, citado por Ribeiro [48], no que diz respeito à busca por modelos parcimoniosos, foi considerado como melhor modelo dentre os pré-selecionados aquele que apresentou os menores valores para os critérios de seleção AICC e BIC.

Na Tabela 4.2 é possível visualizar os modelos analisados e os respectivos valores do critério AICc e critério BIC. Os modelos com * obtiveram os melhores desempenho e índice.

Tabela 4.2: Modelos de Série Temporal de Influenza, Curitiba-PR

| **MODELO** | | **BIC** |
|---|---|---|
| $*SARIMA(1,0,1)X(0,1,1)_{53}$ | 2.275.38 | 2.292.83 |
| $SARIMA(1,0,1)X(1,1,0)_{53}$ | 2.341.75 | 2.359.20 |
| $*SARIMA(1,0,1)X(1,0,1)_{53}$ | 2.276.63 | 2.272.58 |
| $SARIMA(1,0,1)X(0,1,0)_{53}$ | 2.526.42 | 2.540.02 |
| $*SARIMA(1,0,0)X(1,1,1)_{53}$ | 2.288.86 | 2.246.31 |
| $*SARIMA(1,0,0)X(0,1,1)_{53}$ | 2.274.60 | 2.288.21 |
| $SARIMA(1,0,0)X(1,1,0)_{53}$ | 2.341.29 | 2.354.90 |
| $SARIMA(1,0,0)X(0,1,0)_{53}$ | 2.524.91 | 2.534.66 |
| $*SARIMA(1,0,0)X(0,0,1)_{53}$ | 2.244.83 | 2.262.84 |
| $SARIMA(0,0,1)X(1,1,1)_{53}$ | 2.324.64 | 2.342.10 |

| **MODELO*** | **AICc** | **BIC** |
|---|---|---|
| $SARIMA(0,0,1)X(0,1,1)_{53}$ | 2.360.48 | 2.374.09 |
| $SARIMA(0,0,1)X(1,1,0)_{53}$ | 2.424.73 | 2.438.34 |
| $SARIMA(0,0,1)X(0,1,0)_{53}$ | 2.601.40 | 2.611.15 |
| $SARIMA(0,0,0)X(1,1,1)_{53}$ | 2.478.33 | 2.491.94 |
| $SARIMA(0,0,0)X(0,1,1)_{53}$ | 2.503.17 | 2.512.92 |
| $SARIMA(0,0,0)X(1,1,0)_{53}$ | 2.572.28 | 2.582.03 |
| $SARIMA(0,0,0)X(0,1,0)_{53}$ | 2.739.67 | 2.745.55 |
| $SARIMA(0,0,2)X(0,1,4)_{53}$ | 2.282.98 | 2.261.908 |

Fonte: Tabela desenvolvida pelo autor

A Tabela 4.2 elucida os modelos de Séries Temporais (SARIMA) que foram testados e seus respectivos valores do critério *Critério de Informação de Akaike Corrigido (AICc) e critério Critério de Informação Bayesiano (BIC).*

Cinco destes modelos (destacados com *) são os mais razoáveis e mostraram-se mais adequado por possuir o menor valor para o AICc e BIC, quando comparado com os demais pares, e menor número de parâmetros na modelagem.

A escolha do modelo foi confirmada pelo cálculo do *valor médio* (que obteve a média dos resíduos é igual a 0,023 e, com p-valor=0,9) e do *teste de Ljung e Box*, que não é possível rejeitar a hipótese de que os resíduos são *"não correlacionados"*, ou seja, os resíduos têm comportamento semelhante ao de um Ruído Branco.

Após a modelagem, foi realizado a *Previsão e o Monitoramento dos dados refentes a influenza.* Este processo está descrito em detalhes na Seção 4.4.7.

d) **Processamento das Informações do Gráfico de Controle:** o modelo do Gráfico de Controle (GC) escolhido para os *Resíduos* gerados durante o processo de Séries Temporais foi o de Média Móvel Exponencial Ponderada (MMEP). Ele foi escolhido por possuir características mais flexíveis, não supõem distribuições dos valores a serem controlados e é ideal para a detecção de pequenas alterações nos valores alvos monitorados.

   Os parâmetros do Gráfico de Controle foram escolhidos de acordo com as recomendações feitas por Montgomery [21], assim como a definição de *"zero"* como valor alvo dos resíduos e considerou que quanto maior o valor atribuído ao parâmetro $\lambda$ maior é o peso dado às observações mais recentes e menos restritivos serão os limites de controle [21].

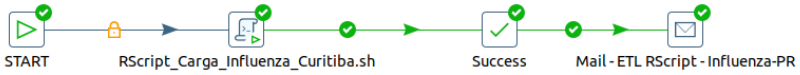
O Gráfico de Controle Média Móvel Exponencial Ponderada com os parâmetros de $\lambda$=**0,5** e $L$=**2,86** (plotagem dos resíduos do modelo suavizados que descreve o comportamento do indicador) foram escolhidos para realizar o monitoramento das observações da Influenza do município de Curitiba-PR, utilizando o (Série Temporal (SARIMA)) com base no indicador e erro de previsão), conforme Tabela 4.3.

Cabe ressaltar, que as observações passadas (dados de Influenza do município de Curitiba-PR) quando foram plotadas no Gráfico de Controle refletiu e coincidiu com a realidade enfrentada pelo município no processo de monitoramento e controle da influenza, conforme demonstram os informes epidemiológicos publicados pelo Centro de Informações Estratégicas em Vigilância (CIEVS), do Estado do Paraná.

Estes confirmam uma elevação considerável no número de casos de influenza nos anos de 2012, 2013 e 2016, o que corrobora para uma situação fora de controle e uma indicação de que pode ter ocorrido surto [64] e, por fim, reforça que o modelo está adequado para o monitoramento da influenza.

e) **Gravação dos Dados no banco de dados *PostgreSQL*, para o Monitoramento e Controle da Influenza em Curitiba-PR:** após o processo de *Extração, Tratamento e Carga*, aplicação da Série Temporal e geração dos parâmetros do Gráfico de Controle, o *RScript* captura as informações que são utilizadas pelo *software R* durante a montagem da visualização do Gráfico de Controle e gera um *data-frame*, seguido da invocação do pacote *"RPostgreSQL"* que restabelece a conexão com o banco de dados *PostgreSQL* e grava os dados do *data-frame*, utilizando a *linguagem SQL*, nas tabelas da modelagem e do monitoramento *ve-influenza-curitiba-qcc* e *ve-influenza-curitiba-monitoramento*) e também, as informações de violação dos Limites do Controles na tabela *ve-influenza-curitiba-violacao*. Esta última, armazena as informações das semanas, municípios e alterações no perfil epidemiológico da influenza do município de Curitiba-PR.

   As informações gravadas nessas tabelas são utilizadas pelo software *Elasticsearch, Logstash e Kibana* durante o processo de plotagem do Gráfico de Controle da influenza, permitindo, assim, que todos os técnicos de Vigilância autorizados possam monitorar o comportamento da doença de forma *online*.

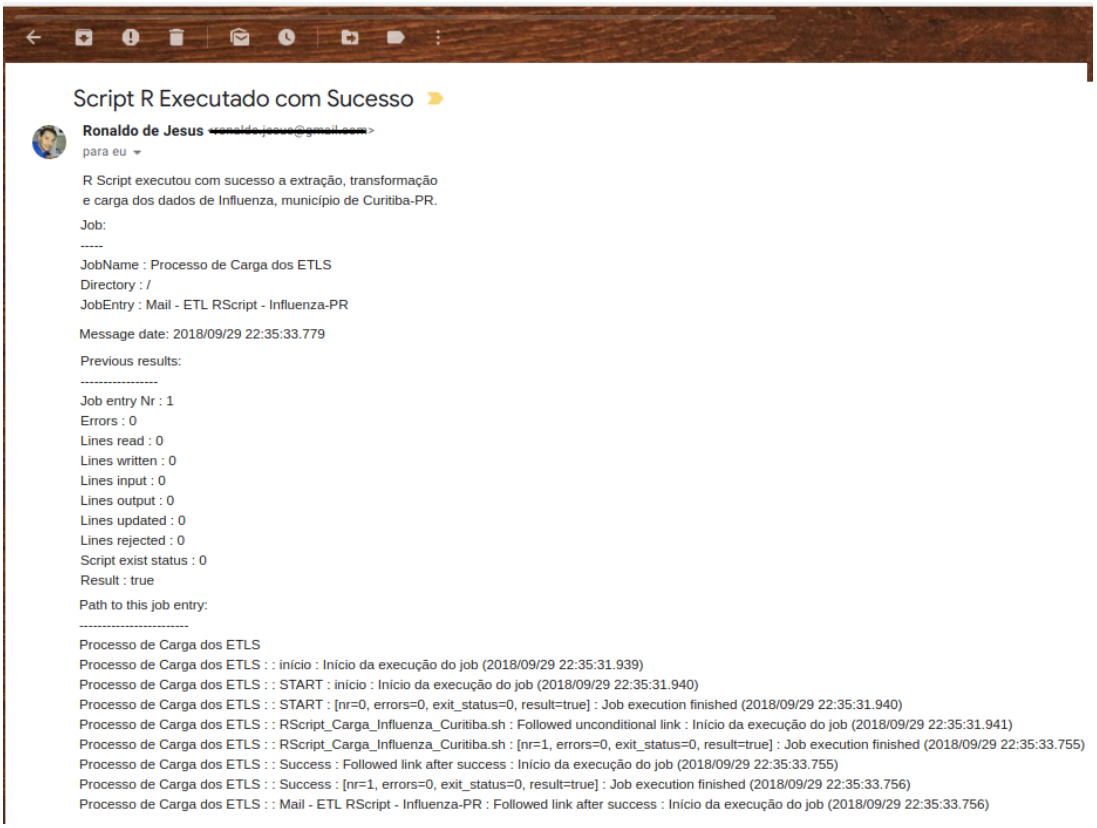



ecution Results

History | Logging | Job metrics | Metrics

| b / Entrada do Job | Coment&#65533;rio | Resultado | Raz&#65533;o | Filename |
|---|---|---|---|---|
| Processo de Carga dos ETLS | | | | |
| Job: Processo de Carga dos ETLS | Início da execução do job | | início | |
| START | Início da execução do job | | início | |
| START | Job execution finished | Sucesso | | |
| RScript_Carga_Influenza_Curitiba.sh | Início da execução do job | | Followed unconditional link | /home/ronaldo/Documentos/linux/rstudio/Scripts, |
| RScript_Carga_Influenza_Curitiba.sh | Job execution finished | Sucesso | | /home/ronaldo/Documentos/linux/rstudio/Scripts, |
| Success | Início da execução do job | | Followed link after success | |
| Success | Job execution finished | Sucesso | | |
| Mail - ETL RScript - Influenza-PR | Início da execução do job | | Followed link after success | |
| Mail - ETL RScript - Influenza-PR | Job execution finished | Sucesso | | |
| Job: Processo de Carga dos ETLS | Job execution finished | Sucesso | finished | |

Figura 4.15: Passo - ETL RScript - Série Temporal e Gráfico de Controle (Fonte: [18]).



Script R Executado com Sucesso

**Ronaldo de Jesus** <ronaldo.jesus@gmail.com>
para eu

R Script executou com sucesso a extração, transformação
e carga dos dados de Influenza, município de Curitiba-PR.

Job:
---
JobName : Processo de Carga dos ETLS
Directory : /
JobEntry : Mail - ETL RScript - Influenza-PR

Message date: 2018/09/29 22:35:33.779

Previous results:
---
Job entry Nr : 1
Errors : 0
Lines read : 0
Lines written : 0
Lines input : 0
Lines output : 0
Lines updated : 0
Lines rejected : 0
Script exist status : 0
Result : true

Path to this job entry:
---
Processo de Carga dos ETLS
Processo de Carga dos ETLS : : início : Início da execução do job (2018/09/29 22:35:31.939)
Processo de Carga dos ETLS : : START : início : Início da execução do job (2018/09/29 22:35:31.940)
Processo de Carga dos ETLS : : START : [nr=0, errors=0, exit_status=0, result=true] : Job execution finished (2018/09/29 22:35:31.940)
Processo de Carga dos ETLS : : RScript_Carga_Influenza_Curitiba.sh : Followed unconditional link : Início da execução do job (2018/09/29 22:35:31.941)
Processo de Carga dos ETLS : : RScript_Carga_Influenza_Curitiba.sh : [nr=1, errors=0, exit_status=0, result=true] : Job execution finished (2018/09/29 22:35:33.755)
Processo de Carga dos ETLS : : Success : Followed link after success : Início da execução do job (2018/09/29 22:35:33.755)
Processo de Carga dos ETLS : : Success : [nr=1, errors=0, exit_status=0, result=true] : Job execution finished (2018/09/29 22:35:33.756)
Processo de Carga dos ETLS : : Mail - ETL RScript - Influenza-PR : Followed link after success : Início da execução do job (2018/09/29 22:35:33.756)

Figura 4.16: Email com informações da Execução do Processo do ETL (Fonte: [18]).

### 4.4.7 Previsão e Monitoramento da Influenza, Curitiba-PR

O modelo de previsão de *Série Temporal* iniciou-se com o processo de análise exploratória, estimação da tendência e sazonalidade, calibração e escolha do modelo, visando minimizar os erros de estimativa dos parâmetros e, consequentemente, minimizando os de previsão.

**Período de Previsão:** a previsão busca entender acontecimento futuro com base no passado. O período de previsão desta pesquisa será de um ano decompostos em semanas epidemiológicas, disponibilizadas no site do SINAN - Ministério da Saúde.

**Comportamento dos Dados:** os dados utilizados para a previsão tem comportamento de tendência, sazonalidade e ciclicidade. A série de dados tornou-se estacionária a partir da aplicação de diferenciação da série e aplicações das técnicas de *SARIMA*.

**Série Temporal:** os dados que compuseram os testes da modelagem da Série Temporal de previsibilidade foram divididos em dois conjuntos de dados, um conjunto para estimação do modelo, compostos por dados referentes a série hatórica dos exames positivos de influenza, município de Curitiba-PR, dos anos de 2010 a 2016 e outro conjunto de experimentação do modelo, dados de Teste referentes as observações do ano 2017.

Após a realização do teste de previsão, foi aplicada a previsão dos dados para o ano de 2018, sempre prevendo as próximas 52 semanas epidemiológicas, com início da previsão a partir da data atual (último dia da execução dos processos (script de Série Temporal)). As informações das previsões geradas pela Série Temporal anteriores à data atual, são incorporadas automaticamente aos dados utilizados para a geração da Série Temporal.

***Package:*** neste etapa, os principais *Package* utilizados na modelagem e previsão dos dados da série para o monitoramento foram: *SARIMA, SARIMA.FOR, ASTSA*.

O *SARIMA* adapta-se aos modelos ARIMA (incluindo diagnósticos aprimorados) em um comando curto. Os resultados são estimativas de parâmetros, erros de previsão, AIC, AICc, BIC e diagnósticos.

**Modelo:** foi selecionado o modelo $SARIMA(1,0,0)X(0,0,1)_{53}$, por possuir componente de eliminação da sazonalidade, adequado a série de dados.

A Figura 4.17 exibe os resíduos padronizados, o *FAC* dos resíduos, um *boxplot* dos resíduos padronizados e os *p-valores* associados à *estatística-Q*.

Os resultados dos resíduos padronizados na Figura 4.17 mostra padrões esperados. Existem alguns *outliers*, no entanto, com alguns valores excedendo 3 desvios padrão em magnitude. O FAC dos resíduos padronizados mostra um pequeno desvio (insignificante) aparente dos pressupostos do modelo. O gráfico *Q-Q* normal dos resíduos mostra a partida da normalidade, com pouca variação, nas caudas devido aos *outliers* que ocorreram nas semanas epidemiológicas com números de positivos mais elevadas.

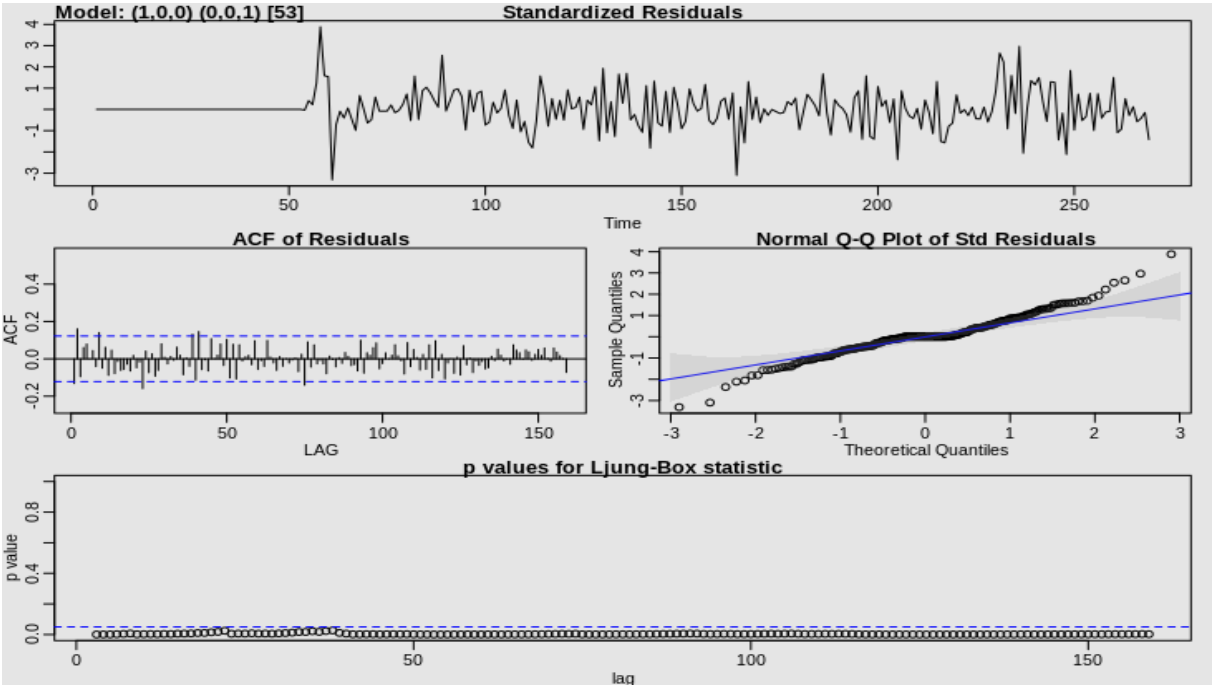



Figura 4.17: Série Temporal - Detalhamento da Previsão
(Fonte: [18]).

O ***SARIMA.FOR*** prevê os pontos futuros, ajustado-os. A saída imprime as *Previsões* e os *Erros das Previsões*, além de fornecer um gráfico da previsão com limites de erro de predição (Tabela 4.3 e Figura 4.18).

**Erro de previsão:** refere-se a *"x"* período de acontecimentos futuros no tempo *"t"*. É a diferença entre o valor monitorado (real) no instante *"t"* e a previsão deste valor em *"k"* período antes. Estes valores encontra-se na Tabela 4.3.

Tabela 4.3: Tabela de previsão para Influenza, ano de 2018, Curitiba-PR

| **Semana Epi.** | **Data Média** | **Ano-SEpi** | **Previsão (prev)** | **Erro (se)** | **Indicador Real** | **Resultado** |
|---|---|---|---|---|---|---|
| 46 | 2018-11-14 | 2018-46 | 4.056487 | 3.426357 | 2 | -2.0564866 |
| 47 | 2018-11-24 | 2018-47 | 2.600505 | 4.250799 | 2 | -0.6005048 |
| 48 | 2018-11-28 | 2018-48 | 3.261283 | 4.634472 | 2 | -1.2612825 |
| 49 | 2018-12-05 | 2018-49 | 3.167256 | 4.827866 | 1 | -2.1672563 |
| 50 | 2018-12-12 | 2018-50 | 2.696401 | 4.927387 | 0 | -2.6964014 |
| 51 | 2018-12-19 | 2018-51 | 3.851517 | 4.977235 | 0 | -3.8515172 |
| 52 | 2018-12-26 | 2018-52 | 2.260497 | 4.998372 | 0 | -2.2604965 |

Fonte: Tabela desenvolvida pelo autor

A Figura 4.18 demonstra a série dos dados, com a previsão para o ano de 2018, referente aos acontecimentos futuro da influenza no município de Curitiba-PR.

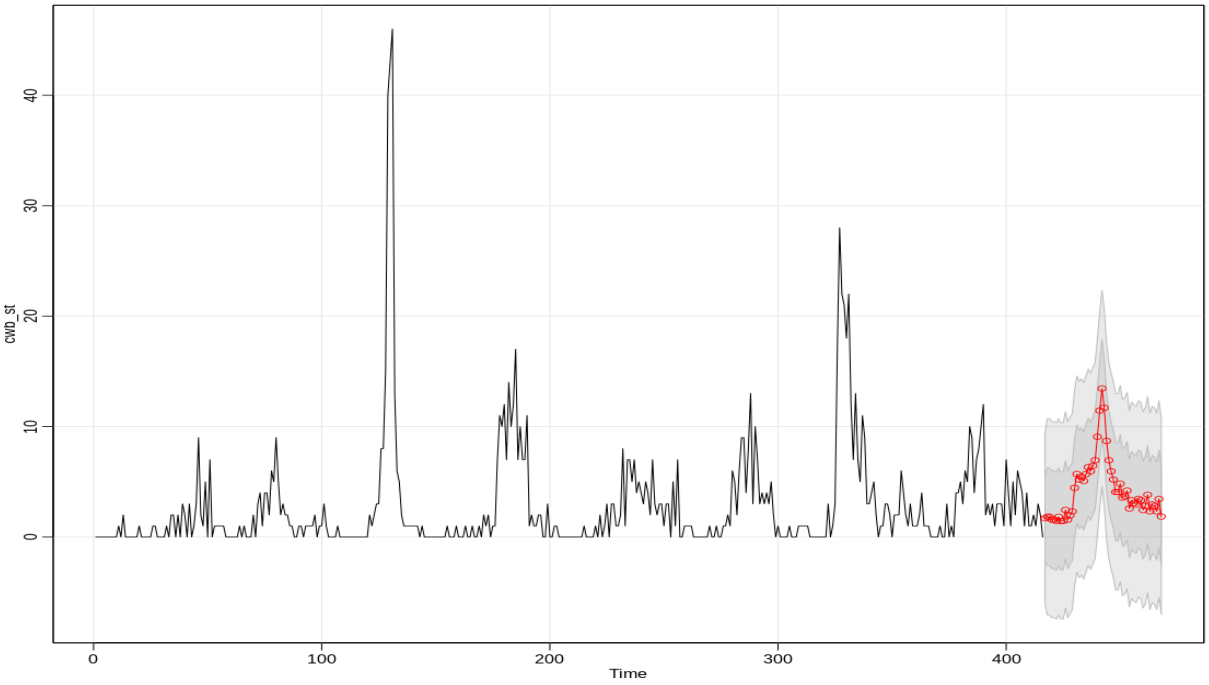



Figura 4.18: Previsão para a Influenza Curitiba-PR (2018) (Fonte: [18]).

### Monitoramento da Influenza

O monitoramento foi realizado a partir da modelagem de previsão da influenza, utilizando os dados reais dos exames positivos de influenza no município de Curitiba-PR.

A partir do *indicador real* foi calculada a série de dados com base nos dados da Tabela 4.3. Estes são concatenados com os resíduos gerados durante a modelagem da Série Temporal e submetidos a função de Média Móvel ***ewma*** com **lambda** ($\lambda$)= 0.5 e **sigma** ($\sigma$)= 2.85.

**Resumidamente**: a cada nova semana o modelo subtrai uma semana da tabela de **Previsão**, realizando o cálculo do *Indicador* e adiciona as informações ao *vetor* da Série Temporal (resíduos). Estes dados são utilizados pelo Gráfico de Controle.

Os resultados destes processos são gravados no Banco de Dados PostGreSQL, e, em seguida, são carregados para o *ElasticSearch* por meio de rotinas programadas no *Logstash*. Por fim, os Gráfico de Controle são disponibilizados em plataforma Web e *online* pelo *Kibana* (Dashboard).

Os dados monitorados, de acordo com a Figura 4.19, mostraram que pelo menos treze pontos indicaram uma situação fora de controle.

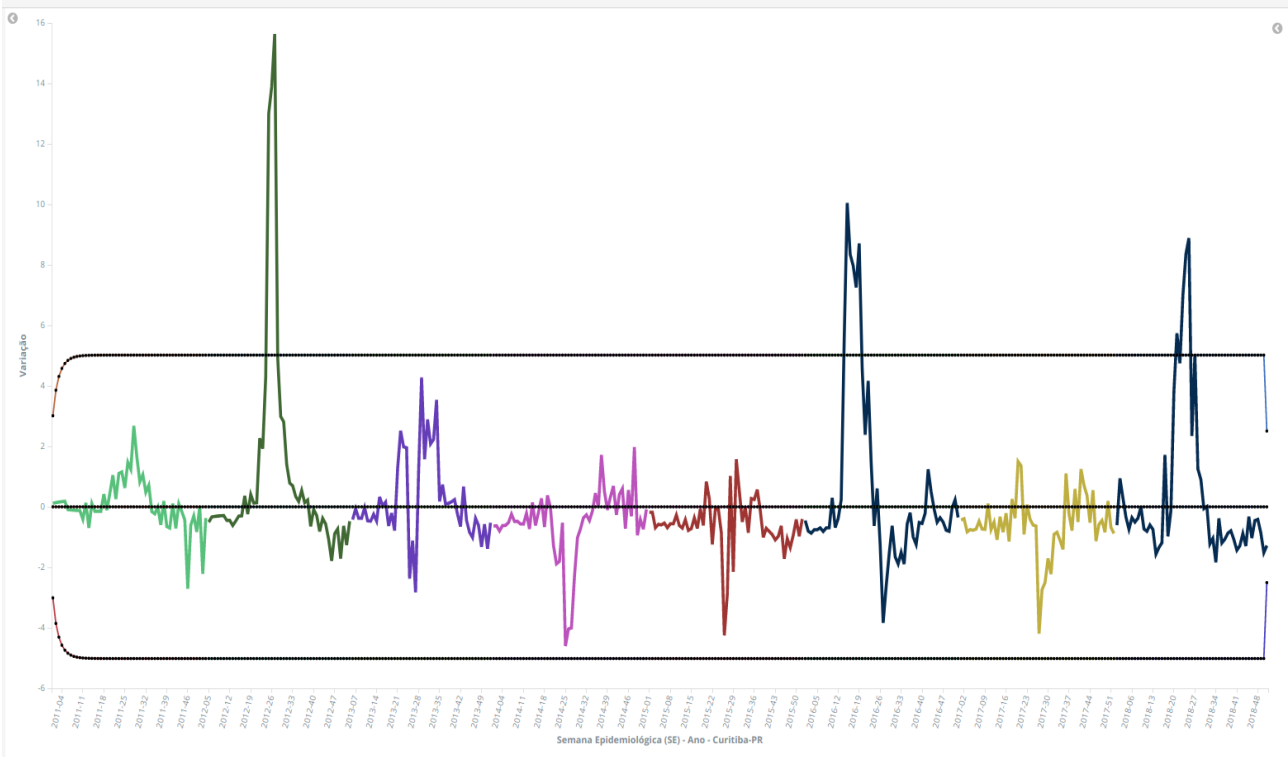



Figura 4.19: Monitoramento da Influenza, Curitiba-PR
(Fonte: [18]).

### 4.4.8 Visualização dos Dados com ELK

A visualização dos dados é a interface final dos usuários e, portanto, a apresentação dos dados deve refletir com exatidão os dados que são atribuídos a ela, de forma ágil e sempre disponível. Além disso, é preciso que seu visual seja interativo, limpo e o mais simples possível, visando sempre o atendimento dos objetivos do negócio.

A partir dessa concepção, objetivos deste trabalho e pesquisas realizadas, será utilizado o *Elastic* para a apresentação dos dados. O Elastic é um conjunto de *software* formado por Elasticsearch, Logstash e Kibana (ELK). Juntos, eles proporcionam a coleta, processamento e visualização dos dados, além de contar com mais de 200 *pluguins*. Conforme descritos na Seção 3.4.4.

As instalações do Elasticsearch, Logstash e Kibana (ELK) foram realizadas de acordo com as instruções descrita no site do *Elastic*.

#### Aplicação do *Logstash* na carga dos dados

As cargas dos dados foram realizadas com o auxílio do *pluguim Java Database Connectivity (JDBC)*, que realizou as interações com o Banco de Dados *PostgreSql* através da SQL.

Quanto aos parâmetros utilizado neste *pluguin* cabe destaque aos *statement e schedule*, o primeiro é responsável por realizar a consulta SQL nos dados, o segundo é responsável pelo agendamento da execução (data e hora ou período) dos processos *input* dos dados.

As cargas dos dados foram divididas em 4 arquivos de configurações *Logstash*, sendo:

a) **Dados de Influenza do Brasil - Georreferenciados:** responsável pela carga dos dados de **Influenza** do Brasil, *georreferenciadas por municípios* com base na *Latitude e Longitude* disponibilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

   Os parâmetros da configuração da carga dos dados estão definidos na Seção II.0.2. Entre esses parâmetros destacam-se o *schedule => "0 4 * * *"*, que executará todos os dias às 04:00hs da manhã; e, também, no processo de *Filter*, destacam-se as conversões dos campos *Latitude e Longitude* para o tipo *float* e a alteração e junção destes para *geopoint*, que é o tipo exigido para a plotagem de mapas no ELK.

   Por fim, os dados são enviados ao *Elasticsearch* para ser indexado, processado e utilizado posteriormente pelo *Kibana*. Destaca-se neste processo o *document-id* que funciona como uma *"chave primária para os dados"*. Por meio dela o *Elasticsearch* realiza o *update* nos dados.

b) **Dados de Produção de Exames do Brasil:** responsável pela carga dos dados de produção laboratorial de exames de todo o Brasil.

c) **Dados de Influenza do Brasil:** responsável pela carga dos dados de Influenza do Brasil.

d) **Dados do Gráfico de Controle:** responsável pela carga dos Dados do Gráfico de Controle da Influenza, município de Curitiba-PR.

### Visualização dos dados em Dashboard *Kibana*

O *Kibana* é uma ferramenta extremamente ágil, capaz de criar *Dashboard* com variados tipos de gráficos. Além das funções *default*, este permite a inclusão de diversos outros *pluguins* que sofisticam o processo de exploração de dados.

O *Elasticsearch, Logstash e Kibana (ELK)* nasceu com foco nos processos de buscas rápidas e monitoramentos de *logs*. Mas, a comunidade alavancou o desenvolvimento e atualmente há muitos *pluguins* disponíveis para o processo de exploração e visualização de dados. Dentre esses (alguns são pagos), encontram-se: *machine learning, graph, watcher, reporting, visualize, dashbord, timelion, index patterns, security settings.* O *kibana* tornou-se uma ferramenta muito interessante para explorar e visualizar de dados.

O *ELK* oferece muitas formas de tratar os dados, todavia, os desta pesquisa passaram por processos de Extração, Tratamento e Carga, sendo disponibilizados ao *Elasticsearch, Logstash e Kibana (ELK)* praticamente prontos para serem visualizados.

De qualquer forma, foi necessário indexar e mapear os dados no *kibana* que foram processados pelo *Logstash*.

| Na... | Health | Status | Primaries | Replicas | Docs ... | Storage s... | Primary ... |
|---|---|---|---|---|---|---|---|
| br_influenza | yellow | open | 5 | 1 | 280136 | 57.7mb | 57.7mb |
| exames | yellow | open | 5 | 1 | 13189 | 2.4mb | 2.4mb |
| geopoint_flu_br | yellow | open | 5 | 1 | 2487 | 513.6kb | 513.6kb |
| qcc_flu_cwb_a | yellow | open | 5 | 1 | 361 | 124.3kb | 124.3kb |

Figura 4.20: Gestão de Index - Painel *Kibana*
(Fonte: [18]).

Após a indexação, os dados são disponibilizados para serem utilizados. Por *default* o *kibana (versão 6.4.2)* disponibiliza sete utilitários (pluguins), dos quais foram utilizados nesta pesquisas os seguintes:

a) ***Discover:*** permitiu acessar todos índices em formato de dados padronizados, ampliando os processos de consultas, filtros, geração de relatórios em formatos *CSV)* e compartilhamento via url (Figura 4.21). Ainda, gerar histograma rápido a partir dos dados.

b) ***Visualize:*** permitiu a criação de visualizações (gráficos) dos dados desta pesquisa, a partir dos índices, consultas do *Elasticsearch* e das pesquisas salvas no *Discover*, usando o processo de agregações dos dados por soma (*sum*) ou contagem (*count*). Os gráficos gerados por esta pesquisa encontram-se no Anexo II.0.3.

c) ***Dashboard:*** foram criados e compartilhados três painéis, o do gráfico de controle de Influenza Curitiba-PR, de produção de exames e do monitoramento da influenza Brasil. Ambos foram criados a partir da coleções de gráficos gerados no *Visualize*.

d) ***Timelion:*** foi criado uma série de dados de Influenza (Brasil), distribuídas por semanas epidemiológicas, plotando os resultados do exame em positivo e negativo.

e) ***Management:*** área que permitiu gerenciar os indexes, mapeamento dos dados e configurar o *ELK*.

O Painel do kibana contam com diversos tipos de métricas, podendo citar: contagem, média, soma, mínimo, máximo, contagem única (retorna valores exclusivos), desvio padrão, top *hits* (principais valores), percentis, classificação percentual, derivada, soma cumulativa, média móvel, difusão serial, e ainda métricas personalizadas (uso do Json).

No Painel estão disponíveis opções para criação de gráficos do tipo mais básicas ao mais avançado, por exemplo: gráfico linha, àrea, pizza, barra, mapas georreferenciados, e ainda, permite utilizar a linguagem *Vega*, que dá suporte para gráficos personalizados. Ela é escrita em *JSON* e geram visualizações interativas usando HTML5, Canvas e Scalable Vector Graphics (SVG). Há também, funções para criação de tabelas (*Data table*) de forma agregadas.

### 4.4.9 Controle Estatístico de Processo - Influenza, Curitiba-PR

O **Controle Estatístico de Processo (CEP)** da Influenza, Curitiba-PR, foi construído a partir das informações que foram processadas e gravadas pelo processo Extração, Tratamento e Carga, executado pelo Software *R*.

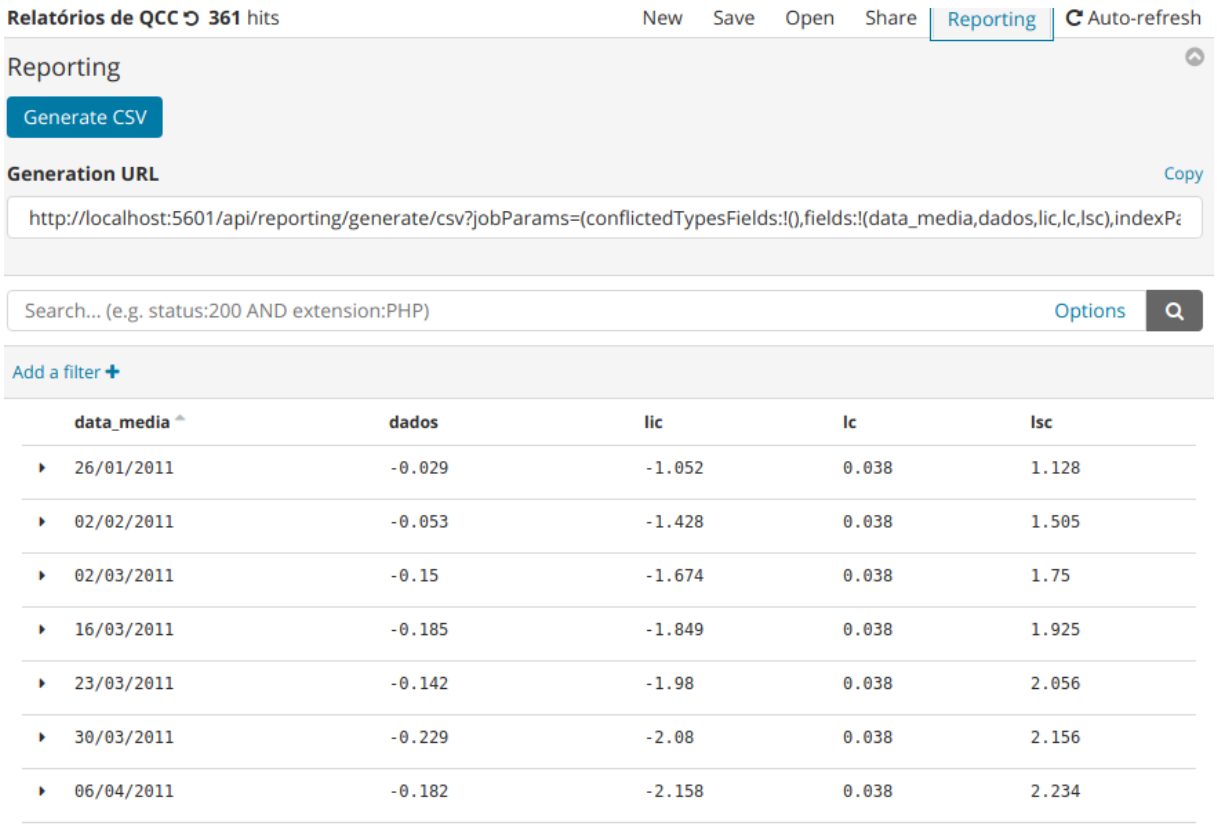
**Carregamento dinâmico dos Dados do CEP de Influenza:** os dados são acessados e carregados para o *kibana* através do *Logstash*, seguido do processo de indexação.

**Estrutura dos dados do CEP:** a Figura 4.21 traz a estrutura dos dados que foram indexados e disponibilizados pelo *ELK*, para serem utilizadas, neste caso, na geração do Gráfico de Controle Estatístico de Processo da influenza.

**Entendendo as variáveis:** as variáveis dos dados carregados e necessárias para o Controle Estatístico de Processo da Influenza, são:

1. **Data média**: refere-se ao ponto médio das datas da semana epidemiológica. Ela é uma variável do tipo *Date*. A Média foi calculada a partir das variáveis *data de início e data final* da semana epidemiológica. Elas são disponibilizadas no site do Ministério da Saúde.
2. **Dados**: esta variável armazena as informações da Influenza (*os Resíduos* gerados pela Série Temporal), do município de Curitiba-PR. Ela é uma variável do tipo *Flaot*.
3. **LIC, LC e LSC**: estas variáveis armazenam os parâmetros do *Limite Inferior do Controle (LIC), Limite Central do Controle (LC) e Limite Superior do Controle* do Controle Estatístico de Processo (CEP), que foram gerados a partir dos **Dados**, dos valores de ***lambda*** ($\lambda$) e de ***sigma*** ($\sigma$), pelo Gráfico de

Controle *Média Móvel Exponencial Ponderada (MMEP)*. Estas variáveis são do tipo *Float*.




| data_media | dados | lic | lc | lsc |
|---|---|---|---|---|
| 26/01/2011 | -0.029 | -1.052 | 0.038 | 1.128 |
| 02/02/2011 | -0.053 | -1.428 | 0.038 | 1.505 |
| 02/03/2011 | -0.15 | -1.674 | 0.038 | 1.75 |
| 16/03/2011 | -0.185 | -1.849 | 0.038 | 1.925 |
| 23/03/2011 | -0.142 | -1.98 | 0.038 | 2.056 |
| 30/03/2011 | -0.229 | -2.08 | 0.038 | 2.156 |
| 06/04/2011 | -0.182 | -2.158 | 0.038 | 2.234 |

Figura 4.21: Informações do Gráfico de Controle - Influenza, Curitiba-PR (Fonte: [18]).

**Métrica:** o Gráfico de Controle gerado pelo software *R* plotam as informações direto das variáveis, que são armazenados em vetores, não sendo necessário gerar agregação ou soma.

Entretanto, o *kibana* exige pelo menos uma *Métrica, um Eixo "X", um Eixo "Y"e um tipo (linha, barra etc)* para a geração de gráfico.

A geração do gráfico de CEP da influenza, na aplicação *kibana*, foi possível a partir da criação de *Quatro Métricas e Buckets* (uma agragação do tipo *Terms* e um *Split de Série*), conforme Figura II.1:

1. **Dados:** agregação das informações armazenadas na variável *Dados* por *Soma* no *Eixo "Y"*;
2. **Limite Inferior do Controle (LIC):** agregação das informações armazenadas na variável em *LIC* por *Soma* no *Eixo "Y"*;

3. **Limite Central do Controle (LC):** agregação das informações armazenadas na variável *LC* por *Soma* no *Eixo "Y"*;
4. **Limite Superior do Controle (LSC):** agregação das informações armazenadas na variável *LSC* por *Soma* no *Eixo "Y"*;
5. ***Buckets (Baldes):*** as métricas foram agregadas no *Eixo "X"* a *"Terms"* (termos do mesmo tipo) utilizando a variável *"Data Média"*, que é a representação da *Semana Endemiológica*. Foi adicionado um *Split de Série*, com os intervalos divididos por anos (2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018). O *Split de Série* definiu uma cor na linha para cada ano da métrica dos *Dados*.
6. **Ordenação:** os dados foram ordenados por Semana Epidemiológica (variável *"Data Média"*) e por especificação *Alfabética* crescente.

**Obs.:** A utilização da agregação a *Terms* por *Data Média* e as *Soma das Métricas* não afetam a geração do Gráfico de Controle, pois há apenas uma informação para cada métrica em cada *Bucket* da *Data Média*. Por exemplo: no *Bucket Data Média 22/11/2017* tem-se os seguintes valores: LIC: -2.462, LC: 0.038, LSC: 2.538 e Dados: -0.922, ou seja, não há nenhuma outra informação neste *Bucket* para ser somada em qualquer das métricas.

**Agregação a *Terms***: permite que se especifique os elementos de um determinado campo para exibição de termos do mesmo tipo, ordenados por contagem ou por uma métrica personalizada.

**Tipo de Gráfico**: assim como no gráfico de MMEP, que é formado por linhas, o gráfico: "Visualização dos dados do Controle Estatístico de Processo (CEP) da Influenza"é do tipo ***linha***.

Visando facilitar o acompanhamento anual da Influenza, foram criados Gráficos de Controles do Controle Estatístico de Processo da Influenza separados por anos. Eles estão destacados nas *Figura II.3, Figura II.9, Figura II.8, Figura II.7, Figura II.6, Figura II.5 e Figura II.4 do Anexo II.0.3*.

O Gráfico de Controle identificou três momentos (ano: 2012, 2016 e 2018) em que a influenza, no município de Curitiba-PR, ultrapassou o Limite Superior do Controle. Estes pontos, foram armazenados na ***tabela de Violação dos Limites de Controles*** e a partir delas foi gerado um relatório do tipo *Table Kibana*, conforme demostrado na Figura 4.23.

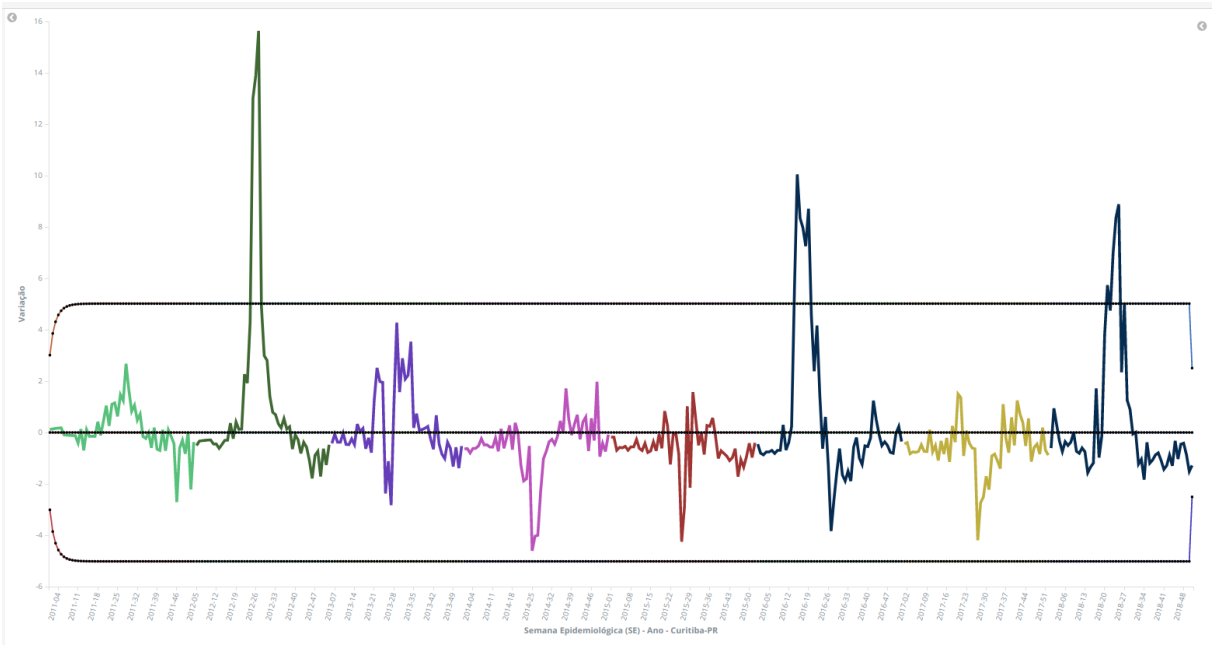



Figura 4.22: CEP da Influenza (2011 a 2018), Curitiba-PR (Fonte: [18]).

| Cód IBGE | Municíoio Residência | Ano-Semana_Epi | Dados | LIC | LC | LSC |
|---|---|---|---|---|---|---|
| 4106902 | Curitiba-PR | 2012-25 | 6.50168917350558 | -2.50664498682565 | 0 | 2.50664498682565 |
| 4106902 | Curitiba-PR | 2012-26 | 6.94875271187535 | -2.50664498682565 | 0 | 2.50664498682565 |
| 4106902 | Curitiba-PR | 2012-27 | 7.81798651834548 | -2.50664498682565 | 0 | 2.50664498682565 |
| 4106902 | Curitiba-PR | 2016-14 | 2.76480879713435 | -2.50664498682565 | 0 | 2.50664498682565 |
| 4106902 | Curitiba-PR | 2016-15 | 5.02334210485912 | -2.50664498682565 | 0 | 2.50664498682565 |
| 4106902 | Curitiba-PR | 2016-16 | 4.16973641925077 | -2.50664498682565 | 0 | 2.50664498682565 |
| 4106902 | Curitiba-PR | 2016-17 | 3.98600303348959 | -2.50664498682565 | 0 | 2.50664498682565 |
| 4106902 | Curitiba-PR | 2016-18 | 3.62542361920465 | -2.50664498682565 | 0 | 2.50664498682565 |
| 4106902 | Curitiba-PR | 2016-19 | 4.35587893877538 | -2.50664498682565 | 0 | 2.50664498682565 |
| 4106902 | Curitiba-PR | 2018-21 | 2.86815069009916 | -2.50664498682565 | 0 | 2.50664498682565 |
| 4106902 | Curitiba-PR | 2018-23 | 3.49755495163099 | -2.50664498682565 | 0 | 2.50664498682565 |
| 4106902 | Curitiba-PR | 2018-24 | 4.17813326770511 | -2.50664498682565 | 0 | 2.50664498682565 |
| 4106902 | Curitiba-PR | 2018-25 | 4.44229426183233 | -2.50664498682565 | 0 | 2.50664498682565 |
| 4106902 | Curitiba-PR | 2018-27 | 2.51159817289924 | -2.50664498682565 | 0 | 2.50664498682565 |

Export: Raw Formatted

Figura 4.23: Tabela de Violação dos Limites de Controles da Influenza (Fonte: [18]).

# Capítulo 5

# Conclusões

O Controle Estatístico de Processo (CEP) tem sido um método eficaz no processo de gestão da Qualidade da Industria e é, também, aplicada a várias outras áreas.

Os dados laboratoriais produzidos pela Rede Nacional de Laboratórios de Saúde Pública, armazenados no Sistema Gerenciador de Ambiente Laboratorial, mostraram-se fonte importante para a Vigilância em Saúde Pública.

Entretanto, há muitos métodos de diagnósticos com diferentes formas de apresentar os mesmos resultados e, ainda, cada doença possui suas particularidades, tornando mais complexo o processo de modelagem das soluções de monitoramento. Os exames de influenza, por exemplo, são realizados pelos métodos de diagnósticos PCR em Tempo Real, Imunofluorescência Direta/Indireta e Sequenciamento.

Os dados utilizados nesta pesquisa foram extraídos, transformados e carregados para um banco de dados *PostgreSQL* de forma dinâmica, utilizando o *Pentaho Data Integration* e script programado em linguagem *R*. Foram realizados vários processos e procedimentos por estes softwares, cabendo destaque aos processos de seleção dos dados, transformação dos exames, remoção da duplicidade de informações, processamento do *RScript* de forma automática, com aplicações das técnicas de Séries Temporais e Gráfico de Controle.

A remoção da duplicidade de informações foi um fator preponderante observado, pois não havia uma chave única que identificasse corretamente os pacientes e os usuários do sistema GAL cadastram um novo paciente para cada pedido de exames médicos, gerando vários cadastros do mesmo paciente. Visando preservar a identidade dos pacientes, foi utilizada a técnica de *HASH* para a criação de uma chave, com base nas informações dos pacientes (nome, mãe, data de nascimento, município de residência), para eliminar a duplicidade. A limpeza restringiu-se a eliminar duplicidade de pacientes que foram cadastros corretamente em todas as consultas médicas, caso contrário, a técnica considerou dois pacientes distintos. Foi priorizado o resultado positivo em detrimento do negativo.

A partir do banco de dados, com as informações tratadas, foi possível executar a Série de Temporal e Gráfico de Controle, aplicando o trabalho de conclusão do curso em Estatística de Veloso [19].

Entretanto, o processo de automação acarretou em mudanças no código do *RScript*, incluído procedimentos de leitura dinâmica das tabelas, tratamento adicional nos dados, padronização dos tipos de variáveis, captura dos dados e parâmetros do Gráfico de Controle de Média Móvel Exponencial Ponderada. Estas informações e as de *Violações dos Limites de Controles*) foram programadas para serem gravadas automaticamente nas tabelas de Vigilância Epidemiológica do banco de dados *PostgreSQL*.

A visualização dos dados laboratoriais e do Controle Estatístico de Processo da influenza de forma automática exigiu a utilização de software que fosse capaz de disponibilizá-los de forma *online*, tornando-os acessíveis a vários usuários ao mesmo tempo. Portanto, um servidor de aplicação *Elasticsearch, Logstash e Kibana* foi configurado, dando espaço para o processo de *Input, Filter e Output* dos dados por meio do *Logstash*, suportando o processo de busca dos dados direto do *PostgreSQL* e os transferindo para o *Elastic* e *Kibana*.

O desenvolvimento do Gráfico de Controle de Média Móvel Exponencial Ponderada, do Controle Estatístico de Processo da influenza de forma dinâmica e *online* no *Kibana*, foi viabilizado a partir da criação de quatro métricas (Limite Inferior do Controle, Limite Central do Controle, Limite Superior do Controle e Dados) do tipo soma, que passaram a compor o *"Eixo Y"*. Os *Buckets* (baldes) compuseram a métrica do *Eixo "X"*, por agregação a *"Terms"* (termos do mesmo tipo) e *Split de Série* (grupos), a partir da variável *"Data Média"*, que é a representação da Semana Endemiológica.

O monitoramento das informações laboratoriais e do Gráfico de Controle Estatístico de Processo da influenza foi disponibilizado em *Dashboard*, via URL (web) por meio do *Kibana*.

Este trabalho alcançou os objetivos propostos, criando uma arquitetura e proposta de monitoramento e controle da doença de influenza no município de Curitiba-PR. Com o monitoramento foi possível detectar três momentos (2012, 2016 e 2018) em que a doença de influenza ultrapassou os limites de controle, o que indicou uma situação fora do controle e possível surto. Ademais, foi construído uma aplicação de monitoramento do processo de produção de exames que auxilia no monitoramento de doenças e agravo à saúde.

Porém, durante o processo de modelagem da Série Temporal e do Gráfico de Controle observou-se que vários modelos e parâmetros poderiam ser aproveitados para a geração dos resíduos que são utilizados no Gráfico de Controle. Entretanto, encontrar o melhor modelo para monitorar e controlar todas as doenças, torna-se um desafio (limitador), pois, é necessário um modelo específico para cada doença e município.

Embora existam muitos desafios no processo de formação de bases de dados com informações confiáveis e que respondam à realidade atual, este trabalho demostrou de forma efetiva a aplicação de técnicas e procedimentos para o Controle Estatístico de Processo da influenza de forma online e acessível a vários usuários ao mesmo tempo, indicando um caminho para a aplicação dinâmica do CEP a outras doenças.

Por fim, outras pesquisas precisam ser realizadas, como por exemplo, a modelagem de esquema para o armazenamento e acesso a dados de suporte ao Controle Estatístico de Processo de forma dinâmica e que atenda várias doenças simultaneamente e a aplicação de outras técnicas de Séries Temporais e Gráfico de Controle na detecção de alterações do comportamento de doenças.

# Referências


[1] Carvalho Fortes, Paulo Antônio de e Helena Ribeiro: *Saúde global em tempos de globalização*. Saúde e Sociedade, 23(2):366–375, 2014. http://www.scielo.br/pdf/sausoc/v23n2/0104-1290-sausoc-23-2-0366.pdf. vi, vii, 1

[2] McMichael, Anthony J.: *Globalization, climate change, and human health*. New England Journal of Medicine, 368(14):1335–1343, 2013, ISSN 0028-4793, 1533-4406. http://www.nejm.org/doi/abs/10.1056/NEJMra1109341, acesso em 2017-04-27. vi, vii, 1

[3] Woodward, David, Nick Drager, Robert Beaglehole e Debra Lipson: *Globalization and health: a framework for analysis and action*. Bulletin of the World Health Organization, 79(9):875–881, 2001. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2566657/pdf/11584737.pdf. vi, vii, 1

[4] Silva, Luiz Jacintho da: *Vigilância epidemiológica: uma proposta de transformação*. Saúde e Sociedade, 1(1):7–14, 1992. http://www.scielo.br/scielo.php?pid=S0104-12901992000100003&script=sci_arttext&tlng=es, acesso em 2017-04-14. vi, vii, 1, 15

[5] Arreaza, Antonio Luis Vicente e José Cássio de Moraes: *Vigilância da saúde: fundamentos, interfaces e tendências*. Ciência & Saúde Coletiva, 15(4):2215–28, 2010. https://www.researchgate.net/profile/Antonio_Arreaza/publication/250028604_Vigilancia_da_saude_fundamentos_interfaces_e_tendencias/links/56a65a3d08aeca0fddcb50e8.pdf, acesso em 2017-04-14. 1, 14, 15, 18, 19

[6] Saúde, Ministério da: *Ministério da saúde: Atuação - vigilância em saúde*, 2017. http://portalms.saude.gov.br/vigilancia-em-saude/atuacao, acesso em 2017-08-25. 1

[7] Bates, Mary: *Tracking Disease: Digital Epidemiology Offers New Promise in Predicting Outbreaks*. IEEE Pulse, 8(1):18–22, 2017, ISSN 2154-2287. http://ieeexplore.ieee.org/document/7831538/, acesso em 2017-04-22. 2

[8] Bardak, B. e M. Tan: *Prediction of influenza outbreaks by integrating Wikipedia article access logs and Google flu trend data*. IEEE 15th International Conference on Bioinformatics and Bioengineering - BIBE, páginas 1–6, 2015. https://ieeexplore.ieee.org/document/7367640. 2

[9] Rodríguez, Emilio: *El proceso de toma de decisiones estratégicas en las universidades públicas*. Calidad en la Educación, (24):47–63, 2006. https://calidadenlaeducacion.cl/index.php/rce/article/view/267/271. 2

[10] Carvalho, Gilson: *A saúde pública no brasil*. Estudos avançados, 27(78):7–26, 2013. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40142013000200002. 2

[11] Laguardia, Josué at al.: *Sistema de informação de agravos de notificação em saúde (sinan): desafios no desenvolvimento de um sistema de informação em saúde*. Epidemiologia e Serviços de Saúde, 13(3):135–146, 2004. http://scielo.iec.pa.gov.br/scielo.php?pid=S1679-49742004000300002&script=sci_arttext&tlng=pt, acesso em 2017-04-22. 3

[12] Bonamigo, Elcio Luiz e Guilherme Afonso Fabiani Campos Soares: *Sub-notificação de doenças de notificação compulsória: Aspectos Éticos, jurídicos e social*. Anais de Medicina, 2015. http://editora.unoesc.edu.br/index.php/anaisdemedicina/article/view/9435, acesso em 2017-06-15. 3

[13] Barbosa, Daniely Aleixo e Andréa Maria Ferreira Barbosa: *Avaliação da completitude e consistência do banco de dados das hepatites virais no estado de pernambuco, brasil, no período de 2007 a 2010*. Epidemiologia e Serviços de Saúde, 22(1):49–58, 2013, ISSN 1679-4974. http://scielo.iec.pa.gov.br/scielo.php?script=sci_abstract&pid=S1679-49742013000100005&lng=pt&nrm=iso&tlng=en, acesso em 2017-04-22. 3

[14] Jesus, Ronaldo de at al.: *Sistema gerenciador de ambiente laboratorial: relato de experiência de uma ferramenta transformadora para a gestão laboratorial e vigilância em saúde*. Epidemiologia e Serviços de Saúde, 22(3):525–529, 2013, ISSN 1679-4974. http://scielo.iec.pa.gov.br/scielo.php?script=sci_abstract&pid=S1679-49742013000300018&lng=pt&nrm=iso&tlng=pt, acesso em 2017-04-15. 4, 18, 19, 54, 55

[15] Saúde, Ministério da: *Informe epidemiológico, ministério da saúde*, 2017. http://portalarquivos.saude.gov.br/images/pdf/2017/novembro/07/Informe-Epidemiol--gico-Influenza-2017-SE-41.pdf, acesso em 2017-10-17. 4, 5

[16] Saúde, Ministério da: *Guia de vigilância da influenza, ministério da saúde*, 2016. http://bvsms.saude.gov.br/bvs/publicacoes/guia_laboratorial_influenza_vigilancia_influenza_brasil.pdf, acesso em 2017-10-17. 4, 5

[17] Silva, Edna Lúcia da e Estera Muszkat Menezes: *Metodologia da pesquisa e elaboração de dissertação*. UFSC, 4a edição, 2001. 7

[18] Jesus, Ronaldo de: *Desenvolvida pelo autor*, 2018. 8, 58, 59, 60, 61, 62, 63, 66, 70, 71, 72, 73, 79, 81, 82, 83, 85, 87, 89, 129, 130, 131, 132, 133, 134, 135, 136, 137

[19] Veloso, Luíza Tuler: *Modelos de séries temporais e gráficos de controle estatístico aplicados a indicadores de vigilância epidemiológica do ministério da saúde*. Relatório técnico de graduação, Departamento de Estatística, Universidade de Brasília-UnB, Brasília-DF, 2018. 11, 34, 35, 67, 75, 91

[20] Paula Baltar, Bruno de: *Análise Temporal dos Preços da Commodity Cobre Usando o Modelo Box & Jenkins*. Tese de Doutoramento, PUC-Rio, 2009. 11

[21] C Douglas, Montgomery: *Introdução ao Controle Estatístico da Qualidade*, volume Único. LTC, 4ª edição, 2009. 12, 76, 77

[22] Choi, Bernard C. K.: *The past, present, and future of public health surveillance*. Scientifica: National Library of Medicine National Institutes of Health, 2012, 2012, ISSN 2090-908X. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3820481/, acesso em 2017-04-29. 14, 15, 18, 19

[23] Faria, Liliam Saldanha e Maria Rita Bertolozzi: *Theoretical assumptions regarding health surveillance: a prospect for its integration*. Acta Paulista de Enfermagem, 22(4):422–427, 2009, ISSN 0103-2100. http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0103-21002009000400012&lng=en&nrm=iso&tlng=pt, acesso em 2017-04-29. 15

[24] Santiago, Alynne da Costa at al.: *Health surveillance based on social and health indicators*. Revista da Escola de Enfermagem da USP, 42(4):798–803, 2008. http://www.scielo.br/scielo.php?pid=S0080-62342008000400025&script=sci_arttext, acesso em 2017-04-28. 15

[25] Velasco, Edward at al.: *Social media and internet-based data in global systems for public health surveillance: A systematic review*. The Milbank Quarterly, 92(1):7–33, 2014, ISSN 0887-378X. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3955375/, acesso em 2017-04-30. 19, 20

[26] Minayo, Maria Cecília de Souza: *Construção de indicadores qualitativos para avaliação de mudanças*. Revista Brasileira de Educação Médica, 33:83–91, 2009, ISSN 0100-5502. http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0100-55022009000500009&lng=en&nrm=iso&tlng=pt, acesso em 2017-04-28. 20, 21

[27] Trevisan, Andrei Pittol e Hans Michael Van Bellen: *Avaliação de políticas públicas: uma revisão teórica de um campo em construção*. Revista de Administração Pública, 42(3):529–550, 2008. http://www.academia.edu/download/34919548/a05v42n3.pdf, acesso em 2017-04-29. 20

[28] Tronchin, D. M. at al.: *Theoretical background for the construction and establishment of quality indicators in health*. Revista gaucha de enfermagem - EENFUFRGS, 30(3):542–546, 2009. http://europepmc.org/abstract/med/20187437, acesso em 2017-05-03. 21, 22, 23

[29] Agostinho, Feni, Enrique Ortega e Ademar Romeiro: *Índices versus indicadores: precisões conceituais na discussão da sustentabilidade de países*. Ambiente & sociedade, 10(2):137–148, 2007. http://www.scielo.br/pdf/asoc/v10n2/a09v10n2, acesso em 2017-05-05. 21

[30] Merchán-Hamann, Edgar, Pedro Luiz Tauil e Marisa Pacini Costa: *Terminologia das medidas e indicadores em epidemiologia: subsídios para uma possível padronização da nomenclatura*. Informe Epidemiológico do Sus, 9(4):276–284, 2000, ISSN 0104-1673. http://scielo.iec.pa.gov.br/scielo.php?script=sci_abstract&pid=S0104-16732000000400006&lng=pt&nrm=iso&tlng=pt, acesso em 2017-05-03. 23

[31] Souza MinayoI, Maria Cecília de: *Construção de indicadores qualitativos para avaliação de mudanças*. Revista Brasileira de Educação Médica, 33(1):83–91, 2009. https://www.researchgate.net/profile/Maria_Minayo/publication/262588214_The_construction_of_qualitative_indicators_for_the_evaluation_of_changes/links/0c96052b2e49cf2008000000.pdf, acesso em 2017-07-23. 23

[32] Vigilância SP, Secretária de: *Conceitos e definições em epidemiologia importantes para vigilância sanitária - sp*, 2004. http://www.cvs.saude.sp.gov.br/pdf/epid_visa.pdf, acesso em 2004-03-01. 23, 24

[33] Saúde, Escritório Regional para as Américas da Organização Mundial da Saúde Organização Pan-Americana da: *Indicadores básicos para a saúde no brasil: conceitos e aplicações*, 2008, ISBN 978-85-87943-65-1. 23, 24

[34] Hernández-Julio, Yamid Fabián at al.: *Framework for the development of business intelligence using computational intelligence and service-oriented architecture*. páginas 1–7. IEEE-12th Iberian Conference on Information Systems and Technologies (CISTI), 2017. http://ieeexplore.ieee.org/abstract/document/7975758/, acesso em 2017-07-28. 26

[35] Magaireah, Asma I., Hidayah Sulaiman e Nor'ashikin Ali: *Theoretical framework of critical success factors (CSFs) for Business Intelligence (BI) System*. páginas 455–463. IEEE - 8th International Conference on Information Technology (ICIT), 2017. 26

[36] Fernández, M., A. Dávila e P. Angeleri: *Data quality applied to an academic business intelligence solution: Lesson learned*. páginas 1–6. IEEE-Colombian Conference on Communications and Computing (COLCOM), agosto 2017. 26

[37] Rodzi, Nur Alia Hamizah Mohamad, Mohd Shahizan Othman e Lizawati Mi Yusuf: *Significance of data integration and ETL in business intelligence framework for higher education*. páginas 181–186. IEEE-International Conference on Science in Information Technology (ICSITech), 2015. 27

[38] Italiano, Isabel Cristina e Joao Eduardo Ferreira: *A hybrid model for data synchronism in data warehouse projects*. páginas 12–21. IEEE - Seventh International

Database Engineering and Applications Symposium, 2003. Proceedings., 2003. 28, 29

[39] Ali, Syed Mohd at al.: *Big data visualization: Tools and challenges*. páginas 656–660. IEEE -2nd International Conference on Contemporary Computing and Informatics (IC3I), 2016. 30

[40] Chang, B. R., Y. A. Wang, Y. D. Lee e C. F. Huang: *Development of multiple big data analysis platforms for business intelligence*. páginas 1930–1933. International Conference on Applied System Innovation (ICASI), 2017. 31

[41] Elastic: *Elasticsearch, logstash, and kibana*, 2018. `https://www.elastic.co/products/elasticsearch`, acesso em 2018-09-19. 31, 32, 33

[42] Elastic: *Arquitetura elasticsearch, logstash, and kibana - elk*, 2018. `https://www.elastic.co/elk-stack`, acesso em 2018-09-19. 32, 34

[43] Antunes, José Leopoldo Ferreira e Maria Regina Alves Cardoso: *Uso da análise de séries temporais em estudos epidemiológicos*. Epidemiologia e Serviços de Saúde, 24(3):565–576, 2015, ISSN 1679-4974. `http://www.iec.pa.gov.br/template_doi_ess.php?doi=10.5123/S1679-49742015000300024&scielo=S2237-96222015000300565`, acesso em 2017-06-21. 34, 36

[44] Lopes, Lívia Rachel Sant' Anna Monteiro Rocha: *Gráficos de controle estatístico de qualidade para indicador estratégico da secretaria da fazenda do governo do piauí*. Tese de Mestrado, Departamento de Estatística, UnB, Brasília-DF, 2014. `http://bdm.unb.br/handle/10483/8147`, acesso em 2017-06-27. 34, 35, 36, 37, 38, 39, 43

[45] Latorre, Maria do Rosário Dias de at al.: *Time series analysis in epidemiology: an introduction to methodological aspects*. Revista Brasileira de Epidemiologia, 4(3):145–152, 2001. `http://www.scielosp.org/scielo.php?pid=S1415-790X2001000300002&script=sci_arttext`, acesso em 2017-06-21. 34, 35, 36, 37, 41

[46] Claro, Fernando Antonio Elias, Antonio Fernando Branco Costa e Marcela Aparecida Guerreiro Machado: *EWMA and x-bar control charts for the monitoring of autocorrelated processes*. Production - UNESP - Guaratinguetá, 17(3):536–546, 2007, ISSN 0103-6513. `http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0103-65132007000300010&lng=en&nrm=iso&tlng=pt`, acesso em 2017-06-28. 37

[47] Sato, Renato Cesar: *Disease management with ARIMA model in time series*. Einstein (Sao Paulo), 11(1):128–131, 2013. `http://www.scielo.br/scielo.php?pid=S1679-45082013000100024&script=sci_arttext`, acesso em 2017-06-26. 39, 40, 41

[48] Werner, Liane e José Luis Duarte Ribeiro: *Previsão de demanda: uma aplicação dos modelos box-jenkins na área de assistência técnica de computadores pessoais*. Gestão e produção. São Carlos, SP. Vol. 10, no. 1 (abr. 2003), p. 47-67, 2003. `http://www.scielo.br/pdf/gp/v10n1/a05v10n1`, acesso em 2017-07-21. 41, 76

[49] Silva, Maria I. S., Ednaldo C. Guimarães e Marcelo Tavares: *Forecast of monthly mean temperatures in uberlândia, minas gerais, brazil using time series models*. Revista Brasileira de Engenharia Agrícola e Ambiental, 12(5):480–485, 2008, ISSN 1415-4366. http://www.scielo.br/scielo.php?script=sci_abstract&pid=S1415-43662008000500006&lng=en&nrm=iso&tlng=pt, acesso em 2017-07-08. 42

[50] Detmann, Edenio e Filho at al.: *Use of regression techniques in the evaluation, in beef cattle, of feed conversion into product: comparison between experimental groups*. Revista Brasileira de Zootecnia, 41(1):138–146, 2012, ISSN 1516-3598. http://www.scielo.br/scielo.php?script=sci_abstract&pid=S1516-35982012000100021&lng=en&nrm=iso&tlng=pt, acesso em 2017-07-08. 42

[51] Chechi, Leonardo e Fábio M. Bayer: *Modelos univariados de séries temporais para previsão das temperaturas médias mensais de erechim, RS*. Revista Brasileira de Engenharia Agricola e Ambiental-Agriambi, 16(12), 2012. http://www.agriambi.com.br/revista/v16n12/v16n12a09.pdf, acesso em 2017-07-09. 43

[52] Thor, Johan at al.: *Application of statistical process control in healthcare improvement: systematic review*. Quality & Safety in Health Care, 16(5):387–399, 2007, ISSN 1475-3898. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2464970/. 43, 44

[53] Henning, Elisa at al.: *Aplicação de gráficos de controle estatístico de processos para o monitoramento dos casos de meningite no município de joinville*. Produção em Foco, 2(1):1–26, 2012. http://qualimetria.ufsc.br/files/2013/04/63-235-1-pb.pdf, acesso em 2017-06-14. 44, 45

[54] Aslam, Muhammad at al.: *Mixed control charts using EWMA statistics*. IEEE Access, 4:8286–8293, 2016, ISSN 2169-3536. http://ieeexplore.ieee.org/document/7744529/, acesso em 2017-06-18. 44, 45

[55] Nidsunkid, S., J. J. Borkowski e K. Budsaba: *The effects of violations of assumptions in multivariate shewhart control charts*. páginas 214–218. Industrial Engineering and Engineering Management (IEEM), 2016 IEEE International Conference on, 2016. http://ieeexplore.ieee.org/abstract/document/7797867/, acesso em 2017-06-19. 45

[56] Nomelini, Quintiliano Siqueira Schroden, Eric Batista Ferreira e Marcelo Silva de Oliveira: *Studies on non-random patterns in shewhart control charts*. Gestão & Produção, 16(3):414–421, 2009, ISSN 0104-530X. http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0104-530X2009000300008&lng=en&nrm=iso&tlng=pt, acesso em 2017-06-19. 45, 46

[57] Omar, M. Hafidz: *Statistical process control charts for measuring and monitoring temporal consistency of ratings*. Journal of Educational Measurement, 47(1):18–35, 2010. http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2009.00097.x/full, acesso em 2017-06-19. 46, 47

[58] Walter, Olga Maria Formigoni Carvalho: *Individual and combined application of CUSUM and shewhart control charts: an application in the metalworking sector*. Gestão & Produção, 20(2):271–286, 2013, ISSN 0104-530X. http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0104-530X2013000200003&lng=en&nrm=iso&tlng=pt, acesso em 2017-06-20. 47

[59] Steiner, Stefan H at al.: *Detecting the start of an influenza outbreak using exponentially weighted moving average charts*. BMC Med Inform Decis Mak, 10, 2010, ISSN 1472-6947. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2909986/. 47

[60] Walter, OMFC at al.: *Aplicação individual e combinada dos gráficos de controle shewhart e CUSUM: uma aplicação no setor metal mecânico*. Gestão e Produção, São Carlos, 20(2):271–286, 2013. https://www.researchgate.net/profile/Robert_Samohyl/publication/282188602_Individual_and_combined_application_of_CUSUM_and_Shewhart_control_charts_An_application_in_the_metalworking_sector/links/570d037908aec783ddcda6b6.pdf, acesso em 2017-07-10. 48, 49

[61] Orssatto, Fábio, Marcio Vilas Boas e Eduardo Eyng: *Gráfico de controle da média móvel exponencialmente ponderada: aplicação na operação e monitoramento de uma estação de tratamento de esgoto*. Engenharia Sanitaria e Ambiental, 20(4):543–550, 2015, ISSN 1413-4152. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-41522015000400543&lng=pt&tlng=pt, acesso em 2017-06-20. 49

[62] Saúde, Ministério da: *Portal da saúde - www.saude.gov.br - histórico*. http://portalsaude.saude.gov.br/index.php/o-ministerio/historico, acesso em 2017-06-11. 51, 52, 53

[63] Santos, Ana Rosa dos: *A rede laboratorial de Saúde Pública e o SUS*. Informe Epidemiológico do Sus, 6(2):7–14, junho 1997, ISSN 0104-1673. http://scielo.iec.pa.gov.br/scielo.php?script=sci_abstract&pid=S0104-16731997000200002&lng=pt&nrm=iso&tlng=pt, acesso em 2017-11-15. 54

[64] SESA-PR: *Informes epidemiológicos - cievs, paraná*, 2018. http://www.saude.pr.gov.br/modules/conteudo/conteudo.php?conteudo=2811, acesso em 2018-06-19. 78

# Anexo I

# Script R dos Dados de Influenza

```
#########################################################################
RScript (Extração, Tratamento e Carga - Dados Influenza, Curitiba-PR)
#-------------------------------------------------------------------------
###SÉRIE TEMPORAL Previsão Table ####
library("lubridate")#munipular datas
library("RColorBrewer")
library("plyr") #Manipulação de DF
library("ggplot2") #Gráfico
library("RColorBrewer") #Paleta de cores
library("forcats") #Series Temporais
library("tseries") #Teste de Dick Fuller
library("TSA") #Periodograma
library("qcc") #Para gráficos de controle
library("WriteXLS")#Alteranartiva - Exportar dados para excel
library("RPostgreSQL")#Driver de conexão com banco de dados PostgresSQL
library("astsa")#Prevision Sarima
library("sarima")#Modelagem Sarima
library("reshape2")#CrossTable Summary
#-------------------------------------------------------------------------
pw <- {
  "new_user_password"
}
drv <- dbDriver("PostgreSQL")
con <- dbConnect(drv, dbname = "django_hl7brasil",
                 host = "localhost", port = 5432,
                 user = "postgres", password = '******')
```

```
rm(pw) # removes the password
dbExistsTable(con, "ve_influenza_brasil")
#-------------------------------------------------------------------------
ve_influenza_cwb <- dbGetQuery(con, "
                               select
                               distinct
                               digest(cast((paciente,data_nasc,mun_resid)
                               as text),'sha256') as Chave
                               , pais_resid as PaisResid
                               , uf_resid as UFResid
                               , mun_resid as MunResid
                               , ibge_mun_resid as IBGEMun
                               , data_coleta as DataColeta
                               , cast (ano as numeric(4,0)) as Ano
                               , cast (mes as numeric(2,0)) as Mes
                               , cast (semana_epi as numeric(2,0))
                               as SemanaEpid, psng_influenza
                               as ResultadoExame
                               from
                               ve_influenza_cwb_gal_2010a11_2018_limpo
                               where mun_resid = 'CURITIBA'
                               and ano between 2009 and 2019 and
                               psng_influenza = '1-Positivo'
                               ")
#-------------------------------------------------------------------------
semana_epi <- dbGetQuery(con, "
                         select
                         semanaepid
                         , inicio
                         , termino
                         , ano
                         , data_media
                         , concat(ano,'-', to_char(semanaepid, 'fm00'))
                         as ano_semanaepi
                         from
                         ve_semana_epi
                         ")
#-------------------------------------------------------------------------
ve_influenza_cwb<-as.data.frame(ve_influenza_cwb)
colClasses = c("character", rep("factor",5),"character",
```

```
            "character",rep("numeric",3), rep("factor",5))
names(ve_influenza_cwb)<-c("Chave","PaisResid","UFResid","MunResid",
    "IBGEMun", "DataColeta","Ano","Mes","SemanaEpid","ResultadoExame")
setClass('mDate')
setAs("character","mDate", function(from) as.Date(from,
format="%d/%m/%Y") )
ve_influenza_cwb$DataColeta<-as.Date(ve_influenza_cwb$DataColeta,
"%d-%m-%Y")
#Semanas Epidemiológicas
colClasses = c("numeric", rep("mDate",2),"numeric","mDate")
names(semana_epi)<-c("SemanaEpid","Inicio","Termino","Ano",
"Data_media","Ano_SemanaEpi")
summary(ve_influenza_cwb)
nrow(ve_influenza_cwb)
nrow(semana_epi)
str(ve_influenza_cwb)
str(semana_epi)
head(semana_epi)
head(ve_influenza_cwb)
head(as.factor(ve_influenza_cwb$Ano))
#-------------------------------------------------------------------------
dados_cwb<-merge(ve_influenza_cwb,semana_epi,by.x=c('Ano','SemanaEpid'),
by.y=c('Ano','SemanaEpid'))
head(dados_cwb)
#-------------------------------------------------------------------------
library(dplyr)
influenza_cwb<-dados_cwb%>%
  select(Ano,SemanaEpid,UFResid,MunResid,IBGEMun,
  DataColeta,Inicio,Termino,Data_media,Ano_SemanaEpi,ResultadoExame)%>%
  filter(ResultadoExame=="1-Positivo" & Ano>'2009' & Ano<'2019'
  & UFResid!='') & SemanaEpid<=53 & Termino<=data)
#View(influenza_cwb)
#-------------------------------------------------------------------------
cwb_all<-influenza_cwb%>%
  select(Ano:ResultadoExame)%>%
  group_by(UFResid, MunResid, Ano, SemanaEpid, Inicio, Termino,Data_media,
  Ano_SemanaEpi)%>%
  summarise(indicador = n())%>%
  filter(UFResid!='' & SemanaEpid<=53)
##View(cwb_all)
```

```
#-----------------------------------------------------------------------
library(reshape2)
indicador_ano<-(dcast(cwb_all, SemanaEpid~Ano, value.var='indicador', sum))
#View(indicador_ano)
#-----------------------------------------------------------------------
# Data Início Monitoramento da Previsão do Ano de 2018
data<-(today()-30)
mysemana<-semana_epi%>%
  filter(data>=Inicio & data<=Termino)
mysemana
mydata<-indicador_ano%>%
  select(SemanaEpid,'2018','2017')
mydata$`2017`<-2018
names(mydata)<-c("SemanaEpid","Indicador","Ano")
mydata_s<-merge(mydata,semana_epi,by.x=c('Ano','SemanaEpid'),
by.y=c('Ano','SemanaEpid'))
mydata_s<-mydata_s[order(mydata_s$Data_media),]
st_mydata_2018<-mydata_s%>%
  filter(SemanaEpid<=mysemana$SemanaEpid)
#-----------------------------------------------------------------------
  # Indicador_2018<-dados_cwb%>%
  #   select(Ano:Ano_SemanaEpi)%>%
  #   group_by(UFResid, MunResid, Ano, SemanaEpid, Inicio,
  Termino, Data_media, Ano_SemanaEpi)%>%
  #   summarise(indicador = n())%>%
  #   filter(Ano=='2018' & SemanaEpid<=mysemana$SemanaEpid)
  # head(Indicador_2018)
  # library(reshape2)
  # comp<-c(1:mysemana$SemanaEpid)
  # Indicador_2018_cross<-(dcast(Indicador_2018,SemanaEpid,
  no,value.var='indicador', sum))
  # View(Indicador_2018_cross)
#-----------------------------------------------------------------------
  # Indicador_2018_prev<-dados_cwb%>%
  #   select(Ano:Ano_SemanaEpi)%>%
  #   group_by(UFResid, MunResid, Ano, SemanaEpid, Inicio,
  Termino, Data_media, Ano_SemanaEpi)%>%
  #   summarise(indicador = n())%>%
  #   filter(Ano=='2018' & SemanaEpid>mysemana$SemanaEpid)
  # View(Indicador_2018_prev)
```

```
#-------------------------------------------------------------------------
PR_ST<-c(indicador_ano$‘2010‘,indicador_ano$‘2011‘,indicador_ano$‘2012‘,
         indicador_ano$‘2013‘,
         indicador_ano$‘2014‘,indicador_ano$‘2015‘,indicador_ano$‘2016‘,
         indicador_ano$‘2017‘,st_mydata_2018$Indicador)
cwb_st<-PR_ST
length(PR_ST)
#-------------------------------------------------------------------------
ggplot(cwb_all, aes(x=Data_media ,y =indicador)) +
  geom_line(size = 0.4)+geom_point(size = 0.7, shape = 22,colour="black",
  fill ="Dark gray") + xlab("Tempo (em Semanas epidemiológicas)")
  + ylab("Indicador")
#-------------------------------------------------------------------------
summary(PR_ST)
sd(PR_ST,na.rm =TRUE) #5.168
sd(PR_ST,na.rm =TRUE)/mean(PR_ST,na.rm =TRUE) #1.931
#-------------------------------------------------------------------------
ddply(cwb_all, .(Ano), summarize,  Minimo=min(indicador,na.rm =TRUE),
Maximo=max(indicador,na.rm =TRUE),
Desvio_Padrao=sd(indicador,na.rm =TRUE),
Media=mean(indicador,na.rm =TRUE), Total_pos=sum(indicador))
#----------------------Séries Temporais---------------------------------
AICC <- function(model){
  n <- model$nobs
  p <- length(model$coef)
  AICC <- model$aic + 2*p*(p+1)/(n-p-1)
  return(AICC)
}
#-------------------------------------------------------------------------
ppp<-periodogram(PR_ST,log='no',plot=TRUE,ylab="Periodogram",
xlab="Frequency",lwd=2)
head(ppp$spec,n=8L)
head(ppp$freq,n=8L)
1/ppp$freq[8]
#-------------------------------------------------------------------------
acf(PR_ST,
    lag.max = 250,
    ylab = "FAC",
    main = "FAC amostral")
#-------------------------------------------------------------------------
```

```
acf(PR_ST,
    lag.max = 250,
    ylab = "FACP",
    main = "FACP amostral", type = "partial")
#-------------------------------------------------------------------------
dif.i<-diff(PR_ST, lag = 52, differences = 1)
plot(dif.i,type="l", main="série do Indicador Diferenciada", ylab='',
xlab="")
mean(dif.i)
#-------------------------------------------------------------------------
acf(dif.i,
    lag.max = 250,
    ylab = "FAC",
    main = "FAC amostral")
acf(dif.i,
    lag.max = 250,
    ylab = "FACP",
    main = "FACP amostral", type = "partial")
#-------------------------------------------------------------------------
# n1<-arima(dif.i, order = c(1, 0, 1),
#           seasonal = list(order = c(1, 0, 1), period = 52),
#           include.mean = TRUE)
# aa1<-AICC(n1)
# aa1
# bb1<-BIC(n1)
# bb1
# #---------------------------------
# n2<-arima(dif.i, order = c(1, 0, 1),
#           seasonal = list(order = c(0, 0, 1), period = 52),
#           include.mean = TRUE)
# aa2<-AICC(n2)
# aa2
# bb2<-BIC(n2)
# bb2
# #---------------------------------
# n3<-arima(dif.i, order = c(1, 0, 1),
#           seasonal = list(order = c(1, 0, 0), period = 52),
#           include.mean = TRUE)
# aa3<-AICC(n3)
# aa3
```

```
# bb3<-BIC(n3)
# bb3
# #---------------------------------
# n4<-arima(dif.i, order = c(1, 0, 0),
#           seasonal = list(order = c(1, 0, 1), period = 52),
#           include.mean = TRUE)
# aa4<-AICC(n4)
# aa4
# bb4<-BIC(n4)
# bb4
#---------------------------------
n5<-arima(dif.i, order = c(1, 0, 0),
          seasonal = list(order = c(0, 0, 1), period = 52),
          include.mean = TRUE)
aa5<-AICC(n5)
aa5
bb5<-BIC(n5)
bb5
#---------------------------------
# n6<-arima(dif.i, order = c(1, 0, 0),
#           seasonal = list(order = c(1, 0, 0), period = 52),
#           include.mean = TRUE)
# aa6<-AICC(n6)
# aa6
# bb6<-BIC(n6)
# bb6
# #---------------------------------
# n7<-arima(dif.i, order = c(0, 0, 1),
#           seasonal = list(order = c(1, 0, 1), period = 52),
#           include.mean = TRUE)
# aa7<-AICC(n7)
# aa7
# bb7<-BIC(n7)
# bb7
# #---------------------------------
# n8<-arima(dif.i, order = c(0, 0, 1),
#           seasonal = list(order = c(0, 0, 1), period = 52),
#           include.mean = TRUE)
# aa8<-AICC(n8)
# aa8
```

```
# bb8<-BIC(n8)
# bb8
# #---------------------------------
# n9<-arima(dif.i, order = c(0, 0, 1),
#           seasonal = list(order = c(1, 0, 0), period = 52),
#           include.mean = TRUE)
# aa9<-AICC(n9)
# aa9
# bb9<-BIC(n9)
# bb9
# #---------------------------------
# n10<-arima(dif.i, order = c(0, 0, 0),
#            seasonal = list(order = c(1, 0, 1), period = 52),
#            include.mean = TRUE)
# aa10<-AICC(n10)
# aa10
# bb10<-BIC(n10)
# bb10
# #---------------------------------
# n11<-arima(dif.i, order = c(0, 0, 0),
#            seasonal = list(order = c(0, 0, 1), period = 52),
#            include.mean = TRUE)
# aa11<-AICC(n11)
# aa11
# bb11<-BIC(n11)
# bb11
# #---------------------------------
# n12<-arima(dif.i, order = c(0, 0, 0),
#            seasonal = list(order = c(1, 0, 0), period = 52),
#            include.mean = TRUE)
# aa12<-AICC(n12)
# aa12
# bb12<-BIC(n12)
# bb12
# #---------------------------------
# n13<-arima(dif.i, order = c(1,0,1),
#            seasonal = list(order = c(0, 0, 0), period = 52),
#            include.mean = TRUE)
# aa13<-AICC(n13)
# aa13
```

```
# bb13<-BIC(n13)
# bb13
# #---------------------------------
# n14<-arima(dif.i, order = c(1, 0, 0),
#            seasonal = list(order = c(0, 0, 0), period = 52),
#            include.mean = TRUE)
# aa14<-AICC(n14)
# aa14
# bb14<-BIC(n14)
# bb14
# #--------------
# n15<-arima(dif.i, order = c(0, 0, 1),
#            seasonal = list(order = c(0, 0, 0), period = 52),
#            include.mean = TRUE)
# aa15<-AICC(n15)
# aa15
# bb15<-BIC(n15)
# bb15
# #--------------
# n16<-arima(dif.i, order = c(0, 0, 0),
#            seasonal = list(order = c(0, 0, 0), period = 52),
#            include.mean = TRUE)
# aa16<-AICC(n16)
# aa16
# bb16<-BIC(n16)
# bb16
#-------------------------------------------------------------------------
# l1<-arima(dif.i, order = c(0, 0, 2),
#           seasonal = list(order = c(0, 0, 4), period = 52),
#           include.mean = TRUE)
# al1<-AICC(l1)
# al1
# bl1<-BIC(l1)
# bl1
#-------------------------------------------------------------------------
# AICC_.<-cbind(aa1,aa2,aa3,aa4,aa5,aa6,aa7,aa8,aa9,aa10,aa11,aa12,aa13,
aa14,aa15,aa16,al1)
# order(AICC_.)
# plot(AICC_.[order(AICC_.)],type="p")
# BIC_.<-cbind(bb1,bb2,bb3,bb4,bb5,bb6,bb7,bb8,bb9,bb10,bb11,bb12,bb13,
```

```
bb14,bb15,bb16,bl1)
# order(BIC_.)
# plot(BIC_.[order(BIC_.)],type="p")
#-------------------------------------------------------------------------
#"Melhores" modelos segundo AICC e BIC:
##SARIMA(1,0,0)x(0,0,1)_53 --> n5 (2 param)
##SARIMA(1,0,0)x(1,1,1)_53 --> n4 (3 param)
##SARIMA(1,0,1)x(0,1,1)_53 --> n2 (3 param)
##SARIMA(1,0,1)x(1,1,1)_53 --> n1 (4 param)
#-------------------------------------------------------------------------
res5<-n5$residuals
plot(res5,type="l")
length(res5)
#-------------------------------------------------------------------------
acf(res5,
    lag.max = 250,
    ylab = "FAC",
    main = "FAC amostral")
#-------------------------------------------------------------------------
Box.test(res5,lag=150 ,fitdf=2,type="Ljung-Box")
#-------------------------------------------------------------------------
ajust5<-dif.i-res5
plot(PR_ST,type="p",cex=.6)
lines(ajust5,type="l",col="red")
mean(res5)
#-------------------------------------------------------------------------
length(res5)
#Anos:----2011(52)----2012(52)------2013(52)-------2014(53)------2015(52)
------2016(52)------2017(52)------2018(52)
####res5<-c(res5[1:52],res5[54:105],res5[107:158],res5[160:212],
res5[214:265],res5[267:318],res5[320:371],res5[373:424])
#-------------------------------------------------------------------------
# Previsão da Série para o próximo ano, neste caso ano de 2018
########## Previsão "A"###########
prev_a<-sarima.for(cwb_st, 10, 1, 0, 1, P = 0, D = 1, Q = 1, S = 52,
                   tol = sqrt(.Machine$double.eps), no.constant = FALSE,
                   plot.all=TRUE, xreg = NULL, newxreg = NULL)
########## Previsão "B"###########
nummy   = length(res5)
n.ahead = 10
```

```
nureg   = time(cwb_st)[nummy] + seq(1,n.ahead)/52
prev_b<-sarima.for(cwb_st,n.ahead,1,0,1,0,1,1,52, xreg=time(cwb_st),
newxreg=nureg)
########## Previsão "C"###########
prev_s<-sarima.for(cwb_st, 10, 1, 0, 0, P = 0, D = 1, Q = 1, S = 52,
                   tol = sqrt(.Machine$double.eps), no.constant = FALSE,
                   plot.all=TRUE, xreg = NULL, newxreg = NULL)
########## Comparando as Previsões ###########
 prev_a$pred
 prev_b$pred
 prev_s$pred
#-------------------------------------------------------------------------
mydata_p<-mydata_s
 hoje_e<-today()# +30 previsão das semanas
hoje<-semana_epi%>%
  filter(hoje_e>=Inicio & hoje_e<=Termino)
mydate_hoje<-mydata_s%>%
    select(Ano:Ano_SemanaEpi)%>%
    filter(SemanaEpid>mysemana$SemanaEpid & SemanaEpid<hoje$SemanaEpid+1)
prev_mydata_2018<-mydata_s%>%
   filter(SemanaEpid>mysemana$SemanaEpid & SemanaEpid<=hoje$SemanaEpid)
mydata_s<-mydata_s[order(mydata_s$Data_media),]
Previsao<-c(prev_s$pred[1:(NROW(prev_mydata_2018))])
Erro_Previsao<-c(prev_s$se[1:(NROW(prev_mydata_2018))])
Monitor <- cbind(prev_mydata_2018,Previsao,Erro_Previsao,
prev_mydata_2018$Indicador-Previsao)
names(Monitor)<-c("Ano","SemanaEpid","Indicador","Inicio",
"Termino","Data_media","Ano_SemanaEpi","Previsao","Erro_Previsao",
"Indicador_cal")
#View(Monitor)
#-------------------------------------------------------------------------
  # plot(Previsao,type="p",cex=.6)
  # lines(Monitor$Indicador,type="l",col="red")
#-------------------------------------------------------------------------
  # Calculando o Lambda (lambda <- BoxCox.lambda(indicador_graf)
  # library(normalr)
  # lambda<-getLambda(indicador_graf_l, parallel = FALSE)
  # lambda
#-------------------------------------------------------------------------
data_monitor<-c(res5,Monitor$Indicador_cal)
```

```
#-------------------------------------------------------------------------
graf_prev<-ewma(data_monitor, lambda = 0.2, nsigmas = 2.860,center=0,
data.name="Previsão do Modelo", plot=FALSE)
plot(graf_prev, add.stats = TRUE, chart.all = TRUE,
     axes.las = 1, digits = getOption("digits"),
     restore.par = TRUE, label.limits=c("LIC","LSC"),
     title=c("Gráfico MMEP para Resíduo do Modelo2"),
     xlab=c("Semana Epidemiológica"), ylab=c(" "))
#-------------------------------------------------------------------------
# dbExistsTable(con, "ve_qcc_previsao_curitiba")
# sql_command_drop_qcc_surto <- "drop table ve_qcc_previsao_curitiba"
# dbGetQuery(con, sql_command_drop_qcc_surto)
# dbWriteTable(con, "ve_qcc_previsao_curitiba", value = Monitor,
append = TRUE, row.names = FALSE)
#-------------------------------------------------------------------------
# #Descobrindo a Semana Epidemiológica atual
# data<-today()
# data<-'2017-05-05'
# mysemana_prev<-monitor%>%
#   filter(data>=inicio & data<=termino)
# mysemana_indicador<-cwb%>%
#   filter(data>=Inicio & data<=Termino)
# ind_monitor<-mysemana_indicador$indicador- mysemana_prev$previsao
# #mysemana<-c(45,78,56,20)
# res5<-c(res5,monitor$indicador_graf)
# graf3<-ewma(res5, lambda = 0.2, nsigmas = 2.860, plot=FALSE)
# plot(graf3, add.stats = TRUE, chart.all = TRUE,
#      axes.las = 1, digits = getOption("digits"),
#      restore.par = TRUE, label.limits=c("LIC","LSC"),
title=c("Gráfico MMEP para Resíduo do Modelo2"),
#      xlab=c("Semana Epidemiológica"), ylab=c(" "))
#-------------------------------------------------------------------------
#Gráfico de Controle
# lambda=0.05 L=2.492 # Quanto maior o L, maior o peso para as observações
#recentes da série!
graf1<-ewma(data_monitor, lambda = 0.05, nsigmas = 2.492,center=0,
data.name="Resíduo do Modelo", plot=FALSE)
plot(graf1, add.stats = TRUE,axes=FALSE, chart.all = TRUE,axes.las = 1,
     restore.par = FALSE, label.limits=c("LIC","LSC"),
     title=c("Gráfico MMEP para Resíduo do Modelo"),
```

```
     xlab=c("Semana epidemiológica"), ylab=c(" "))
#-------------------------------------------------------------------------
# lambda=0.1 L=2.703 #
graf2<-ewma(data_monitor, lambda = 0.1, nsigmas = 2.703,center=0,
data.name="Resíduo do Modelo", plot=FALSE)
plot(graf2, add.stats = TRUE, chart.all = TRUE,
     axes.las = 1, digits = getOption("digits"),
     restore.par = TRUE, label.limits=c("LIC","LSC"),
     title=c("Gráfico MMEP para Resíduo do Modelo"),
     xlab=c("Semana Epidemiológica"), ylab=c(" "),
     axes=(breaks=c(seq(from=1,to=365,by=52))))
#-------------------------------------------------------------------------
# lambda=0.2 L=2.860 #
graf3<-ewma(data_monitor, lambda = 0.2, nsigmas = 2.860,
center = 0, data.name="Resíduo do Modelo", plot=FALSE)
plot(graf3, add.stats = TRUE, chart.all = TRUE,
     axes.las = 1, digits = getOption("digits"),
     restore.par = TRUE, label.limits=c("LIC","LSC"),
     title=c("Gráfico MMEP para Resíduo do Modelo2"),
     xlab=c("Semana Epidemiológica"), ylab=c(" "))
#-------------------------------------------------------------------------
## Exportando informações
graf3$MunResid<-"Curitiba-PR"
graf3$cod_ibge_mun<-4106902
#-------------------------------------------------------------------------
ve_qcc_curitiba<-cbind(graf3$x,graf3$y,graf3$limits[,1],
graf3$center,graf3$limits[,2],graf3$cod_ibge_mun,graf3$MunResid)
ve_qcc_curitiba<-as.data.frame(ve_qcc_curitiba)
#View(ve_qcc_curitiba)
names(ve_qcc_curitiba)<-c("seq_semana","dados","lic","lc",
"lsc","cod_ibge_mun","mun_resid")
#Incluindo a SemanaEpi nos residuos (Data.média Semana Epidemiológica)
semana_cep<-semana_epi%>%
  filter(Ano>2010 & Ano<=2018 & SemanaEpid<53 &
  Ano_SemanaEpi<=hoje$Ano_SemanaEpi)
seq_semana<-c(1:NROW(ve_qcc_curitiba))
semana_cep<-cbind(semana_cep,seq_semana)
View(semana_cep)
#View(data_monitor)
#View(res5)
```

```
ve_qcc_curitiba<-merge(ve_qcc_curitiba,semana_cep,by=c("seq_semana",
"seq_semana"))
ve_qcc_curitiba<-ve_qcc_curitiba[order(ve_qcc_curitiba$data_media),]
ve_qcc_curitiba$seq_semana<-graf3$x
names(ve_qcc_curitiba)<-c("seq_semana","dados","lic","lc","lsc",
"cod_ibge_mun","
mun_resid","semana_epid","inicio","termino","ano","data_media",
"data_str")
head(ve_qcc_curitiba)
str(ve_qcc_curitiba)
#View(ve_qcc_curitiba)
#------------------------------------------------------------------------
#Tabela de Violação dos Limites de Controles
ve_qcc_violacao_curitiba<-cbind(graf3$violations)
ve_qcc_violacao_curitiba<-as.data.frame(ve_qcc_violacao_curitiba)
names(ve_qcc_violacao_curitiba)<-c("seq_semana")
str(ve_qcc_violacao_curitiba)
ve_qcc_violacao_curitiba<-merge(ve_qcc_violacao_curitiba,
ve_qcc_curitiba,
by=c("seq_semana","seq_semana"))
#------------------------------------------------------------------------
##Escrevendo os Dados no Banco de Dados PostgreSQL
dbExistsTable(con, "ve_qcc_curitiba")
sql_command_drop_qcc <- "drop table ve_qcc_curitiba"
dbGetQuery(con, sql_command_drop_qcc)
dbWriteTable(con, "ve_qcc_curitiba", value = ve_qcc_curitiba,
append = TRUE,
row.names = FALSE)
#------------------------------------------------------------------------
#Tabela de monitoramento e alertas
dbExistsTable(con, "ve_qcc_violacao_curitiba")
sql_command_drop_qcc_surto <- "drop table ve_qcc_violacao_curitiba"
dbGetQuery(con, sql_command_drop_qcc_surto)
dbWriteTable(con, "ve_qcc_violacao_curitiba",
value = ve_qcc_violacao_curitiba,
append = TRUE, row.names = FALSE)
#------------------------------------------------------------------------
########################################################################
######## FIM DO SCRIPT R - INFORMAÇÕES DE INFLUENZA, CURITIBA-PR #########
########################################################################
```

# Anexo II

# Parte dos Processos Realizados no ELK

### II.0.1 Index e Mapeamento dos Dados

```
-------------------------------------------------------------------------
# Criando o Index (qcc_flu_cwb_a)
PUT qcc_flu_cwb_a
{
  "mappings": {
    "qcc_flu_cwb_a": {
      "properties": {
        "ano":    { "type": "integer"  },
        "cod_ibge_mun":     { "type": "text"  },
        "dados":     { "type": "text"  },
        "data_str":     { "type": "text"  },
        "data_media":  {
          "type":   "date",
          "format": "strict_date_optional_time||epoch_millis"
        },
        "inicio":  {
          "type":   "date",
          "format": "strict_date_optional_time||epoch_millis"
        },
        "lc":     { "type": "text"  },
        "lic":     { "type": "text"  },
        "lsc":     { "type": "text"  },
```

```
        "mun_resid":     { "type": "text"  },
        "semana_epid":     { "type": "integer"  },
        "seq_semana":     { "type": "integer"  },
        "termino":  {
          "type":   "date",
          "format": "strict_date_optional_time||epoch_millis"
        }
      }
    }
  }
}
--------------------------------------------------------------------------
# Mapeamento dos Dados de Influenza do Brasil (index: br_influenza)
        {
          "mapping": {
            "lbr_influenza": {
              "properties": {
                "@timestamp": {
                  "type": "date"
                },
                "@version": {
                  "type": "text",
                  "fields": {
                    "keyword": {
                      "type": "keyword",
                      "ignore_above": 256
                    }
                  }
                },
                "data_coleta": {
                  "type": "text",
                  "fields": {
                    "keyword": {
                      "type": "keyword",
                      "ignore_above": 256
                    }
                  }
                },
                "exame": {
                  "type": "text",
```

```
      "fields": {
        "keyword": {
          "type": "keyword",
          "ignore_above": 256
        }
      }
    },
    "metodo": {
      "type": "text",
      "fields": {
        "keyword": {
          "type": "keyword",
          "ignore_above": 256
        }
      }
    },
    "gestante": {
      "type": "text",
      "fields": {
        "keyword": {
          "type": "keyword",
          "ignore_above": 256
        }
      }
    },
    "grupo_agravo": {
      "type": "text",
      "fields": {
        "keyword": {
          "type": "keyword",
          "ignore_above": 256
        }
      }
    },
    "pais": {
      "type": "text",
      "fields": {
        "keyword": {
          "type": "keyword",
          "ignore_above": 256
```

```
    }
  }
},
"regiao": {
  "type": "text",
  "fields": {
    "keyword": {
      "type": "keyword",
      "ignore_above": 256
    }
  }
},
"sexo": {
  "type": "text",
  "fields": {
    "keyword": {
      "type": "keyword",
      "ignore_above": 256
    }
  }
},
"data_coleta": {
  "type": "date"
},
"status": {
  "type": "text",
  "fields": {
    "keyword": {
      "type": "keyword",
      "ignore_above": 256
    }
  }
},
"id_influenza_br": {
  "type": "long"
},
"mes_coleta": {
  "type": "text",
  "fields": {
    "keyword": {
```

```
      "type": "keyword",
      "ignore_above": 256
    }
  }
},
"matbio": {
  "type": "text",
  "fields": {
    "keyword": {
      "type": "keyword",
      "ignore_above": 256
    }
  }
},
"metodo": {
  "type": "text",
  "fields": {
    "keyword": {
      "type": "keyword",
      "ignore_above": 256
    }
  }
},
"municipio": {
  "type": "text",
  "fields": {
    "keyword": {
      "type": "keyword",
      "ignore_above": 256
    }
  }
},
"altitude": {
  "type": "long"
},
"latitute": {
  "type": "float"
},
"longitude": {
  "type": "float"
```

```
},
"psng_influenza": {
  "type": "text",
  "fields": {
    "keyword": {
      "type": "keyword",
      "ignore_above": 256
    }
  }
},
"qtd": {
  "type": "long"
},
"semana_epi": {
  "type": "long"
},
"sg_uf_residencia": {
  "type": "text",
  "fields": {
    "keyword": {
      "type": "keyword",
      "ignore_above": 256
    }
  }
},
"capital": {
  "type": "text",
  "fields": {
    "keyword": {
      "type": "keyword",
      "ignore_above": 256
    }
  }
},
"fronteira": {
  "type": "text",
  "fields": {
    "keyword": {
      "type": "keyword",
      "ignore_above": 256
```

```
                }
              }
            },
            "subtipagem_influenza": {
              "type": "text",
              "fields": {
                "keyword": {
                  "type": "keyword",
                  "ignore_above": 256
                }
              }
            },
            "tipo_influenza": {
              "type": "text",
              "fields": {
                "keyword": {
                  "type": "keyword",
                  "ignore_above": 256
                }
              }
            }
          }
        }
      }

-------------------------------------------------------------------------
# Mapeamento dos Dados: Produção de exames - Brasil (index: exames)
        {
        "mapping": {
          "l_exames": {
            "properties": {
              "@timestamp": {
                "type": "date"
              },
              "@version": {
                "type": "text",
                "fields": {
                  "keyword": {
                    "type": "keyword",
```

```
      "ignore_above": 256
    }
  }
},
"ano": {
  "type": "long"
},
"exame": {
  "type": "text",
  "fields": {
    "keyword": {
      "type": "keyword",
      "ignore_above": 256
    }
  }
},
"metodo": {
  "type": "text",
  "fields": {
    "keyword": {
      "type": "keyword",
      "ignore_above": 256
    }
  }
},
"metodo": {
  "type": "text",
  "fields": {
    "keyword": {
      "type": "keyword",
      "ignore_above": 256
    }
  }
},
"regiao": {
  "type": "text",
  "fields": {
    "keyword": {
      "type": "keyword",
      "ignore_above": 256
```

```
        }
      }
    },
    "agravo": {
      "type": "text",
      "fields": {
        "keyword": {
          "type": "keyword",
          "ignore_above": 256
        }
      }
    },
    "status": {
      "type": "text",
      "fields": {
        "keyword": {
          "type": "keyword",
          "ignore_above": 256
        }
      }
    },
    "id_producao": {
      "type": "long"
    },
    "mes": {
      "type": "long"
    },
    "producao": {
      "type": "long"
    },
    "sg_uf": {
      "type": "text",
      "fields": {
        "keyword": {
          "type": "keyword",
          "ignore_above": 256
        }
      }
    }
}
```

```
            }
          }
        }

-------------------------------------------------------------------------
# Mapeamento dos Dados: Georreferencia-Influenza(index: geopoint_flu_br)
        {
          "mapping": {
            "l_geopoint_flu_br": {
              "properties": {
                "@timestamp": {
                  "type": "date"
                },
                "@version": {
                  "type": "text",
                  "fields": {
                    "keyword": {
                      "type": "keyword",
                      "ignore_above": 256
                    }
                  }
                },
                "id": {
                  "type": "long"
                },
                "location": {
                  "type": "geo_point"
                },
                "municipio": {
                  "type": "text",
                  "fields": {
                    "keyword": {
                      "type": "keyword",
                      "ignore_above": 256
                    }
                  }
                },
                "total": {
                  "type": "long"
                }
```

```
            }
          },
          "_default_": {
            "properties": {
              "location": {
                "type": "geo_point"
              }
            }
          }
        }
      }

-------------------------------------------------------------------------
# Mapeamento dos Dados: Gráfico de Controle (index: qcc_flu_cwb_a)
# Influenza Curitiba-PR
      {
        "mapping": {
          "qcc_flu_cwb_a": {
            "properties": {
              "@timestamp": {
                "type": "date"
              },
              "@version": {
                "type": "text",
                "fields": {
                  "keyword": {
                    "type": "keyword",
                    "ignore_above": 256
                  }
                }
              },
              "cod_ibge_mun": {
                "type": "float"
              },
              "dados": {
                "type": "float"
              },
              "data_media": {
                "type": "date"
              },
```

```
        "data_str": {
          "type": "text",
          "fields": {
            "keyword": {
              "type": "keyword",
              "ignore_above": 256
            }
          }
        },
        "lc": {
          "type": "float"
        },
        "lic": {
          "type": "float"
        },
        "lsc": {
          "type": "float"
        },
        "seq_semana": {
          "type": "float"
        }
      }
    }
  }
}
```

### II.0.2 Parâmetros das configurações dos arquivos de carga - *Logstash*

```
----------------------------
#Arquivo de configuração do Logstash (Carga dos Dados Georreferenciados)
#Geração de Mapa Espacial
        input {
            jdbc {
                jdbc_connection_string => "jdbc:postgresql://localhost:
                5432/django_hl7brasil"
                jdbc_driver_library => "/home/ronaldo/Documentos/linux/
                logstash/postgresql-9.4.1212.jre6.jar"
                jdbc_user => "postgres"
                jdbc_password => "******"
                jdbc_driver_class => "org.postgresql.Driver"
                statement => "select * from ve_geomaps_psng_flu_br"
                schedule => "0 4 * * *"
            }
        }

        filter {
            mutate { convert => {"lat" => "float"} }
            mutate { convert => {"lon" => "float"} }
            mutate { rename => {"lat" => "[location][lat]"} }
            mutate { rename => {"lon" => "[location][lon]"} }
        }
        output {
            elasticsearch {
                      index => "geopoint_flu_br"
                      document_type => "l_geopoint_flu_br"
                      document_id => "%{[id]}"
                      hosts => ["127.0.0.1:9200"]
                }
        }

----------------------------
#Arquivo de configuração do Logstash (Carga dos Dados de Produção de Exames)
         input {
            jdbc {
                jdbc_connection_string => "jdbc:postgresql://localhost:5432/
```

```
            django_hl7brasil"
            jdbc_driver_library => "/home/ronaldo/Documentos/linux/logstash/
            postgresql-9.4.1212.jre6.jar"
            jdbc_user => "postgres"
            jdbc_password => "*******"
            jdbc_driver_class => "org.postgresql.Driver"
            statement => "SELECT id_producao, regiao, uf, agravo, exame,
            metodo, status, ano, mes, producao
            from ve_producao_brasil"
            schedule => "0 3 * * *"
        }
    }
    output {
        elasticsearch {
                index => "exames"
                document_type => "l_exames"
                document_id => "%{[id_producao]}"
                hosts => ["127.0.0.1:9200"]
            }
    }

----------------------------
#Arquivo de configuração do Logstash (Carga dos Dados de Influenza do Brasil)
    input {
        jdbc {
            jdbc_connection_string => "jdbc:postgresql://localhost:5432/
            django_hl7brasil"
            jdbc_driver_library => "/home/ronaldo/Documentos/linux/logstash/
            postgresql-9.4.1212.jre6.jar"
            jdbc_user => "postgres"
            jdbc_password => "*******"
            jdbc_driver_class => "org.postgresql.Driver"
            statement => "SELECT id_influenza_br, grupoagravo,
            exame, metodo, matbio, anocoleta,
            mescoleta, semanaepidemiologica, datacoleta, sexo, gestante,
            municipio, fronteira, capital, longitude,
            latitute, altitude, uf_resd, regiao,
            pais,  status, psng_influenza, tipo_influenza,
            subtipagem_influenza, qtd
            from ve_influenza_brasil"
```

```
                schedule => "0 3 * * *"
            }
        }
        output {
            elasticsearch {
                    index => "br_influenza"
                    document_type => "lbr_influenza"
                    document_id => "%{[id_influenza_br]}"
                    hosts => ["127.0.0.1:9200"]
                }
        }

----------------------------
#Arquivo de configuração do Logstash (Carga dos Dados do Gráfico de Controle)
        input {
            jdbc {
                jdbc_connection_string => "jdbc:postgresql://localhost:5432/
                django_hl7brasil"
                jdbc_driver_library => "/home/ronaldo/Documentos/linux/logstash/
                postgresql-9.4.1212.jre6.jar"
                jdbc_user => "postgres"
                jdbc_password => "*******"
                jdbc_driver_class => "org.postgresql.Driver"
                statement => "select seq_semana, dados, lic, lc, lsc,
                cod_ibge_mun, data_media, data_str
                from ve_qcc_curitiba"
                schedule => "* 5 * * *"
            }
        }
        output {
            elasticsearch {
                    index => "qcc_flu_cwb_a"
                    document_type => "qcc_flu_cwb_a"
                    document_id => "%{[seq_semana]}"
                    hosts => ["127.0.0.1:9200"]
                }
        }
----------------------------
```

### II.0.3 Gráficos, Mapas - Dashboard - ELK

```
--------------------------------
#Gráfico de Controle de Influenza, Curitiba-PR (2011 a 2018)
#Dashboard de Produção de Exames Laboratoriais do Brasil
#Dashboard: Exames de Influenza - Brasil
#Gerado com auxílio do Kibana
```

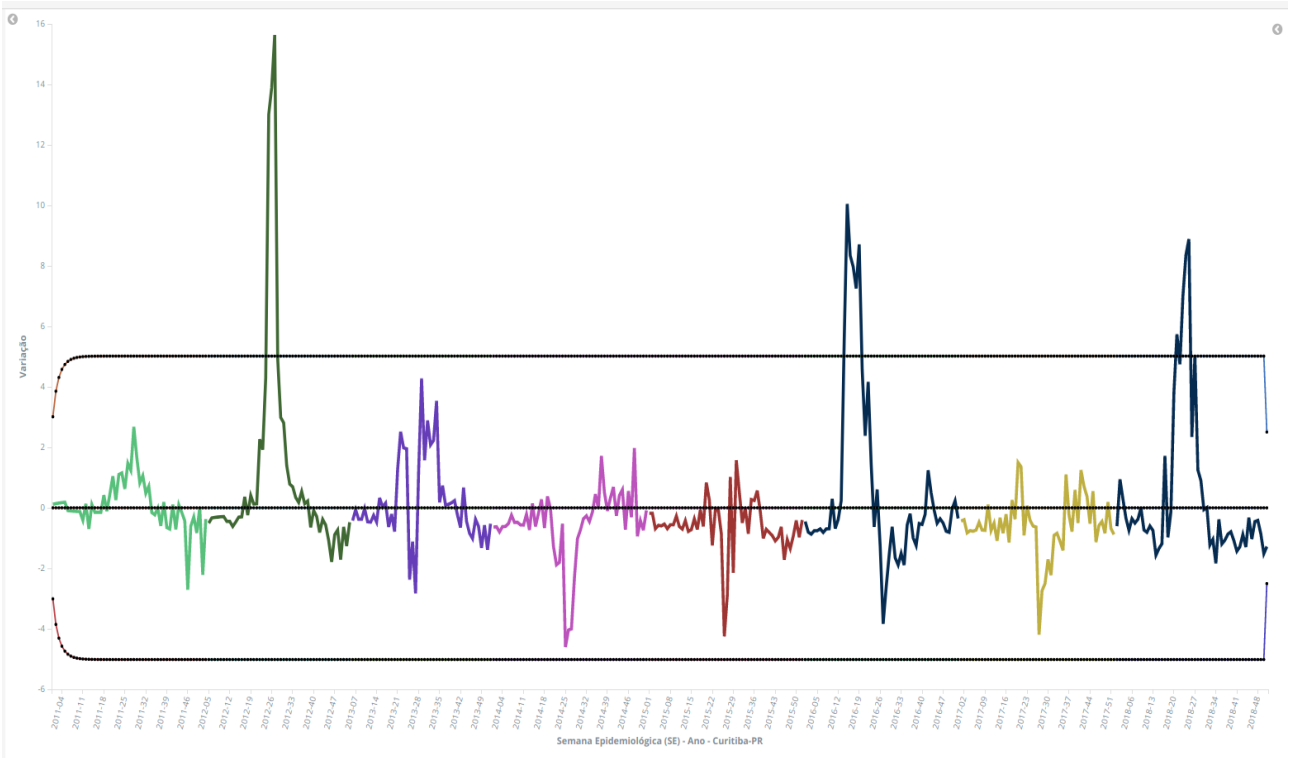



Figura II.1: Kibana: CEP da Influenza (2011 a 2018), Curitiba-PR (Fonte: [18]).

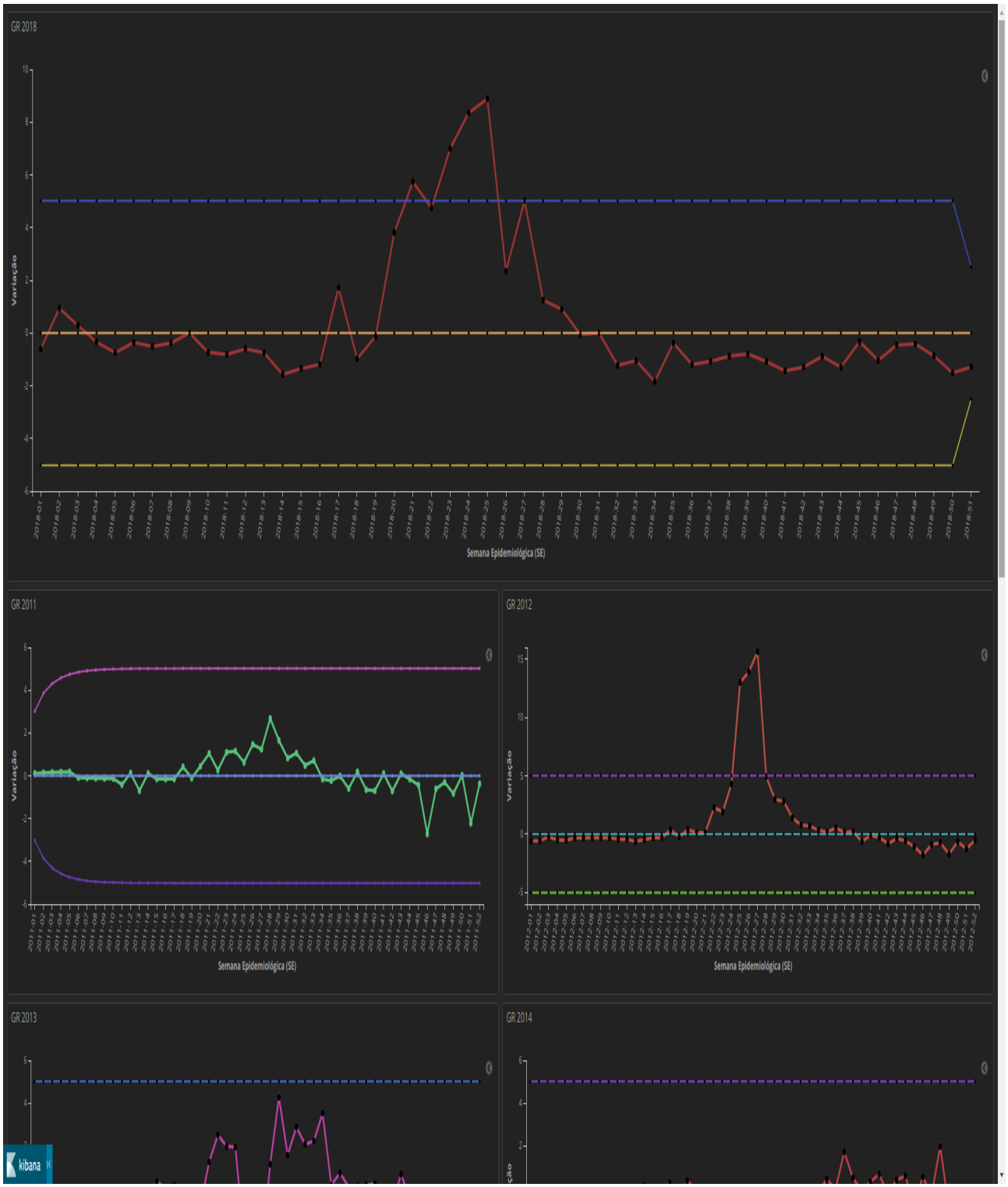



Figura II.2: Kibana: CEP da Influenza (2011, 2012 e 2018), Curitiba-PR (Fonte: [18]).

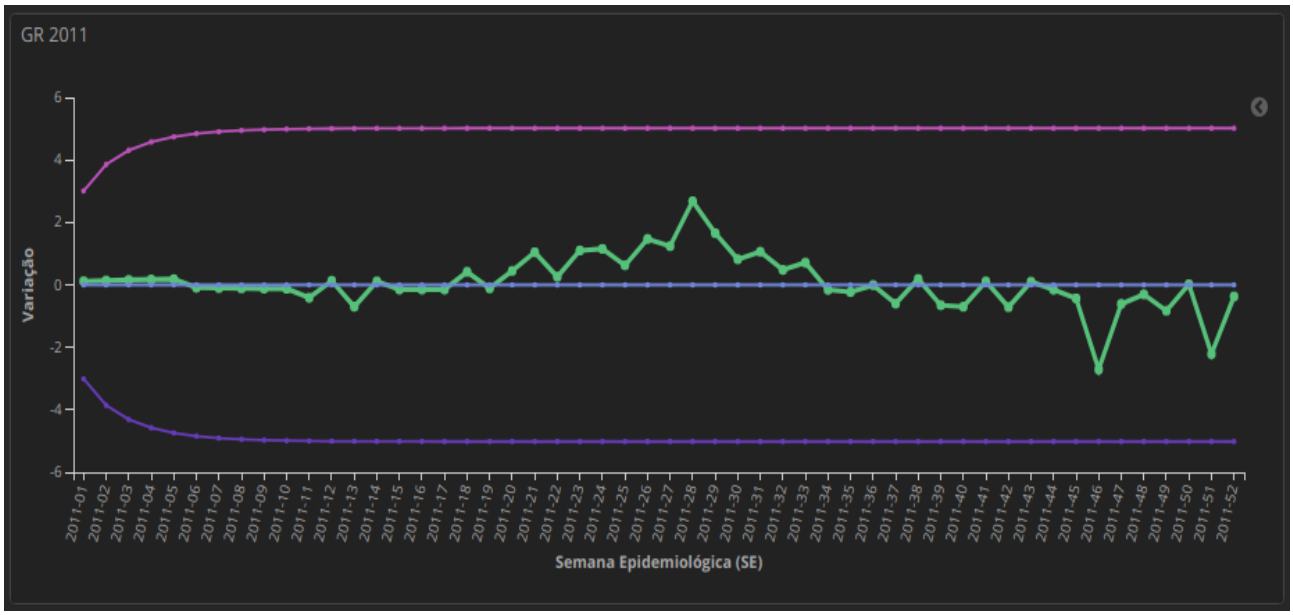



Figura II.3: Kibana: CEP da Influenza (2011), Curitiba-PR (Fonte: [18]).

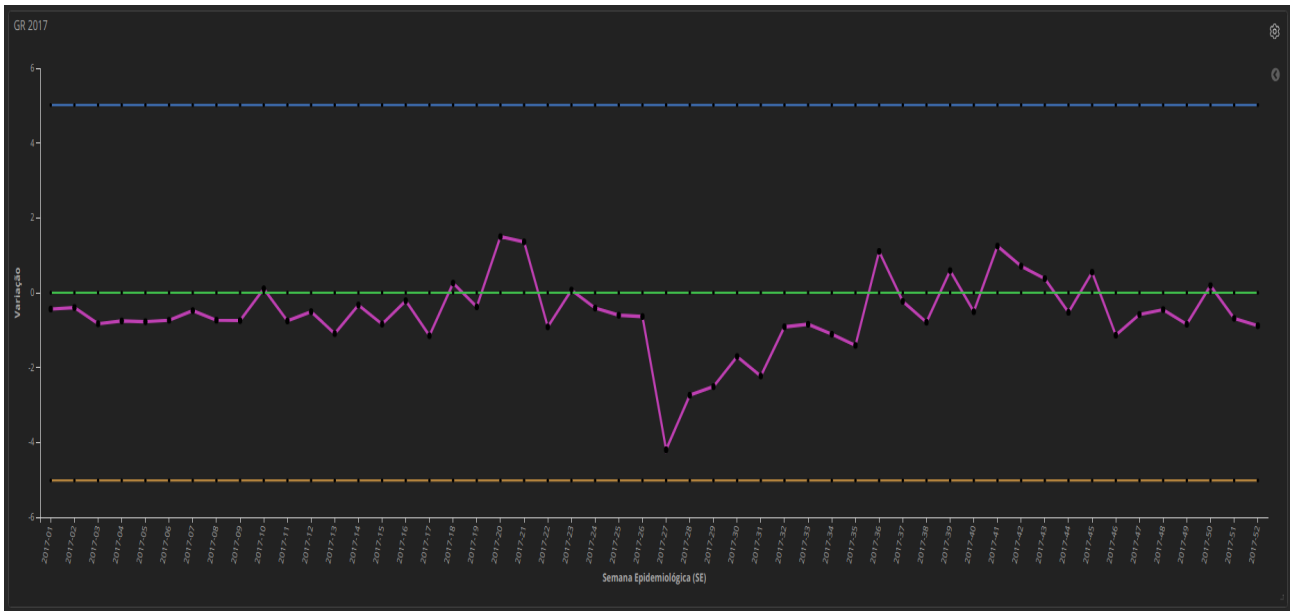



Figura II.4: Kibana: CEP da Influenza (2017), Curitiba-PR (Fonte: [18]).

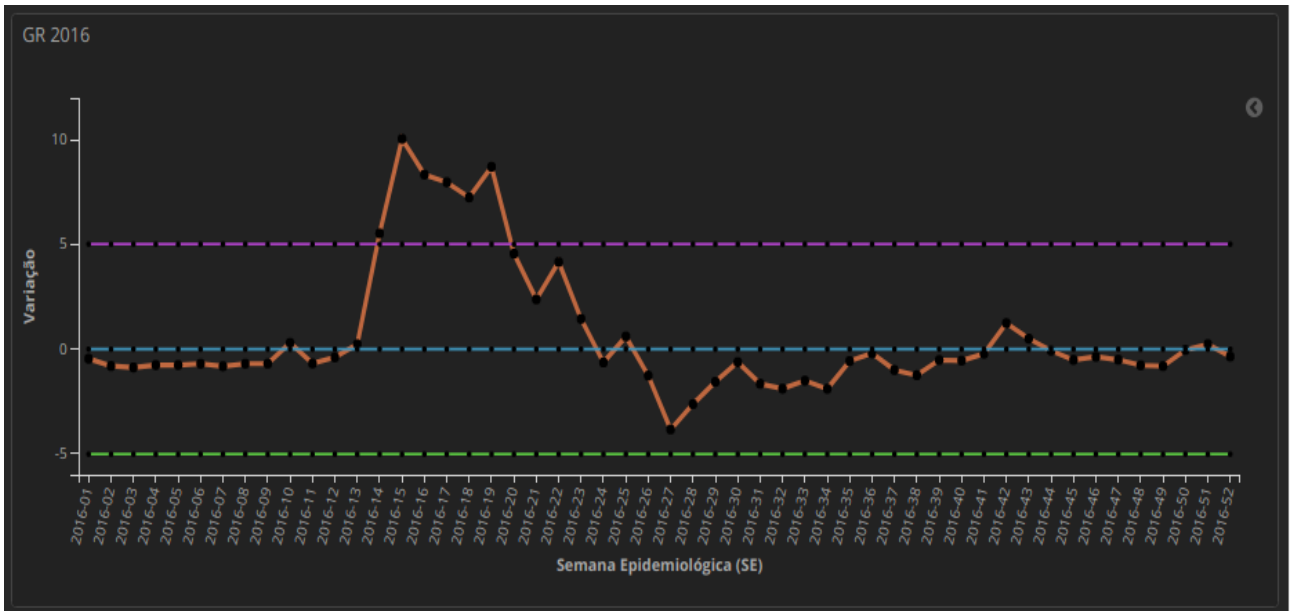



Figura II.5: Kibana: CEP da Influenza (2016), Curitiba-PR (Fonte: [18]).

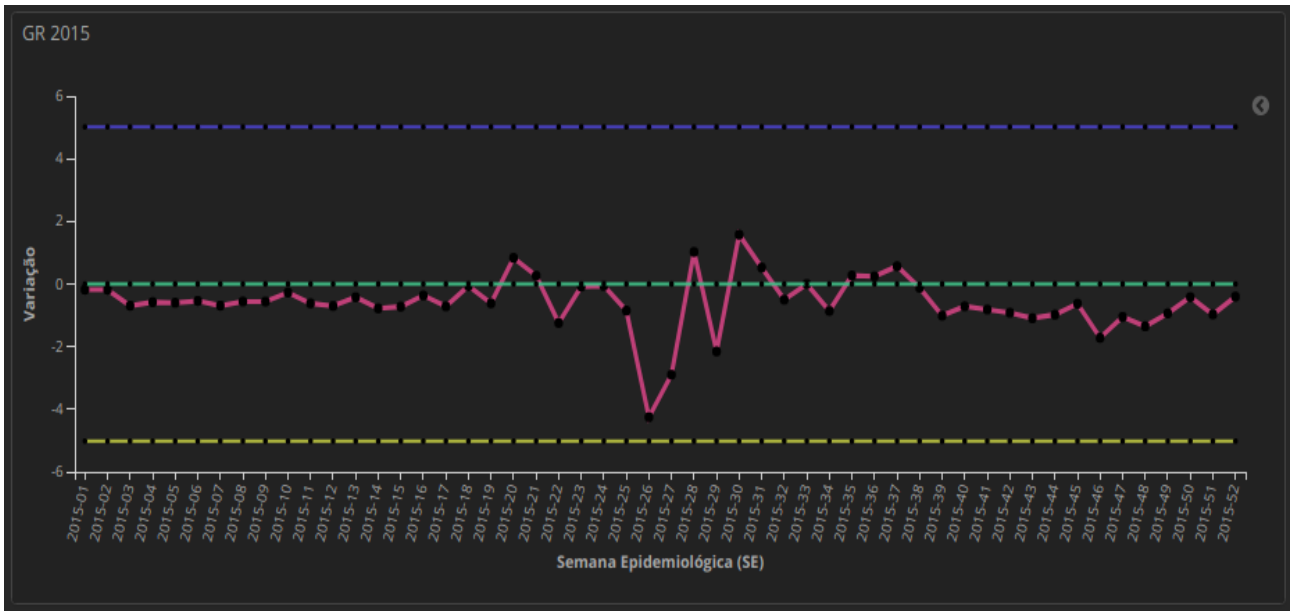



Figura II.6: Kibana: CEP da Influenza (2015), Curitiba-PR (Fonte: [18]).

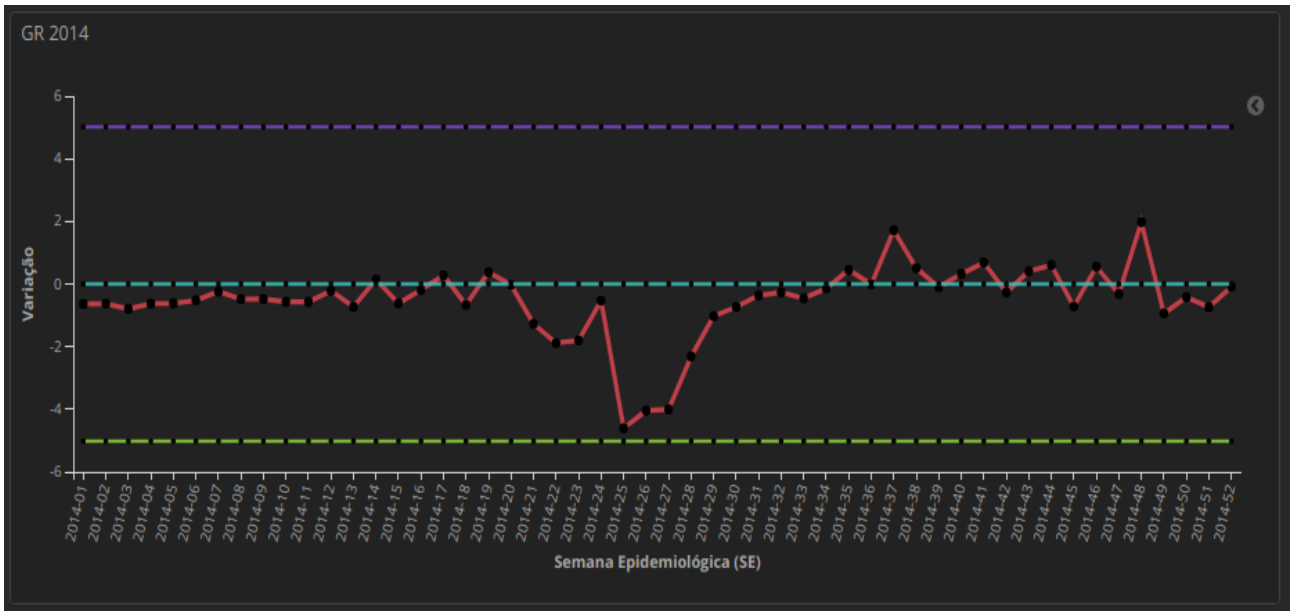



Figura II.7: Kibana: CEP da Influenza (2014), Curitiba-PR (Fonte: [18]).

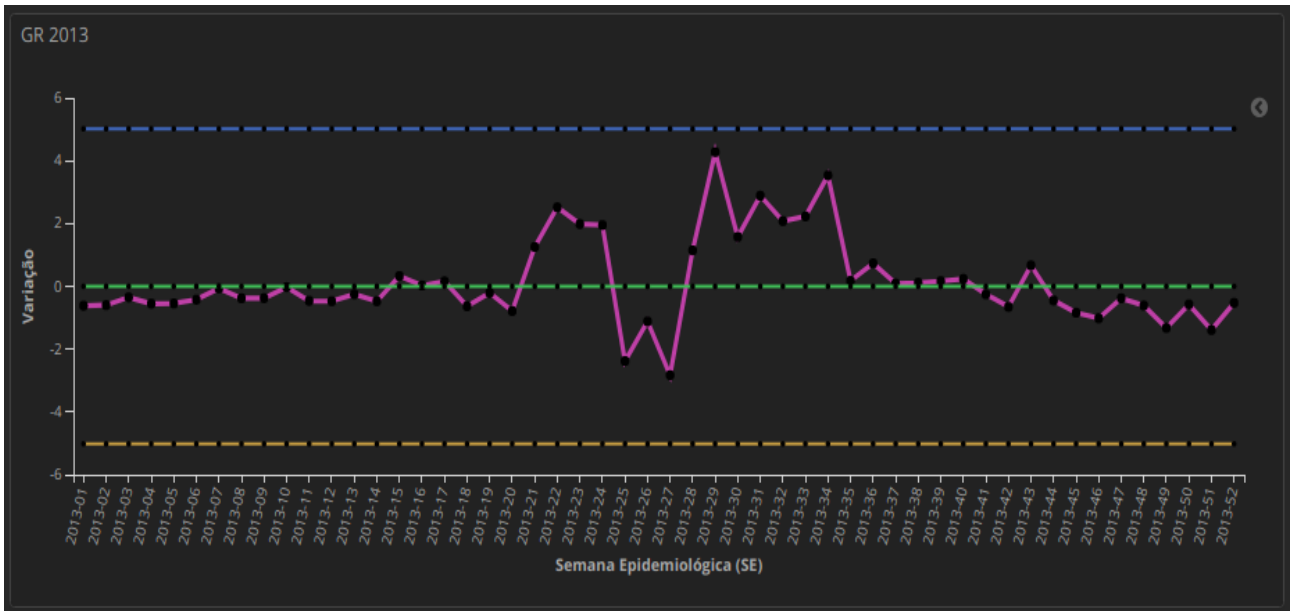



Figura II.8: Kibana: CEP da Influenza (2013), Curitiba-PR (Fonte: [18]).

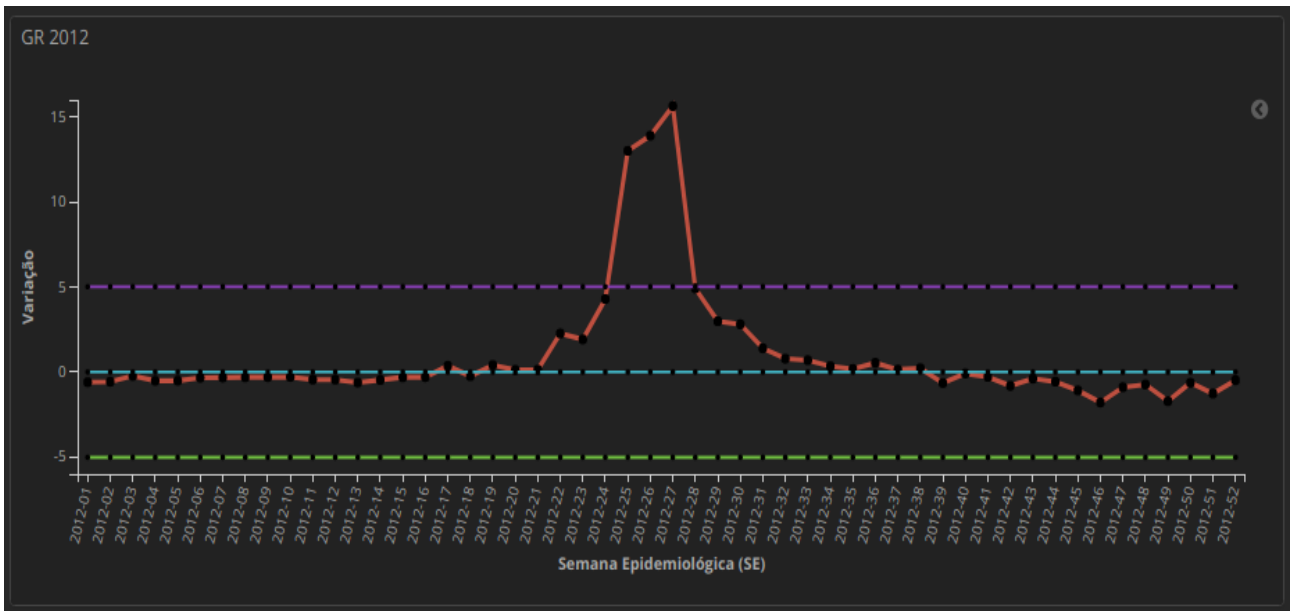



Figura II.9: Kibana: CEP da Influenza (2012), Curitiba-PR
(Fonte: [18]).

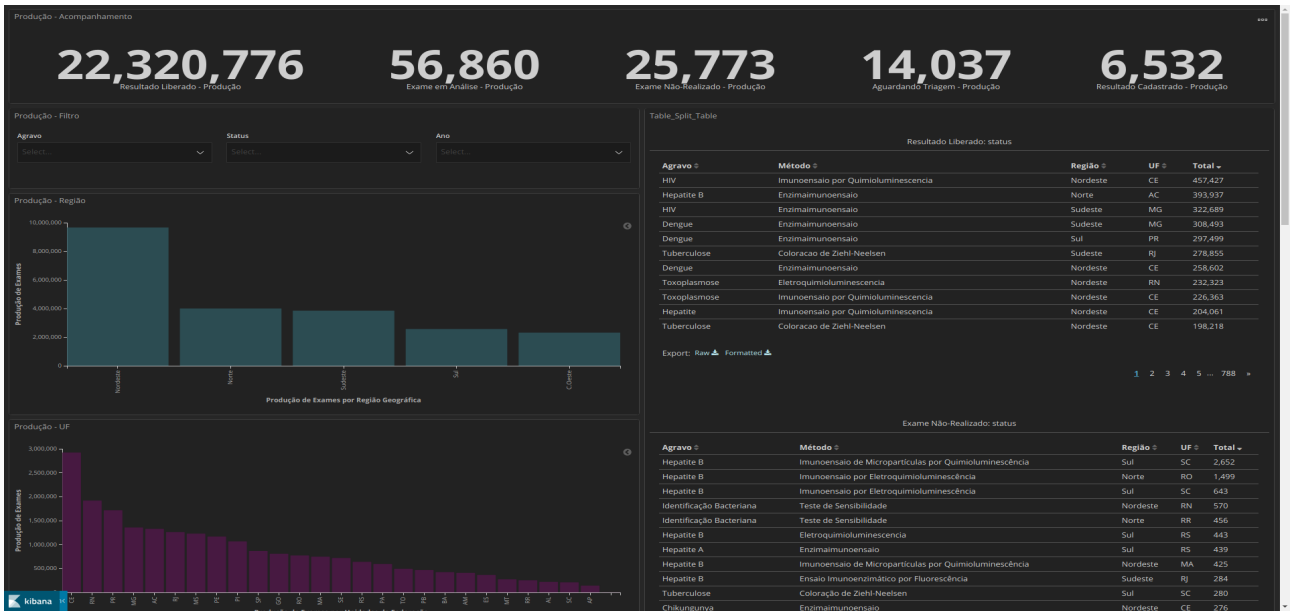



Figura II.10: Kibana: Produção de Exames - Parte 1
(Fonte: [18]).

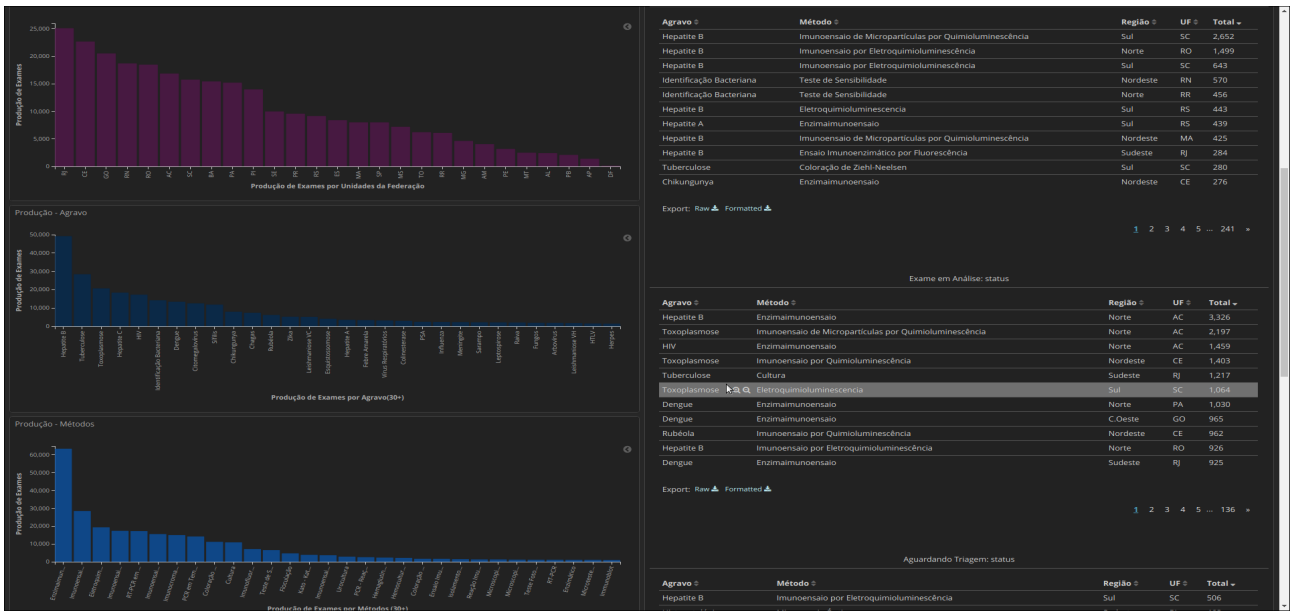



Figura II.11: Kibana: Produção de Exames - Parte 2
(Fonte: [18]).

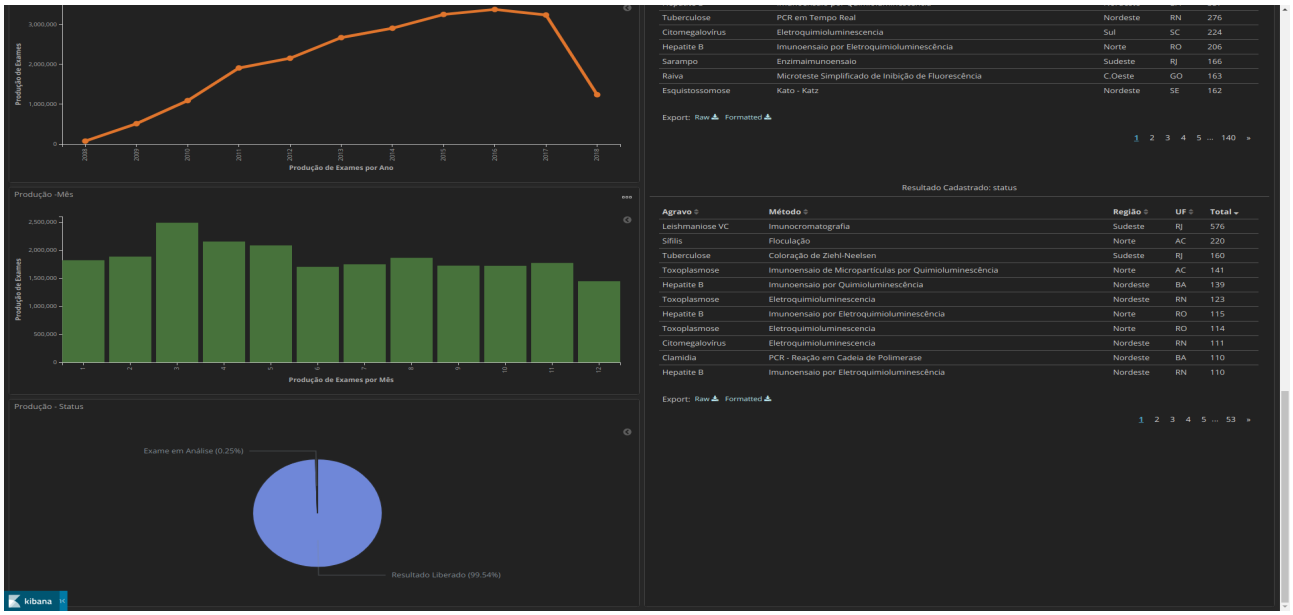



Figura II.12: Kibana: Produção de Exames - Parte 3
(Fonte: [18]).

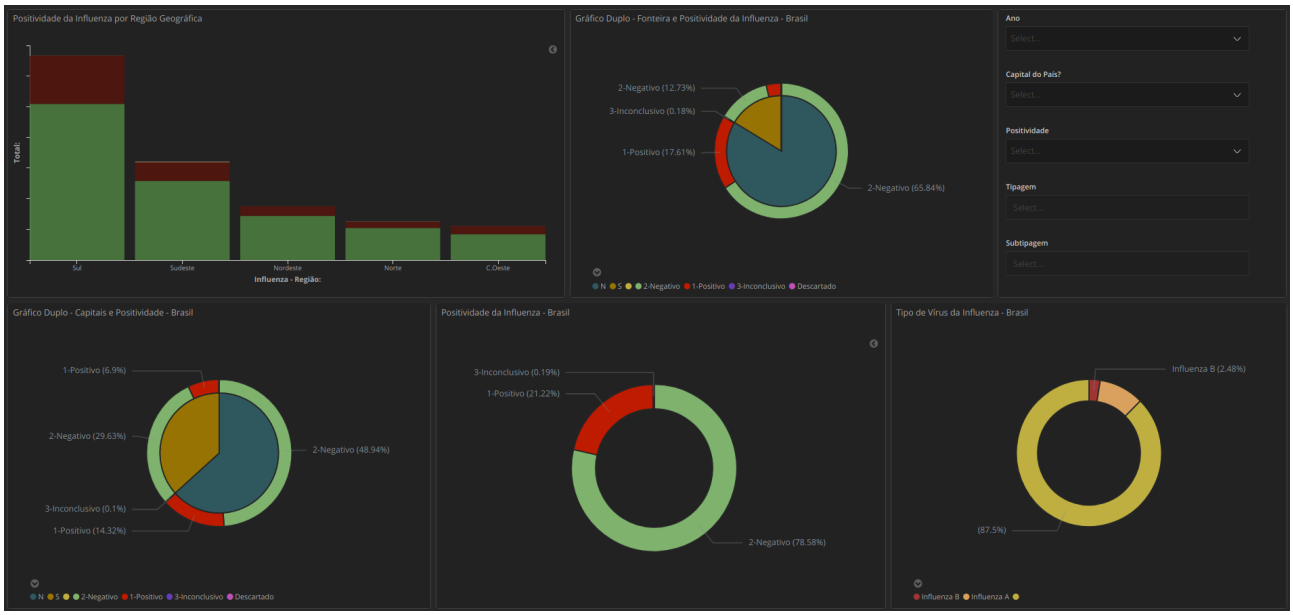



Figura II.13: Kibana: Exames de Influenza, Brasil - Parte 1 (Fonte: [18]).

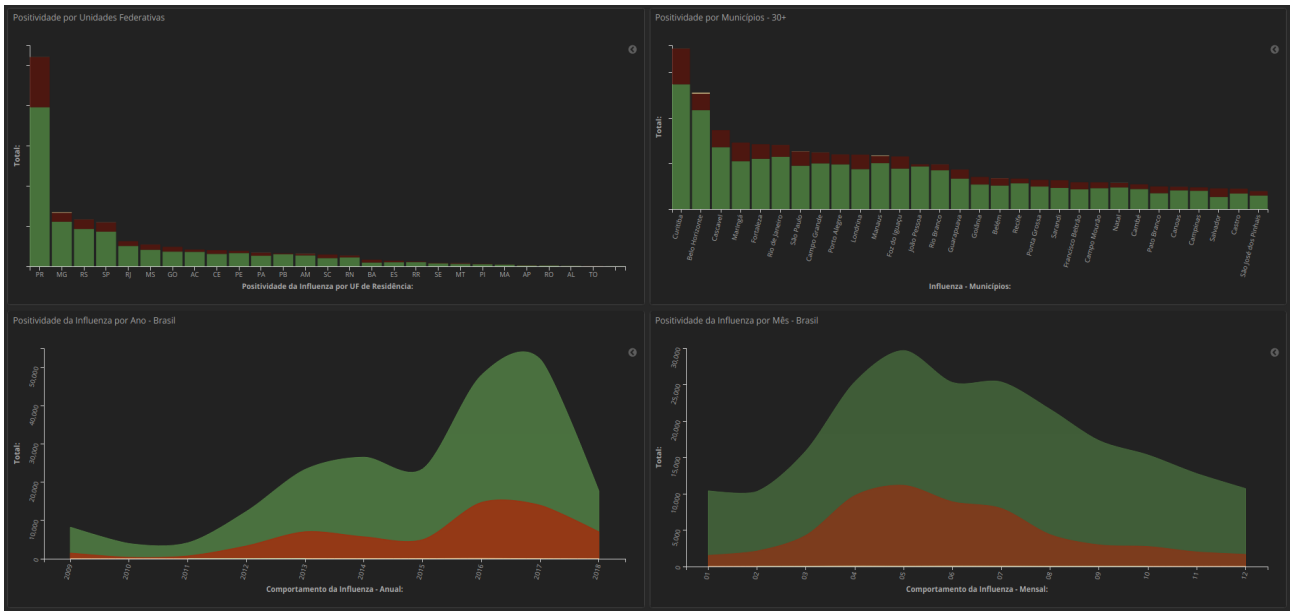



Figura II.14: Kibana: Exames de Influenza, Brasil - Parte 2 (Fonte: [18]).

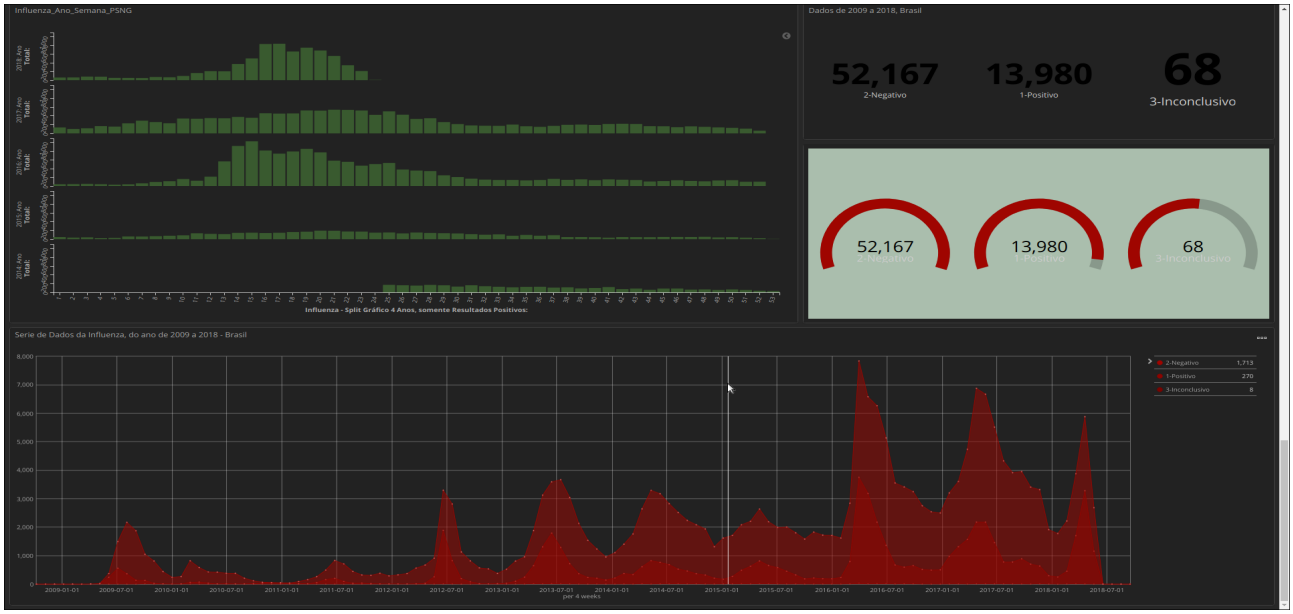



Figura II.15: Kibana: Exames de Influenza, Brasil - Parte 3
(Fonte: [18]).

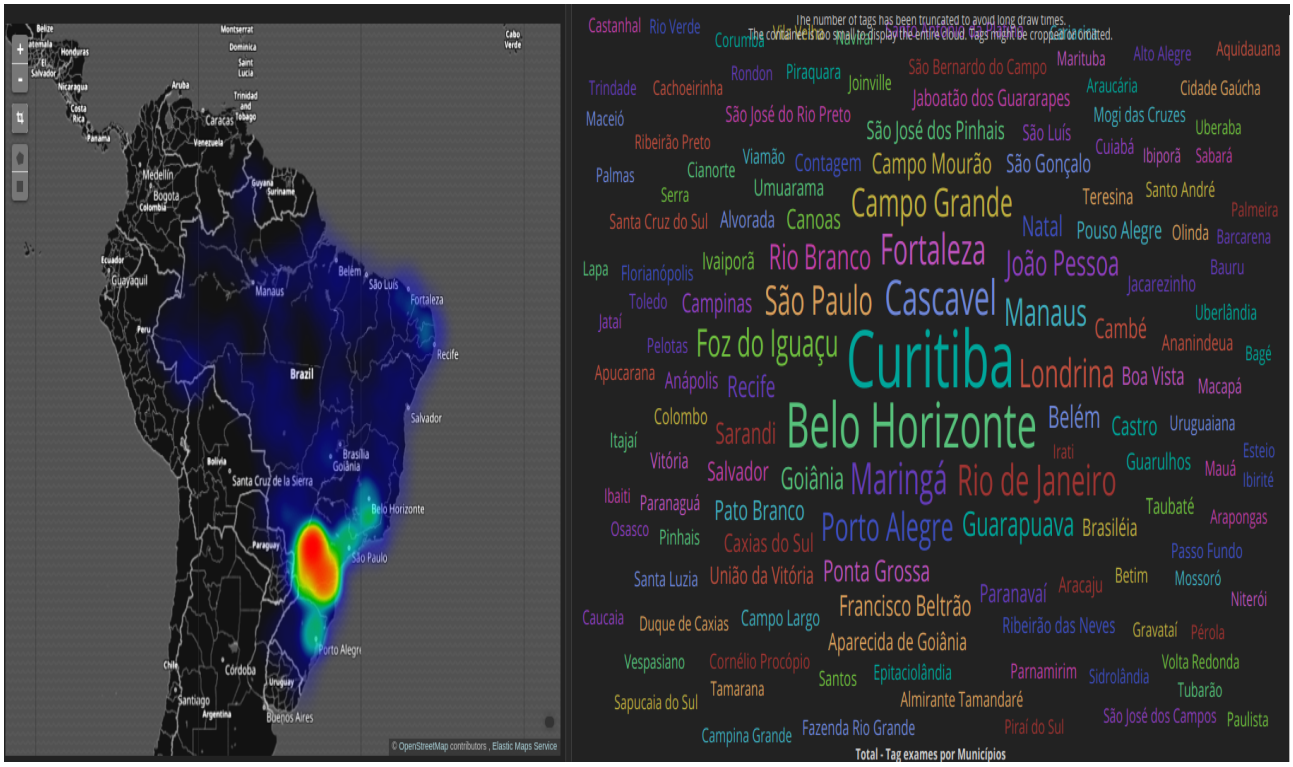



Figura II.16: Kibana: Exames de Influenza, Mapa e Tag - Brasil
(Fonte: [18]).

# Anexo III

# Parte do ETL - Influenza

*Step 4*: Scritp SQL do ETL - Padronização dos Resultados da Influenza

```
-------------------------------------------------------------------------
#### Exemplo da Padronização dos resultados de Influenza
ALTER TABLE ve_influenza ADD COLUMN psng_influenza character varying(20);
ALTER TABLE ve_influenza ADD COLUMN tipo_influenza character varying(18);
ALTER TABLE ve_influenza ADD COLUMN subtipagem_influenza character
            varying(50);

UPDATE ve_influenza
   SET psng_influenza =
    CASE resultado
        WHEN 'Resultado: Detectavel ' THEN '1-Positivo'
        WHEN 'Resultado: Nao Detectavel ' THEN '2-Negativo'
        WHEN 'Resultado: Inconclusivo ' THEN '3-Inconclusivo'
        WHEN 'Resultado: Reagente ' THEN '1-Positivo'
        WHEN 'Resultado: Nao Reagente ' THEN '2-Negativo'
    ELSE 'Descartado'
END;

 UPDATE VE_INFLUENZA
   SET tipo_influenza =
   CASE
        WHEN psng_influenza = '1-Positivo' and resultado2 =
        'Virus: Influenza "A" ' THEN 'Influenza A'
        WHEN psng_influenza = '1-Positivo' and resultado2 =
        'Virus: Influenza "B" ' THEN 'Influenza B'
        WHEN psng_influenza = '1-Positivo' and resultado13 =
```

```
        'Influenza A: Detectavel ' THEN 'Influenza A'
        WHEN psng_influenza = '1-Positivo' and resultado14 =
        'Influenza B: Detectavel ' THEN 'Influenza B'
    ELSE ''
END;

UPDATE VE_INFLUENZA
   SET subtipagem_influenza=
   CASE
        WHEN psng_influenza = '1-Positivo' and tipo_influenza =
        'Influenza B' and resultado5 = 'Linhagem: Victoria '
        THEN 'Linhagem: Victoria'
        WHEN psng_influenza = '1-Positivo' and tipo_influenza =
        'Influenza B' and resultado5 = 'Linhagem: Yamagata '
        THEN 'Linhagem: Yamagata'
        WHEN psng_influenza = '1-Positivo' and tipo_influenza =
        'Influenza A' and resultado3 = 'Subtipagem: Influenza
        A H1N1 (pdm09) ' THEN 'Influenza A H1N1 (pdm09)'
        WHEN psng_influenza = '1-Positivo' and tipo_influenza =
        'Influenza A' and resultado3 = 'Subtipagem: Influenza
        A inconclusivo para H1N1 (pdm09) ' THEN 'Influenza A
        inconclusivo para H1N1 (pdm09)'
        WHEN psng_influenza = '1-Positivo' and tipo_influenza =
        'Influenza A' and resultado3 = 'Subtipagem: Influenza A
        Sazonal ' THEN 'Influenza A Sazonal'
        WHEN psng_influenza = '1-Positivo' and tipo_influenza =
        'Influenza A' and resultado3 = 'Subtipagem: Influenza A Sazonal
        / H1 ' THEN 'Influenza A Sazonal/H1'
        WHEN psng_influenza = '1-Positivo' and tipo_influenza =
        'Influenza A' and resultado3 = 'Subtipagem: Influenza A Sazonal
        / H3 ' THEN 'Influenza A Sazonal/H3'
    ELSE ''
END;
-------------------------------------------------------------------------
```

```
--------------------------------------------------------------------------
Habilitando a função de Hash no SGBD PostgrSql:
->  extension if not exists pgcrypto;
--------
Criando o campo na tabela para armazenar o Hash:
->  atler table ve_influenza_brasil add column hash bytea;
--------
Incluindo o Hash na tabela:
->  ve_influenza_brasil
    update ve_influenza_brasil
        set hash = digest(cast((paciente, nascimento, municipio)
        as text),'sha256')
    where no_paciente <> ''
--------
Resultado SQL Select (hash):
Resultado do HASH das informações do paciente "FELIPE ANTONIO DA SILVA"
a) hash do nome do paciente:
")\244\316\004\274>\355\263C\326%m5{\310kBXaj"\036#Y\360\222\255\223
\360\021\304("
--------
b) hash do nome do paciente e data de nascimento:
"\317\355W\311zd+G\312\005\250F\256\221\330\011\247\244:VN\331\344\
313\225\264\270b\352\212H\254"
--------
c) hash do nome do paciente, município e data de nascimento:
"\002\236\303_\355\226\313\207\330%\367)S\004\307\377s\343.\314\005/
@!\271F\3770S\300Rc"
--------------------------------------------------------------------------
```