UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Gláucio Bezerra Rocha

Abordagem para Otimização do Processo de Alocação de Viaturas Policiais utilizando Registro de Ocorrências Criminais para Suporte no Processo Decisório Gerencial no Âmbito da Segurança Publica.

João Pessoa - PB

Dezembro - 2022

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Gláucio Bezerra Rocha

Abordagem para Otimização do Processo de Alocação de Viaturas Policiais utilizando Registro de Ocorrências Criminais para Suporte no Processo Decisório Gerencial no Âmbito da Segurança Publica.

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação da Universidade Federal da Paraíba por Gláucio Bezerra Rocha, sob a orientação do Prof. Dr. Gilberto Farias de Sousa Filho, como parte dos requisitos necessários para a obtenção do título de Mestre em Informática.

Linha de Pesquisa: Sinais, Sistemas Digitais e Gráficos .

Orientadora: Profº. Drº. Gilberto Farias de Sousa Filho.

João Pessoa - PB

Dezembro – 2022

Catalogação na publicação
Seção de Catalogação e Classificação

R672a Rocha, Gláucio Bezerra.
Abordagem para otimização do processo de Aalocação de viaturas policiais utilizando registro de ocorrências criminais para suporte no processo decisório gerencial no âmbito da segurança publica. / Gláucio Bezerra Rocha. - João Pessoa, 2022.
72 f. : il.

Orientação: Gilberto Farias de Sousa Filho.
Dissertação (Mestrado) - UFPB/CI.

1. Informática. 2. Clusterização - Inteligência policial. 3. Segurança Pública - Ocorrências policiais. 4. Viaturas policiais - Problema de localização. I. Sousa Filho, Gilberto Farias de. II. Título.

UFPB/BC CDU 004(043)

Elaborado por ANNA REGINA DA SILVA RIBEIRO - CRB-15/24

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de Gláucio Bezerra Rocha, candidato ao título de Mestre em Informática na Área de Sistemas de Computação, realizada em 02 de dezembro de 2022.

No dia dois do mês de dezembro de dois mil e vinte e dois, às quatorze horas, no Centro de Informática da Universidade Federal da Paraíba, em Mangabeira, reuniram-se os membros da Banca Examinadora constituída para julgar o Trabalho Final do Sr. Gláucio Bezerra Rocha, vinculado a esta Universidade sob a matrícula nº 20201006089, candidato ao grau de Mestre em Informática, na área de "Sistemas de Computação", na linha de pesquisa "Sinais, Sistemas Digitais e Gráficos", do Programa de Pós-Graduação em Informática, da Universidade Federal da Paraíba. A comissão examinadora foi composta pelos professores: Gilberto Farias de Sousa Filho (PPGI-UFPB), Orientador e Presidente da Banca, Alisson Vasconcelos de Brito (PPGI-UFPB), Examinador Interno, Lucidio dos Anjos Formiga Cabral (PPGI-UFPB), Examinador Interno, Bruno Jefferson de Sousa Pessoa (UFPB), Examinador Externo ao Programa. Dando início aos trabalhos, o Presidente da Banca cumprimentou os presentes, comunicou a finalidade da reunião e passou a palavra ao candidato para que ele fizesse a exposição oral do trabalho de dissertação intitulado "Abordagem para Otimização do Processo de Alocação de Viaturas Policiais utilizando Registro de Ocorrências Criminais para Suporte no Processo Decisório Gerencial no Âmbito da Segurança Publica". Concluída a exposição, o candidato foi arguido pela Banca Examinadora que emitiu o seguinte parecer: "**aprovado**". Do ocorrido, eu, Fernando Menezes Matos, Coordenador do Programa de Pós-Graduação em Informática, lavrei a presente ata que vai assinada por mim e pelos membros da banca examinadora. João Pessoa, 02 de dezembro de 2022.

Prof. Fernando Menezes Matos

Prof. Gilberto Farias de Sousa Filho
Orientador (PPGI-UFPB)

Prof. Alisson Vasconcelos de Brito
Examinador interno (PPGI-UFPB)

Prof. Lucidio dos Anjos Formiga Cabral
Examinador Interno (PPGI-UFPB)

Prof. Bruno Jefferson de Sousa Pessoa
Examinador Externo ao Programa (UFPB)

Abordagem para Otimização do Processo de Alocação de Viaturas Policiais utilizando Registro de Ocorrências Criminais para Suporte no Processo Decisório Gerencial no Âmbito da Segurança Publica.

Resumo

Considerando a segurança pública no Brasil como uma área de grande importância da sociedade atual, sendo um setor crítico nos Estados brasileiros estando diretamente ligada ao maior bem jurídico do ser humano, a vida, além do patrimônio, faz necessário a realização de estudo que permita mitigar o aumento da criminalidade por meio de uso de inteligência policial. Em análise realizada no âmbito da segurança pública do Estado da Paraíba, um dos problemas identificados e trazido para esse trabalho foi a má distribuição de viaturas policiais por toda área geográfica das cidades, ocasionando atrasos e ineficiências quando em atendimento a população. Atualmente as viaturas policiais são distribuídas em bases fixas localizadas em bairros, onde praticamente cada bairro possui algumas viaturas a disposição para atender aquela área, e percebe-se que atual alocação dessas patrulhas não consideram algumas características do fenômeno criminal para serem definidas suas localizações, a exemplo de número de ocorrências existentes e distância de locais com maior incidência criminal. Considerando todo o cenário do problema apresentado, esse trabalho propõe um processo de clusterização de áreas por meio de registros de ocorrências policiais, para posteriormente prever a quantidade de crimes nessas regiões identificadas, e finalmente, precreve uma solução através da resolução do Problema de Localização de Máxima Cobertura Capacitada adaptada aos dados criminais.

Palavras-Chave: Clusterização, Series Temporias, Problema de Localização, Análise Criminal, Segurança Pública.

Abordagem para Otimização do Processo de Alocação de Viaturas Policiais utilizando Registro de Ocorrências Criminais para Suporte no Processo Decisório Gerencial no Âmbito da Segurança Publica.

Abstract

Considering public security in Brazil as an area of great importance in today's society, being a critical sector in the Brazilian states and directly linked to the greatest legal asset of human beings, life, in addition to heritage, it is necessary to carry out a study that allows mitigating the increase in crime through the use of police intelligence. In an analysis carried out within the scope of public security in the State of Paraíba, one of the problems identified and brought to this work was the poor distribution of police vehicles throughout the geographic area of the cities, causing delays and inefficiencies when serving the population. Currently, police vehicles are distributed in fixed bases located in neighborhoods, where practically each neighborhood has some vehicles available to serve that area, and it is clear that the current allocation of these patrols does not consider some characteristics of the criminal phenomenon to define their locations, the example of number of existing occurrences and distance from places with higher crime incidence. Considering the entire scenario of the problem presented, this work proposes a process of clustering areas through records of police occurrences, to later predict the amount of crimes in these identified regions, and finally, prescribes a solution through the resolution of the Location Problem of Maximum Qualified Coverage adapted to criminal data.

Keywords: Clustering, Temporary Series, Location Problem, Criminal Analysis, Public Security.

SUMÁRIO

Introdução 1
1.1 Objetivo Geral 3
1.2 Objetivos Específicos 4
1.3 Organização da Proposta 4
Fundamentação Teórica 6
2.1 Series Temporais 6
2.1.1 Modelos de Series Temporais 10
2.1.2 Previsão de dados por meio de Series Temporais 12
2.2 Problema de Localização 15
2.3 Clusterização de Dados 18
2.4 Business Intelligence 21
2.5 Segurança Pública e Análise Criminal 24
Trabalhos Relacionados 28
Metodologia 33
4.1 Extração, Transformação e Business Intelligence 37
4.2 Clusterização 42
4.4 Problema de Localização de Máxima Cobertura Capacitado 46
Resultados 48
5.1 Clusterização de Regiões 48
5.2 Previsão de Crimes 51
5.3 Alocação de Viaturas utilizando o modelo matemático 55
Referências 61

LISTA DE FIGURAS

Figura 1 - Gráfico de previsão apresentado pela API Prophet
Figura 2 - Exemplo de aglomeração de elementos por meio do método heuristíco hierárquico
Figura 3 - Dados originais na forma que foram coletados
Figura 4 - Fase de clusterização dos registros criminosos. (a) Localização dos registros. (b) A grupamentos dos registros por proximidade.
Figura 5 - Quantidade de crimes registrados em um dia para cada região criminal em dia especifico.
Figura 6 - Previsão dos crimes em uma região baseado no seu histórico criminal.
Figura 7 - Solução do modelo proposto de localização de viaturas policiais para melhor atender as regiões
Figura 8 - Fluxo referente aos processo de Extração, Transformação e Business Intelligence
Figura 9 - Modelagem dimensional estrela do Business Intelligence.
Figura 10 - Dashboard desenvolvido com o auxílio da ferramenta QlikSense.
Figura 11 - Gráfico exibindo o ponto de menor ângulo definido como Cotovelo
Figura 12 - Gráfico exibindo a curva suave plotada pelo método Elbow.
Figura 13 - Gráfico da série temporal por dia do período de jan/2017 a dez/2021 da região 1.
Figura 14 - Gráfico Elbow indicando o melhor K igual a 3.
Figura 15 - Relação do raio médio entre os pontos e os centroides dos clusters.
Figura 16 - Gráfico de Elbow considerando o raio médio máximo de 4.0Km.
Figura 17 - Clusterização dos dados considerando o número de cluster K=13

LISTA DE TABELAS

Tabela 1 – Dados Reais Região 1
Tabela 2 – Dados Previsto Região 1
Tabela 3 - Dados Reais
Tabela 4 – MAPE das Regiões
Tabela 5 – MAPE Geral
Tabela 6 - Todos os Crimes – Mensal
Tabela 7 - Todos os Crimes – Diário
Tabela 8 - Crimes Contra o Patrimônio – Mensal
Tabela 9 - Crimes Contra o Patrimônio – Diária
Tabela 10 - Crimes Contra a Pessoa – Mensal
Tabela 11 - Crimes Contra a Pessoa – Mensal
Tabela 12 – Centroídes dos Bairros
Tabela 13 – Centroídes das Regiões
Tabela 14 – Matriz de Distâncias entre Bairros
Tabela 15 – Resultado de Alocação de Viaturas por Dia
Tabela 16 – Instância de alocação de viaturas
Tabela 17 – Resultado de Alocação de Viaturas por Semana

LISTA DE ABREVIATURAS E SIGLAS

AC – Análise Criminal

AR – Auto-regressico

ARIMA – Auto-regressivo Integrados de Média Móvel

ARMA – Auto-regressivo e Média Móvel

API – *Application Programming Interface*

BI – *Business Intelligence*

BO – Boletim de Ocorrência

DW – *Data Wharehouse*

IA – Inteligência Artificial

MAE – Erro Médio Absoluto

MCLP – *Maximum Coverage Location Problem*

MSE – Erro Quadrático Médio

ML – *Machine Learning*

OLAP – *Online Analytical Processing*

OLTP – *Online Transaction Processing*

RMSE – Raiz Quadrada do Erro Médio

SCP – *Set Convering Problem*

SPP – Sistema de Procedimento Policial

Capítulo

1

Introdução

A segurança pública no Brasil, área de grande importância da sociedade atual, é um setor crítico nos Estados brasileiros. Ferreira e Rigueira (2013) constatam que a segurança vive no limite de crises decorrentes de problemas estruturais, uma mídia pesada utilizada para denegrir sua imagem, e quanto mais a criminalidade aumenta os Estados não conseguem conter e controlar os problemas relacionados à segurança. É uma área de alta relevância pelo fato de estar diretamente ligada ao maior bem jurídico do ser humano, a vida, além do patrimônio. É responsável por proteger a vida, cessar e punir criminosos. De um lado temos um Estado desaparelhado sem conseguir dá uma resposta a contento à sociedade, de outro, um sistema organizado controlado por criminosos, denominado de Crime Organizado. Em Lum e Greene (2007) o crime organizado é definido como um empreendimento complexo, centralizado, constituído para a prática de atividades ilegais. Atualmente, para se combater o crime é necessário investimento em inteligência policial, por meio da Análise Criminal (AC). A análise criminal é responsável por produzir conhecimento policial, onde se avalia e acompanha sistematicamente o fenômeno da criminalidade em massa, instrumentando os operadores da segurança pública e proporcionando um suporte administrativo, tático, estratégico para a prevenção e repressão do crime e da violência (FERRO, 2006).

No Estado da Paraíba a realidade não é diferente. Pouca evolução percebe-se na estruturação da segurança pública, principalmente no tocante a tecnologia da informação como ferramenta de gestão e solução para subsidiar a atividade de análise criminal na produção de inteligência policial. Observa-se a ausência de vários indicadores e/ou informações pertinentes e relevantes para a gestão de segurança

pública que não estão acessíveis nos dias atuais para os gestores. Saber quais os crimes de maior incidência, bairros e/ou regiões mais perigosos, gêneros das vítimas, horários mais propícios do acontecimento de crime, locais que estão relacionados a determinados tipos de crimes, previsão de quantitativo de crimes em determinadas regiões para uma atuação mais planejada de efetivo policial, dentre outros questionamentos que não têm respostas, são exemplos de uma gestão precária, onde, caso o gestor da segurança pública tivesse posse desse conhecimento poderia gerir com maior eficácia um setor de tamanha relevância. A Polícia Civil, instituição ligada diretamente a segurança pública, e responsável por apurar todos os crimes ocorridos em território paraibano, no desenvolvimento de sua função de polícia judiciária, é responsável pelo registro, por meio das informações das vítimas, de todas as ocorrências de crimes ocorridas. Esses registros, que são formalizados em Boletins de Ocorrências (BO), contém todos os dados sobre o fato criminoso sofrido pela vítima, e caso bem utilizados e submetidos a uma metodologia de análise criminal pode responder muitas perguntas que hoje estão sem respostas, impactando diretamente na gestão e colhendo resultados mais assertivos no combate à criminalidade.

Nesse cenário, a ciência oferece diversas metodologias e ferramentas para auxiliar na solução e gestão de problemas nesse contexto, dentre elas estão Business Intelligence, Inteligência Artificial, Séries Temporais, Programação Matemática e Clusterização de Dados.

Em análise realizada no âmbito da segurança pública do Estado da Paraíba, um dos problemas identificados e trazidos para esse trabalho foi a mal distribuição de viaturas policiais por toda área geográfica das cidades, ocasionando atrasos e ineficiências quando em atendimento à população. Atualmente as viaturas policiais são distribuídas em bases fixas localizadas em bairros, onde praticamente cada bairro possui algumas viaturas a disposição para atender aquela área. Percebe-se que a atual alocação dessas patrulhas não considera algumas características do fenômeno criminal para serem definidas suas localizações, a exemplo de número de ocorrências existentes e distância de locais com maior incidência criminal. Dessa forma, é possível identificar bairros com maior índice criminal equipados com menos viaturas em detrimento de bairros com menor números de ocorrências, resultando em um superdimensionamento e/ou subdimensionamento de recursos e aparato policial, além de perceber que, as viaturas ficam circulando de forma aleatória nas suas respectivas regiões aguardando serem acionadas, tentando descobrir de forma empírica o local

que ocorrerá um crime, e muitas vezes no momento do acionamento estão em locais adversos daqueles que normalmente ocorre crimes com maior frequência.

Considerando todo o cenário do problema supracitado, esse trabalho propõe uma abordagem que utiliza os milhares de registros de ocorrências policiais registrados pela Polícia Civil que são ignorados, subutilizados, mais precisamente da cidade de João Pessoa-PB, que não são utilizados para fins de estudo do fenômeno do comportamento da criminalidade local, para fornecer um suporte a gestão da segurança pública. Para fins de gerar mais conhecimento policial, os dados históricos das ocorrências foram agrupados (clusterização), mapeando regiões com maior índice de criminalidade, independente de fronteiras legais existentes (bairros), mas com foco no número de ocorrências, apresentando a análise desses dados para que o gestor possa identificar padrões, tendência e sazonalidade por meio de séries temporais, e assim realizar previsões de ocorrências policias nestas regiões identificadas no processo de clusterização, para um melhor planejamento. Posteriormente, identificadas as regiões com maiores índice de criminalidade e a previsão de crimes nessas regiões, propomos uma fase de prescrição da localização das viaturas policiais com o objetivo de cobrir as regiões com maior demanda criminal previsto pelas fases anteriores. Para isto, propomos um modelo matemático para o Problema de Localização de Máxima Cobertura Capacitada adaptado aos dados criminais, considerando distância de cobertura e capacidade de atendimento das viaturas.

1.1 Objetivo Geral

Esse trabalho tem como objetivo geral o desenvolvimento de uma solução que para definir uma melhor estratégia de posicionamento de viaturas policiais, visando proporcionar uma maior cobertura das áreas identificadas por meio das incidências criminais, considerando a capacidade de atendimento de ocorrências por cada viatura e um raio de cobertura pré-estabelecido, possibilitando-os a realizar uma gestão mais assertiva.

1.2 Objetivos Específicos

Tendo em vista o objetivo geral desse trabalho, segue abaixo os objetivos específicos a serem atingidos:

- Extração e preparação de dados oriundos do sistema de registro de ocorrências policiais para fim de análise e suporte ao processo decisório;
- Clusterização dos dados de ocorrências policiais para delimitar áreas em que se concentrem maiores índices de criminalidade;
- Criação de séries temporais das regiões mapeadas e previsão de crimes nessas áreas;
- Proposta de um modelo matemático que permita posicionar patrulhas policiais dentro das áreas identificadas, considerando distância e capacidade de números de ocorrências máxima de atendimento de cada viatura.

1.3 Organização da Proposta

Esse estudo de mestrado está estruturado em 6 (seis) capítulos. No Capítulo 1 é apresentado os tópicos referentes a motivação e contextualização do problema, como também o objetivo geral e específico.

O Capítulo 2 é apresentado à fundamentação teórica, onde é realizada a definição e discussão acerca de Séries Temporais, Business Intelligence, Problema de Localização, Clusterização de dados, como também uma explanação sobre Segurança Pública e Análise Criminal, destacando a importância da implementação de uma gestão mais profissional.

O Capitulo 3 é apresentado alguns trabalhos relacionados no âmbito dos temas discutidos. No Capítulo 4 é apresentado a metodologia utilizada para desenvolvimento da solução desse trabalho, detalhando sobre o processo de extração e transformação de dados, clusterização, séries temporais, previsão de dados e o modelo matemático do problema de localização de máxima cobertura capacitada.

O capítulo 5 apresenta os resultados obtidos no decorrer do desenvolvimento do trabalho. E por fim, o capítulo 6 apresenta conclusões do trabalho.

Capítulo

2

Fundamentação Teórica

2.1 Series Temporais

Série temporal é definido por Garcia (2000) como uma sequência de dados ordenados cronologicamente. Esses dados são oriundos de eventos registrados sempre em função de uma variável tempo, podendo ser agrupados de forma diária, semanal, mensal, anual, etc. Esses eventos são vistos em diversas áreas de conhecimento, tais como: economia, meteorologia, saúde, comércio, etc. Morettin e Toloi (1981) trás alguns exemplos de séries temporais:

- Estimativa trimestral do Produto Nacional Bruto;
- Valores diários de temperatura;
- Índices diários da Bolsa de Valores;
- Quantidade mensal de chuva;
- Valores mensais de vendas de veículos;
- Registro de marés.

Outros exemplos de séries temporais apresentadas por Morettin e Toloi (2004) são:

- Valores mensais de consumo de energia elétrica;
- Emissão diária de poluentes de uma cidade;
- Índices do Produto Interno Bruto;
- Índices de custos de vida de uma cidade;

Como pode ser observado, as séries temporais estão bem mais presentes no dia-a-dia do que possamos imaginar. O objetivo de uma série temporal é analisar os dados registrados e identificar padrões, e a partir dessa análise poder entender alguns comportamentos, como exemplo, como um produto está sendo comercializado, se está sendo bem aceito, épocas do ano que se aumentam as vendas, assim como permitir a realização de previsões de vendas futuras. Morettin e Toloi (1981) apresenta alguns objetivos da análise de uma série temporal:

- Investigar o mecanismo que gerou a série temporal;
- Fazer previsões de valores;
- Descrever comportamentos;
- Procurar periodicidade relevantes nos dados.

Para Ehlers (2007), os objetivos de analisar uma série temporal, de modo geral, podem ser:

- Descrever as propriedades da série, tais como: tendência, sazonalidade, discrepâncias, alterações estruturais, etc;
- Predição, ou seja, com base em dados do passado prevê valores futuro.
- Controlar qualidade de processos;
- Explicar a variação de uma série temporal baseado em outra série temporal.

No tocante a predição de valores, Ehlers (2007) enfatiza que a previsão do futuro envolve o fator incerteza, onde as elas não são perfeitas, mas tem que buscar sempre reduzir os erros dessa previsão.

Para Ehlers (2007) a ordem dos dados apresentados em uma série temporal é determinante, assim como aponta que a principal característica a ser observada numa série temporal é que os dados vizinhos são dependentes, onde se faz necessário estudar, analisar e modelar essa dependência. Outras características apresentadas afirmam que as observações correlacionadas em séries temporais são difíceis de análise necessitando de técnicas específica para esse fim, assim como ratifica a dificuldade de lidar com observações perdidas e dados discrepantes devido à natureza sequencial.

Segundo Morettin e Toloi (2004), os modelos que descrevem séries temporais são processos estocásticos. Ehlers (2007) define processos estocásticos como uma coleção de variáveis aleatórias que são ordenadas no tempo e definidas como um conjunto de pontos que podem ser discreto ou contínuos. A definição de um modelo de classificação de série temporal vai depender de vários fatores, tais como: comportamento do fenômeno em análise ou conhecimento a priori sobre a natureza e do objetivo da análise. Uma das classificações de processos estocásticos são os processos estacionários ou não-estacionário. Um processo estacionário se desenvolve quando a origem do tempo não é importante na análise, apresentando uma constância, diferentemente do processo não-estacionário.

Morettin e Toloi (2004) classifica os tipos de modelos de séries temporais em dois, quais sejam: modelos paramétricos e os não paramétricos. Os modelos paramétricos são aqueles que os números de parâmetros considerados são finitos, já os não paramétricos envolvem um número infinitos de parâmetros.

Os modelos paramétricos toda a análise é realizada no domínio do tempo, e os modelos mais usados são os modelos de regressão, modelos auto-regressivo e de média móvel (ARMA), modelos auto-regressivo integrados e de médias móveis (ARIMA), modelos de memória longa (ARFIMA), modelos estruturais e modelos não-lineares. Os modelos não paramétricos mais usados são os a função de auto-covariância, ou autocorrelação, e sua transformada de Fourier, o espectro.

Quando falamos em série temporal temos que considerar alguns componentes desta que são fundamentais para seu entendimento, quais sejam: a sazonalidade, o ciclo da série, a aleatoriedade e a tendência (GARCIA, 2000).

Para Garcia (2000) e Ehlers (2007), a sazonalidade diz respeito a um padrão identificado que se repete a cada *p* período de tempo idêntico, onde p é denominado de fator de sazonalidade. Um exemplo prático acontece na venda de protetores solar no período de verão, onde pode ser percebido um aumento nas vendas desses produtos. Ehlers (2007) apresenta dois tipos de sazonalidade, a Aditiva e a Multiplicativa. A sazonalidade aditiva a série temporal apresenta flutuações sazonais próxima a uma constância não considerando o nível global a série. Supondo que seja esperado um aumento de 1 milhão de reais de venda em um determinado mês do ano, esse valor deve ser somado a média anual naquele mês especificamente. Já a sazonalidade multiplicativa as flutuações sazonais vão depender do nível global as séries. Considerando o cenário de venda no qual espera-se um aumento de 10% em relação da média anual em um período específico, e seu valor dependerá do valor da média anual, por isso denominada de sazonalidade multiplicativa.

Garcia (2000) define com ciclo da série como a variação cíclica de um determinado padrão. Esse componente difere da sazonalidade por não ter um intervalo frequente dessa variação, apresentando uma desregularidade.

A tendência de uma série temporal mostra o crescimento ou declínio dos valores dessa série. É um componente que apresenta mudanças graduais em longo prazo. Para Ehlers (2007) a tendência em uma série pode apresentar vários padrões, quais sejam:

- Crescimento linear: onde espera-se um aumento nas vendas a cada período de 1 milhão de reais;
- Crescimento exponencial: onde espera-se um aumento nas vendas a cada período em um percentual de 30%, resultando em um fator de 1,3.
- Crescimento amortecido: onde espera-se um aumento nas vendas a cada período em um percentual sobre o aumento de vendas registrado no período anterior. Se no período x foi registrado um aumento de 1 milhão de reais que correspondeu a 70% de aumento, nos próximos períodos o aumento esperado será que 70% de 1 milhão de reais, resultando em 700

mil, e assim acontecerá nos demais períodos, mantendo o percentual, mas o aumento bruto sendo amortecido.

Garcia (2000) define a aleatoriedade de uma série temporal como um comportamento identificado que não se encontra explicação, não sendo possível realizar nenhum tipo de previsão, justificando até uma possível falha de previsão. Muitas vezes são resultantes de fatos inesperados como pandemias, catástrofes naturais etc.

2.1.1 Modelos de Series Temporais

Para Morettin e Toloi (1981) os modelos de séries temporais servem para descrever o comportamento de uma série em particular, e seu desenvolvimento está relacionado a alguns fatores, tais como: o comportamento do fenômeno, conhecimento prévio que o analista tem da sua natureza, do objetivo da análise, da existência de métodos ótimos de estimação e da disponibilização de programas adequados.

Nessa seção vamos apresentar alguns modelos adequados para uso em séries temporais, os denominados de processos estocásticos.

Para Pal e Prakash (2017) os modelos auto-regressivo (AR) são muito bons na captação de tendência, pois os valores a serem projetados são previstos com base em nos valores passados. Os autores afirmam que esse modelo é definido com AR(p), onde p se refere a ordem do componente AR.

- O modelo AR de primeira ordem é denotado por AR(1):

$$X_t = \phi\epsilon_{t-1} + \epsilon_t$$

- O modelo AR de segunda ordem é denotado por AR(2):

$$X_t = \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \epsilon_t$$

- O modelo AR de p^{th} ordem é denotado por AR(p):

$$X_t = \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \ldots + \phi_p\epsilon_{t-p} + \epsilon_t$$

Os modelos ARMA, Auto-regressivo e de média móvel, é apresentado por Ehlers (2007) como uma combinação dos modelos Auto-Regressivo (AR) e Média Móvel (MA) e pode-se obter uma representação mais adequada com um menor número de parâmetros. O modelo auto-regressivo regride em seus próprios termos desfasados para realizar a previsão, e o modelo de média móvel define sua função baseada no erro do passado, que são eventos não normais e imprevisíveis (Pal e Prakash, 2017). Ehlers (2007) apresenta o modelo ARMA(p, q):

$$X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-\mathrm{p}} + \epsilon_t + \beta_1 \epsilon_{t-1} + \ldots + \beta_q \epsilon_{t-q}$$

Para Yenidogan et al. (2018) os modelos ARIMA, Auto-Regressivo Integrados de Média Móvel, são bem conhecidos em aplicações de séries temporais e de ciência de dados. Eles são classificados em modelos sazonais, onde feriados e outras datas específicas são aplicadas diretamente no conjunto de dados, e os modelo não sazonais, que removem os efeitos sazonais, feriados e outras datas, do conjunto de dados e usam linhas de tendências puras. Os modelos ARIMA são aplicados a séries temporais não estacionárias, são denotado ARIMA(p, d, q) e é dado por Ehlers (2007) logo abaixo, onde p representa no número de parâmetros auto-regressivos, d o número de diferenciações para que a série torne-se estacionárias, e o q representa o número de parâmetros de média móveis.

$$W_t = \alpha_1 W_{t-1} + \ldots + \alpha_p W_{t-\mathrm{p}} + \epsilon_t + \beta_1 \epsilon_{t-1} + \ldots + \beta_q \epsilon_{t-q}$$

Ehlers (2007) destaca que para dados reais um modelo ARIMA é no máximo uma aproximação para o verdadeiro processo gerador dos dados.

Morettin e Toloi (2004) apresenta a abordagem Box e Jenkinks como uma metodologia bastante utilizada na análise de dados que consiste em ajustar modelos ARIMA a um determinado conjunto de dados, e utilizam uma estratégia de construção de modelo baseada no ciclo iterativo, no qual a escolha da estrutura do modelo é baseada nos próprios dados. Morettin e Toloi (2004) apresenta os seguintes estágios desse ciclo iterativo:

- Especificação: classe geral de modelos considerada para a análise;
- Identificação: modelo, com base na análise de autocorrelação, autocorrelações parciais e outros critérios;

- Estimação: parâmetros do modelo identificado são estimados;
- Verificação ou Diagnóstico: modelo ajustado, através de uma análise de resíduos, para se saber se este é adequado para a previsão.

2.1.2 Previsão de dados por meio de Series Temporais

Como já mencionado, as séries temporais são formadas por dados oriundo de eventos e dispostos em relação a variável tempo. Elas estão presentes nas mais diversas áreas de sociedade, e um dos objetivos da análise de séries temporais é a previsão de valores futuros.

A previsão de valores por meio de séries temporais, segundo Taylor e Letham (2017)**,** ajuda na gestão das organizações, permitindo que seja realizado um melhor planejamento, tais como: capacidade suportada, definição de metas e detecção de anomalias.

Apesar da importância da previsão de dados baseado em séries temporais, para Taylor e Letham (2017) muitos desafios estão presentes nesse contexto, principalmente na previsão de dados confiáveis e de alta qualidade, tendo em vista a grande variedade dessas séries temporais e a raridade de profissionais com experiências em modelagem.

Para Ahmed et al. (2020) os modelos mais populares para previsões de séries temporais são o ARIMA e LSTM RNN. Nesse contexto o Facebook desenvolveu um modelo baseado em aprendizado de máquina, denominado Prophet, para previsão de séries temporais.

Segundo Ahmed et al. (2020) o Prophet é um modelo baseado em regressão aditiva simples $y(t)$, compostos de três principais componentes, quais sejam: tendência, sazonalidade e efeitos de feriados. Ele foi desenvolvido para facilitar as previsões de alta qualidade e obter uma previsão mais precisa e realista, sendo necessário apenas uma quantidade de dados histórico para melhor modelagem.

Taylor e Letham (2017) apresenta o Prophet como uma abordagem prática para previsão em escala que combina modelos configuráveis. Prophet é um modelo de regressão simples e modular com parâmetros configuráveis que podem ser ajustados

pelo especialista que tem domínio do conhecimento sobre a série temporal, e que funciona bem com parâmetros padrão. É aplicado um modelo de série temporal decomponível com três componentes: tendência, sazonalidade e feriados, conforme equação seguinte:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

A função $g(t)$ representa a tendência, modelando as mudanças não periódicas no valor da série temporal. Já $s(t)$ é a função que representa as mudanças periódicas, como sazonalidade semana, anual etc. Por fim, a função $h(t)$ representa os efeitos dos feriados. O termo ϵ_t representa o erro, que são as mudanças incomuns que não são comportadas pelo modelo.

Em Prophet (2022) demonstra que aplicação prática do modelo *Prophet* se dar de forma simples, por meio da API Prophet, basta instanciar a classe *Prophet*, e posteriormente invocar os métodos *fit()* e *predict()*.

Inicialmente quando é instanciada a classe *Prophet* é possível passar para seu construtor algumas configurações para serem utilizadas na previsão. O método *fit()* recebe um *dataframe* de formado do Pandas, com os dados histórico, possuindo duas colunas, sendo uma chamada *ds*, com formado AAAA-MM-DD (ano-mês-dia), e outra coluna chamada *y*, que representa a medida que se deseja prever. É importante ressaltar que os nomes das colunas deve ser, obrigatoriamente, os nomes supracitados, e qualquer alteração destas a API não proverá o resultado esperado.

É possível a API considerar dados não diários, ou seja, subdiários, e para esses casos a coluna *ds* deverá ser AAAA-MM-DD HH:MM:SS, consequentemente a sazonalidade diária é ajustada automaticamente.

Após a invocação do método *fit()* com os *dataframe* contendo os dados histórico, é necessário gerar um *dataframe* com a quantidade de dias que deseja a previsão. Esse *dataframe* pode ser gerado por meio do método *make_future_dataframe(periods=365)*, passando como parâmetro o *periods*, que representa a quantidade de dias a ser gerada a previsão.

Já com o modelo pronto e o *dataframe* com as datas futuras preparado, invoca-se o método *predict()* com o referido *dataframe*, que resultará em um *dataframe* contendo as seguintes colunas: *ds*, *yhat*, *yhat_lower* e *yhat_upper*. A previsão

encontra-se na coluna *yhat.* É possível plotar a previsão por meio do método *plot(dataframe_futuro)*, como pode ser visualizado exemplo na Figura 1 abaixo:

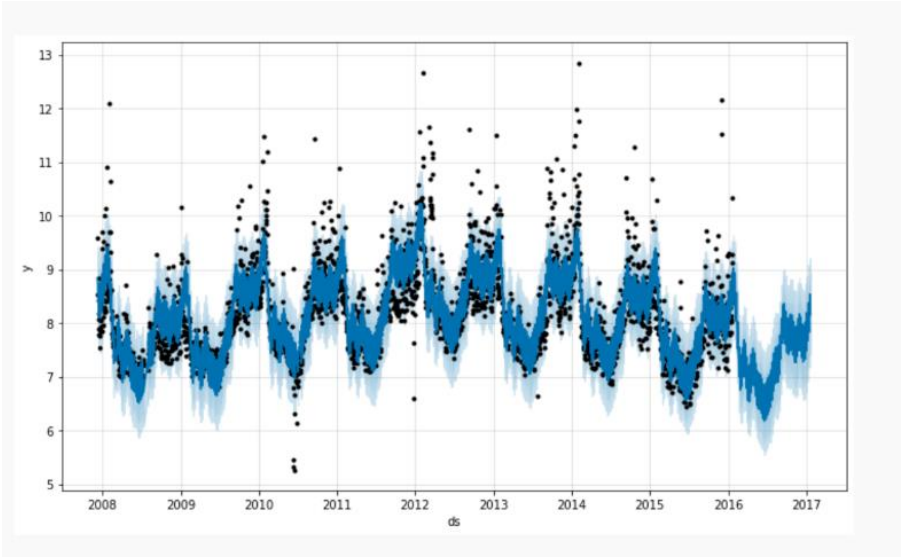

Figura 1. Gráfico de previsão apresentado pela API Prophet.

É possível constatar no gráfico da Figura 01 os dados histórico reais representados por pontilhados de cor preto, e na cor azul escuro os ajustes desses dados histórico reais, e por fim, em dado momento percebe-se que os dados histórico reais são cessados (pontilhado de cor preto) e apenas são plotados os dados do futuro, em continuidade na cor azul escuro, resultado da *API Prophet*.

Por fim, para mensurar o erro do processo de previsão da API *Prophet*, se faz necessário a utilização de métricas de avaliação, dentre as quais destaca-se: Erro Médio Absoluto (MAE), Erro Quadrático Médio (MSE) e a Raiz Quadrada do Erro Médio (RMSE).

O cálculo do Erro Médio Absoluto (MAE) é dado pela seguinte expressão:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|^2$$

O cálculo do Erro Médio Quadrático (MSE) é dado pela seguinte expressão:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

A raiz Quadrada do Erro Médio (RMSE) é dada pela seguinte expressão:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

2.2 Problema de Localização

Para Horner (2009) os problemas de localização é objeto de estudo há muitos anos, onde vários povos nômades procuravam por regiões para se alocarem considerando vários fatores que influenciariam em seu dia-a-dia, tal como alimentação, água, solo produtivo, facilidade de acesso, etc.

Para Oliveira (2020) a caracterização dos problemas de localização sao compostas por quatro componentes, quais sejam: Cliente, facilidades que serão instaladas, area onde os clientes estão que deverão ser coberta pelas facilidades e a distância entre as instalações e clientes.

Ferreira e Ferreira (2012) afirma que os problemas de localização é muito discutido na literatura e que existem várias abordagem para determinar uma melhor localização, seja para negócios, imóveis, equipamentos, ou qualquer dispositivo de baixa locomoção que necessitem permanecer em uma dado local por um longo tempo. Ferreira e Ferreira (2012) diz que um dos primeiros trabalhos sobre o tema tratava-se da melhor localização de atividades agrícolas de uma cidade com objetivo de minimar custos com transportes. Oliveira (2020) afirma que os modelos de localização podem ser classificados em três grupos, a depender do objetivo da otimização, são eles:

- modelos de *p-mediana*, que consiste em minimizar a distancia entre a facilidade e o cliente;
- modelos de *p-central*, que tem visa minimizar a distância máxima da facilidade e do cliente;
- modelos de cobertura, que tem como objetivo garantir a cobertura para os clientes dentro de uma distância (raio) pré-estabelecida.

Isler, Bonassa, Cunha (2012) define que o problema de *p-mediana* tem como finalidade identificar p instalações (medianas) em um conjunto de n pontos de

demandas ($n > p$), alocando os pontos remanescente ($n - p$) as médianas, de modo que a soma da distancia, tempo, etc, entre as medianas e os pontos de demanda seja mínimo. Para Lorena e Senne (2005) o problema de *p-mediana* considera decisões ótimas sobre a localização de p instalações levando-se em conta a distância e capacidade de serviço.

A *p-mediana* é o modelo matemático mais conhecido em relação ao problema de localização de instalações que oferecem prestação de serviços, onde existem diversos modelos que podem ser aplicados a situações do mundo real. O modelo *p-mediana* não capacitado é um dos modelos de p-mediana presente a literatura, sendo um modelo que desconsidera as restrições de capacidades das instalações, ou seja, uma unidade instalada atender a qualquer quantidade de clientes, já o modelo de *p-mediana* capacitado absorve restrições de capacidade, onde a capacidade de atendimento da mediana tem que ser respeitada, não permitindo ser extrapolada. (SANTOS et al., 2020).

Muller e Santos (2006) afirmam que o *Set Convering Problem* (SCP) e *Maximum Coverage Location Problem* (MCLP) são as duas versões mais conhecida para o problema de localização.

Cobrir uma região em sua totalidade muitas vezes tonar-se inviável devido limitações de ordem econômica. Nesse sentido, o Problema de Localização de Máxima Cobertura (PLMC), desenvolvido por Church e Revelle (1974), buscar uma solução que proporciona níveis aceitáveis de cobertura, onde localiza um número determinado de facilidades compatíveis com os recursos disponíveis, de tal forma que o máximo de clientes possível de uma região seja coberto a uma distância crítica S predefinida (GALVÃO et al., 2003). Para Pontin et al. (2010) esse modelo não busca atender toda uma população, mas oferecer o máximo de atendimento com os recursos disponíveis. Ele afirma que o conceito de cobertua está relacionado ao fato de se verificar de um determinado ponto está dentro de uma determinada distancia ou tempo até uma facilidade. Church e Revelle (1974) apresenta o modelo matemático formulado para o Problema da Máxima Cobertura. Segue:

$$I. Maximinize \qquad z = \sum_{i \in I} \alpha_i y_i$$

$$S.T. \qquad \sum_{j \in N_i} x_i \geq y_i \qquad para\ todos\ i \in I \qquad (1)$$

$$\sum_{j \in J} x_i = p \qquad (2)$$

$$x_j = (0,1) \quad para\ todos\ j \in J \qquad (3)$$

$$y_i = (0,1) \quad para\ todos\ i \in I \qquad (4)$$

Onde:

I: indica o conjunto de demanda de nós;

J: indica o conjunto de facilidades;

S : indica a distância máxima que um nó estará coberto;

d_{ij}: sendo a menor distância entre os nós i e j;

x_j: 1 se a facilidade for alocada em j e 0, caso contrário;

N_i: $\{j \in J \mid d_{ij} \leq S\}$

α_i: representa a população a ser atendida pelo nó i;

p: representa o número de facilidades a ser alocadas.

Church e Revelle (1974) ratifica que o objetivo do modelo supracitado é maximinizar o número de clientes atendidos ou coberto, dentro de uma distância máxima S de facilidades desejadas. N_i representa o conjunto de facilidades disponível para realizar a cobertura dos conjuntos de nós I.

O modelo apresentado por Church e Revelle (1974) a restrição (1) só permite que y_i seja igual a 1 apenas quando um nó for coberto por uma ou mais facilidades, ou seja, existe pelo menos uma facilidade dentro da distância máxima S. O numero de facilidades disponíveis aos nós é restrito a p, conforme destaca a restrição (2). De acordo com a restrição (3) quando x_j for igual a 1, a facilidade está atendendo a um nó específico, dessa forma, pode-se afirmar, por meio da restrição (4) que um nó só estará coberto, ou seja y_i igual a 1, quando uma ou mais facilidades estiverem dentro de uma distância máxima S, caso contrário o nó em questão está fora da região de cobertura. Por fim a solução supracitada não objetiva apenas na maior quantidade de nós a ser atendidos, mas como também a quantidade de facilidades instaladas.

Pontin et al. (2010) destaca que a localização de hospitais, atendimentos de emergências ou corpo de bombeiros, e a distancia ou tempo de deslocamento entre pontos de demandas e a facilitadas são fatores importantes para estabelecer um nível de qualidade oferecidos aos usuários. Oliveira (2020) destaca que definir a localizaçao de facilidades é um problema crítico nas organizações porque são decisões que refletem em vários outros fatores e aspectos, em nível operacional e logístico, e decisões erradas podem gerar altos cursos e perda de competitividade.

2.3 Clusterização de Dados

O problema de clusterização é definido por Dias (2004) como um processo em que se agrupa elementos de um conjunto em clusters, de modo que cada cluster (ou grupo) represente uma configuração em que cada elemento pertencente a ele possua uma maior similaridade com outros elementos desse mesmo cluster em relação aos demais elementos de outros clusters. Cruz e Ochi (2011) define clusterização como um processo que une objetivos similares em grupos (ou *clusters*). Para Drummond, Ochi, Rosário (2006) clusterização é uma técnica de agrupar dados utilizando critérios como similaridade ou dissimilaridade, podendo seus métodos serem classificados como hierárquico ou não hierárquico. Campello e Hruschka (2006) define clusterização como uma tarefa no qual tem como objetivo determinar um número finito de conjuntos de acordo com as semelhaças de seus objetos.

Dias (2004) e Cruz e Ochi (2011) afirmam que o número de clusters pode ser ou não conhecido, e caso ele seja um dado de entrada da aplicação, a literatura o referencia como “problema de *k*-clusterização”, onde k indicaria o número de clusters a serem formados, por outro lado, caso k não seja fornecido, estariamos diante de um “problema de clusterização automática”, dessa forma, para Cruz e Ochi (2011) k não sendo conhecido trata-se de um problema bem mais complexo o que aumenta substanciamente o número de soluções possíveis.

O problema de clusterização automática é definido matematicamente por Cruz e Ochi (2011) da seguinte forma: $X = \{x_1, x_2, \dots, x_n\}$, sendo X um conjuntos de vários objetos, onde cada objeto x_i é um ponto no espaço R^p, representado por uma tupla $(x_{i1,}x_{i2}, \dots, x_{ip})$, e cada coordenada x_{ij} trata-se de um atributo j do objeto i. Com

objetivo de encontrar o conjunto $C = (C_1, C_2, \dots, C_k)$ automaticamente, onde k (número de clusters) não ser conhecido, considera-se a similaridade de objetos do mesmo clusters seja maximinizada, por outro lado, a similaridade de objetos de clusters diferentes sejam minimizadas, conforme condições expressas abaixo:

$$C_i \neq \boldsymbol{\phi}, para \quad i = 1, \dots, k \tag{1}$$

$$C_i \cap C_j = \boldsymbol{\phi}, para \quad i, j = 1, \dots, k \; e \; i \neq j \tag{2}$$

$$\bigcup_{i=1}^{k} C_i = X \tag{3}$$

Na condição (1) para cada *clusters* C_i criado ele deve conter, obrigatoriamente, um ou mais objeto, bem como na condição (2) que afirma que a interceção $C_i \cap C_j$ de dois *clusters* tem que ser vazio, dessa forma um objeto não pode fazer parte de mais de um *clusters*. Por fim, na condição (3), após a criação automatica de todos os *clusters*, a união deles deve compreender todo o conjunto X e objetos do espaço R^p.

Dias, Ochi, Soares (2004) afirma que no processo de clusterização a busca de melhores soluções dentre as melhores soluções possíveis, é um problema NP-Dificil. Dessa forma, são apresentados na literatura métodos heuristicos com soluções sub-otimas, porem devido a grande diversidades das aplicações que tratam de problemas de clusterização, as heuristicas são desenvolvidas com foco em determinada classe de problema.

A clusterização utilizando método heuristico hierárquico tem como objetivo gerar os *clusters* gradativamente por meio de aglomerações de elementos, onde cada *clusters* com tamanho maior de 1 pode ser considerado como sendo composto por *clusters* menores, conforme pode ser visualizando na Figura 2. Já clusterização por meio do método heuristico não hierárquico o algoritmo divide o conjunto de elementos em *k* conjuntos (podendo *k* ser conhecido ou não), onde cada configuração obtida é submetida a uma dada função, e caso a avaliação da clusterização sinalize que a configuração não atende o problema, nova configuração é obtida realizando a migração de elementos entre os *clusters*. O processo continua, iterando até que uma condição de parada seja atendida. (Dias; Ochi; Soares, 2004)

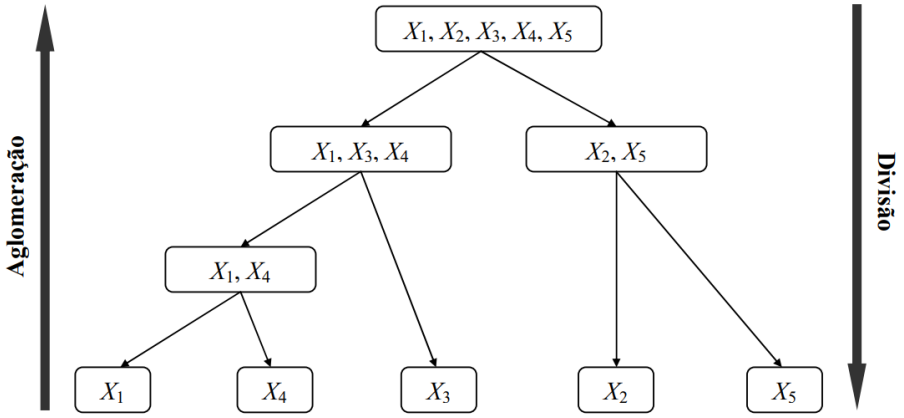

Figura 2. Exemplo de aglomeração de elementos por meio do método heuristíco hierárquico

Fonte: Dias, Ochi, Soares (2004)

O processo de agrupamento, segundo Söküt Açar e Ayman Öz (2020)**,** visa agrupar elementos baseado nas suas características verificando quanto as suas semelhanças e dissimilaridades. Para PRIMEIRO *k-means* é o método mais popular da técnica hierárquica, baseado no algoritmo *sharp set*, onde ele separa os objetos n em grupos k como $S = (S_1, S_2, \dots, S_k)$, e a atribuição dos objetos aos grupos se dão utilizando a média de grupos mais próximos, permitindo apenas que cada objeto pertença a apenas um agrupamento. Chartier e Morissette (2013) apresenta três algoritmos de clusterização baseados no *k-means*, que são os mais usados, com objetivos e resultados poucos diferentes, quais sejam: *Forgy/Lloyd, MacQueen* e o *Hartigan & Wong*.

Para avaliar o desempenho de um algoritmo de clusterização, a literatura dispõe de várias métricas, também chamadas de medidas de validades, que foram propostas. Campello e Hruschka (2006) afirma que dentre as métricas mais difundidas estão o *Fuzzy Hypervolume and Average Partition Desinty,* o *Xie-Beni index, Average Within-Cluster Distance* e o *Average Silhouette Width Criterion*. Campello e Hruschka (2006) destaca que essas medidas de validades de *clusters* mencionadas possui características particulares que quando aplicadas em determinadas classes de problemas podem se comportar de forma de uma superar a outra, alternando seus resultados quando na mudança da classe do problema.

Söküt Açar e Ayman Öz (2020) apresenta o índice *Silhouette (SIL)* como sendo uma das formas de validar um *cluster,* de bom desempenho, aplicado a várias medições de distância, cujo objetivo é determinar a adequação de cada objeto a um *cluster.* Vejamos:

$$Sil(x_i) = \frac{b(x_i) - a(x_i)}{maks(a(x_i), b(x_i))}$$

Conforme definição presentada por Söküt Açar e Ayman Öz (2020), $a(x_i)$ representa a dissimilaridade média do objeto i para todos os outros objetos s no mesmo *cluster*, e $b(x_i)$ como sendo a mínima dissimilaridade média do objeto i para todos os objetos em *cluster* mais próximo. $Sil(x_i)$ fornece a desigualdade de $-1 \leq Sil(x_i) \leq 1$, indicando que quando seu valor está próximo de 1 significa que o objeto i está bem classificado. Quando $Sil(x_i)$ está próximo de 0 pode-se afirmar que o objeto i_n está entre dois *clusters*, e quando $Sil(x_i)$ está próximo de -1 indica que o objeto i_n está classificado incorretamente. Para o autor, a média de $Sil(x_i)$ obtida para o número de *clusters* relevantes indica sua validade, de forma geral, se o seu valor médio estiver acima de 0,5, então os *clusters* esperados serão alcançados.

2.4 Business Intelligence

Business Intelligence (BI), ou Inteligência de Negócio, é definida por Braghittoni (2015) como sendo um conjunto de conceitos e métodos que tem como objetivo melhorar, no âmbito da gestão, a tomada de decisão, utilizando-se de sistemas que tenha como arquitetura estrutural a base de fatos e dimensões, para aglutinar e processar os dados de um domínio específico. É importante destacar que o BI não é uma ferramenta, e sim uma forma de manipular um conjunto de dados para obter vantagem de gestão, e pode ser implementado com várias ferramentas.

Atualmente as organizações estão sendo pressionadas a agir conforme a evolução, e para isso elas tem que inovarem para responderem as demandas de tomadas de decisões operacionais frequentes, onde essas demandas devem ser respondidas rapidamente, e em alguns casos até em tempo real, porém devido a grande quantidade de dados e informações, há necessidade de uso de tecnologia computadorizada (RAMESH, 2019). O BI tem como foco especificamente ajudar o gestor da tomada de decisão, podendo, através da análise dos dados entender fenômenos ocorridos, assim como realizar um planejamento estratégico mais fidedigno.

Os principais objetivos de um sistema de BI, conforme Ramesh (2019), são:

- Permitir o acesso interativo dos dados, em muitas vezes em tempo real;
- Proporcionar a manipulação de dados e fornecer ao gestor e analistas de negócio a capacidade de realizar a análise adequada;
- Ao analisar os dados, situações e desempenhos históricos e atuais, os tomadores de decisões conseguem valiosos *insights* que podem servir como base para decisões melhores e mais informadas.

Para um melhor entendimento no contexto do BI, é importante a definição de três conceitos relacionado, que são: dado, informação e conhecimento. Dado é definido de forma simples e objetiva como sendo o registro de um evento, porém, expressado sozinho não tem significado. A Informação pode ser definida como um conjunto de dados relacionados e contextualizado, com um significado expressivo. Por fim, o Conhecimento é a interpretação da informação, com resultado de ações e tratamento específicos, que subsidie a condução de algo de forma mais assertiva (FACELI et al., 2011).

A grande maioria dos sistemas de informações que comumente está presente no dia-a-dia na sociedade são chamados de sistemas transacionais, que dão suporte, em nível operacional, as empresas e seus negócios, a exemplo de sistemas que registram vendas em supermercados, que realizam saques bancários, pagamentos de contas etc, e consequentemente alimentam diariamente uma base de dados denominada de base de dados operacionais. Os sistemas transacionais são denominados de OLTP (*online transaction processing*). Diferente dos sistemas OLTP, os sistemas de processamento analíticos, denominado de OLAP (*online analytical processing*) são bem mais incomuns de serem encontrados, porém é de uma grande relevância, tendo em vista ser capaz de proporcionar aos gestores do negócio a possibilidade e examinar dados em busca de informação e conhecimento a respeito do negócio para embasar decisões gerenciais.

Os sistemas de Business Intelligence devem ser desenvolvidos utilizando uma metodologia correta para que seja eficaz e cumpra o seu papel. Para Primak (2008), seis passos devem ser seguidos para que um BI seja implantando de forma correta. São eles:

- Identificar as necessidades a serem endereçadas na solução de BI, onde estas devem ser relevantes aos objetivos e estratégias do negócio;

- Identificar as fontes de dados já existentes na organização. Normalmente as organizações têm várias informações em bancos de dados, planilhas e arquivos, e em muitos casos é necessário criar mais informações, porém, é de extrema importância mapear aquelas já existentes;

- Extrair, transformar e carregar os dados para criar uma base de dados multidimensional orientada por assunto. Este processo deve garantir que todas as informações relevantes sejam contempladas consistentes;

- Ajudar a organização a escolher a ferramenta de apresentação para visualizar e analisar as informações resultantes do item anterior;

- Criar relatórios padrões, permitir análise sob demanda e mineração de dados (Data Mining) visando obtenção de *insights* sobre os indicadores chaves de desempenho;

- Planejar uma implantação de forma abrangente para toda a corporação, de forma a garantir que os tomadores de decisão tenham a informação adequada quando e onde eles precisarem.

Algumas ferramentas fazem parte do projeto de um sistema de BI, dentre elas estão o *Data Wharehouse (DW)*, *Data Mining (DM)* e a ferramenta de OLAP. Os DW e DM são ferramentas especiais de armazenamento de dados, responsável por dar sustento a camada de inteligência de negócios (PRIMAK, 2008). Os dados presentes nessas estruturas estão modelados utilizando metodologia de componentes de Fato e Dimensão. No *Data Warehouse* estão contidos os dados organizados por assunto e por data, já os *Data Mining* contém um subconjunto lógico e físico do DW, pertencentes a uma área específica de uma organização, por exemplo: finanças, vendas, contabilidade, marketing etc (BRAGHITTONI, 2015). Braghittoni (2015) define Data Warehouse como um conjunto de dados orientado por assunto, conciso e integrado, variável com o tempo e não volátil. Já as ferramentas de OLAP (Online Analytical Processing) trabalham com os dados, dos DM, com operadores dimensionais, permitindo dessa forma uma abordagem múltipla e combinada de análise. Nesse cenário conclui então que a dinâmica para a construção do BI é realizada com a obtenção dos dados dos sistemas transacionais presentes nas organizações, posteriormente esses dados são submetidos ao processo de extração,

transformação e carregamento (ETL), gerando dessa forma os DW e DM, modelados na perspectiva de Fato e Dimensões.

O primeiro passo de um projeto de BI é a criação do Data Warehouse (DW) responsável por conter os dados das ocorrências que serão manipulados. O DW é o coração do projeto de BI. Ele é criado em um banco de dados relacional como qualquer sistema transacional (OLTP), mas com um formato que possa responder as pesquisas de forma mais performática possível. É importante destacar que as ferramentas de BI podem trabalhar consultando os dados que estão armazenados nas bases dos sistemas de uso operacional, porém, não é a forma indicada e adequada, isso por vários motivos, inclusive de performance, que pode afetar a rotina diária no ambiente operacional. O DW ele segue uma estrutura arquitetural, regras de design e boas práticas específica de sua metodologia. Ele é baseado em uma arquitetura dimensional, com foco nos fatos e dimensões. Dessa forma, teremos as tabelas referentes as dimensões e a tabela de fato. As tabelas de dimensões deverão ser as primeiras a serem criadas, isso porque elas representam as chaves primárias dos dados que estão dentro da tabela fato que armazenarão as chaves estrangeiras.

Por fim, quando pensamos em sistemas transacionais (OLTP), aqueles que estão presentes com mais frequência em nosso dia-a-dia, realizando as atividades operacionais das organizações, ou seja, vendas, consultas médicas, saques bancários, e que os dados registrados por eles vão dar o suporte aos sistemas de BI, a forma que os dados registrados por eles são organizados são em estruturas denominado de ER (entidade x relacionamento), onde utilizam-se de tabelas formadas por linhas (que representam os registros) e colunas (que representa os atributos), onde cada tabela representa uma entidade existente naquela sistema, e que se relacionam entre si. Já quando falamos em sistemas analíticos (OLAP), que são as ferramentas de BI, os dados são modelados utilizando-se o conceito de Fato e Dimensão.

2.5 Segurança Pública e Análise Criminal

A segurança pública é um dos serviços mais relevantes presentes na gestão pública de uma sociedade moderna organizada. Para Carvalho (2011) a segurança da sociedade é o principal requisito à garantia de direitos e ao cumprimento de deveres estabelecidos nos ordenamentos jurídicos. Os Estados são responsáveis, por meio de

seus órgãos ou instituições, de implementar ações que garanta a segurança da população, por meio de políticas públicas, definindo ações em planos e programadas implementadas para garantir a segurança individual e coletiva dos indivíduos.

Em Carvalho (2011) constata que a segurança vive no limite de crises decorrentes problemas estruturais, uma mídia pesada utilizada para denegrir sua imagem, e quanto mais a criminalidade aumenta os Estados não consegue conter e controlar os problemas relacionados à segurança. Em contraste a esse cenário estatal temos um sistema organizado e controlado por criminosos, denominado de Crime Organizado. Em Leal (2016) o crime organizado é definido como um empreendimento complexo, centralizado, constituído para a prática de atividades ilegais. É possível identificar no cenário supracitado, onde o Estado demonstra diariamente sua precariedade e atuação pífia no âmbito da segurança pública, e por outro lado, o crime organização em ascensão.

A gestão da segurança pública nos Estado e no Governo Federal carece de um melhor planejamento, monitoramento, avaliação de resultados, eficiência dos recursos financeiros (LEAL, 2016). É impossível combater o crime organizado sem ter o mínimo de organização estatal, sem ter investimentos estrutural e principalmente intelectual. A profissionalização da gestão da segurança pública é fundamental para mitigar a criminalidade.

Atualmente o sistema de segurança pública dos Estados são formados por forças policiais, quais sejam: Polícia Militar, Polícia Civil e Polícia Penal. A Polícia Militar é a instituição responsável por manter a ordem pública por meio da atuação de policiais fardados, ostensivos, para evitar a atuação criminosa, e quando essas ações ocorrem, eles são responsáveis por agirem de imediato ao momento do crime, em busca de capturar os responsáveis e apresentá-los a polícia judiciária para realização dos trâmites pertinentes. A Polícia Civil, instituição estatal e denominada de polícia judiciária, tem como função a manutenção da ordem pública e atuação, especificamente, em elucidações de crimes ocorridos, produzindo e formalizando as provas e materialização dos delitos, posteriormente submetendo os procedimentos para o sistema judiciário. Por fim a Polícia Penal é responsável pela custódia de pessoas acusadas e sentenciada, garantindo a integridade física e mental, além de apuração de delitos ocorridos dentro das estruturas físicas dos presídios. Equipar e modernizar esses órgãos, seja estruturalmente como intelectualmente por meio de seus agentes, é primordial para o combate à criminalidade. Nesse contexto, é

imprescindível o investimento em inteligência policial para uma maior e melhor eficácia das estruturas policiais estaduais.

Para se combater o crime é necessário investimento em inteligência policial, por meio da Análise Criminal (AC). A análise criminal é responsável por produzir conhecimento policial, onde se avalia e acompanha sistematicamente o fenômeno da criminalidade em massa, instrumentando os operadores da segurança pública e proporcionando um suporte administrativo, tático, estratégico para a prevenção e repressão do crime e da violência (FERRO, 2006). De acordo com a Enciclopédia de Ciências Policial:

> A Análise Criminal é a coleta e o manejo de informações relacionadas ao crime e com dados para discernir padrões dentro desses mesmos dados, com o objetivo de prever, compreender, ou explicar empiricamente o crime e a criminalidade, para avaliar o desempenho de uma agência de justiça criminal do governo, ou para estabelecer o desdobramento tático e estratégico dos recursos humanos de tais agências. (Lum e Greene, 2007)

Para que a AC seja realizada de forma eficiente o estado tem que utilizar, principalmente, a tecnologia a seu favor, isso porque o grande fluxo de informações e variáveis em estão envolvidos diretamente e indiretamente, no contexto da segurança pública, são enormes e impossível a percepção, processamento e ação por parte dos agentes de segurança sem o uso de tecnologias que lhe deem apoio. Para Ferro (2006) a atividade de Inteligência em Segurança Pública é apontada no Brasil como um instrumento fundamental para o enfrentamento a criminosos, fato esse que só cresce. Considerando a definição de Inteligência como atividade policial que tem como finalidade obter, analisar e disseminar conhecimento entre os atores interessados, os gestores, em seus processos decisórios, precisam estar subsidiado com conhecimento que esteja em consonância com a realidade vivida naquele ambiente/território de gestão, ou seja, alguns fenômenos são inerentes a determinadas regiões, para realizar uma repressão a criminalidade mais eficaz.

Para entender a relevância do tema, o Decreto nº 3.695 de 21 de dezembro de 2000, em seu Art. 1º, parágrafo 2º, define o termo inteligência no âmbito do território brasileiro, de forma e uso geral, e por conseguinte, o Decreto nº 3.695 de 21 de dezembro de 2000, em seu Art. 1º cria o Subsistema de Inteligência de Segurança Pública, definindo a Inteligência no contexto da Segurança Pública, qual seja, a

Inteligência de Segurança Pública é um sistema que obtém, produz, sistematiza, analisa e dissemina conhecimento que subsidiem os governos na tomada de decisões no que tange a Segurança Pública (Williams e Bailey, 1995).

Em Leal (2016) é definido Inteligência Policial como a produção de conhecimento com objetivo de suportar a tomada de decisões nos níveis políticos estratégico, tático e operacional, ratificando que ela atua nos seguintes objetos: crime, criminoso, criminalidade e questões conexas. O Subsistema de Inteligência de Segurança Pública é importante para fazer a paridade ao crime organizado, uma vez que o crime organizado já faz um uso, de forma cirúrgica e tecnológica, de conhecimento a seu favor.

Para a produção de inteligência policial existem duas vertentes por parte das instituições: ir coletar os dados ou ir buscar os dados. A ação de coletar dados está relacionada ao fato das instituições coletarem dados em fontes de dados abertas e públicas, por exemplo, quando utilizam de dados públicos postados nas redes sociais para obter conhecimento sobre um determinado criminoso. Por outra lado, a ação de buscar dados é quando, por exemplo, acontece uma interceptação telefônica, infiltração policial dentre outros, sendo uma atividade mais complexa e que muitas vezes beiram os direitos fundamentais das pessoas.

Capítulo

3

Trabalhos Relacionados

Nesse capítulo serão expostos alguns trabalhos relacionados que explora o uso de ferramentas de Business Intelligence, Séries Temporais, Previsão de dados, Clusterização e Problema de Localização, relatando seus objetivos e resultados.

O estudo de Ribeiro, Oliveira, Pedrosa (2021) destaca a importância do uso de Business Intelligence (BI) na Administração Pública Portuguesa para analisar a eficiência, desempenho de serviços e prevenção de fraudes. O trabalho avalia e expõe vários projetos de Business Intelligence aplicado a Administração Pública que contribuiu para melhoria de desempenho nas atividades apoiada pela ferramenta, ratificando que o BI aplicado de forma correta trás grandes benefícios a Administração Pública, como também relatando que BI ainda é pouco utilizado no cenário público. O estudo considerou 15 (quinze) trabalhos encontrados nos repositórios Scopus e B-on, todos eles relacionados a administração pública e uso de Business Inteligence.

Para Khan et al. (2020) os sistemas de Business Intelligence têm atraído a atenção das empresas, profissionais e pesquisadores, proporcionando, dentro das organizações, uma melhoria no processo de suporte a tomada de decisões, assim como no desenvolvimento de produtos e serviços. Destaca a grande quantidade de dados geradas no mundo contemporâneo e a necessidade de habilidades específica para análise desses dados. Fornece uma pesquisa relatando o impacto do uso de BI nas organizações.

Milán et al. (2020) apresenta uma revisão bibliográfica sobre o estado da arte do uso de Business Intelligence para o gerenciamento de performance nas corporações. Destaca que devido a grande quantidade de informações provenientes de vários tipos de fontes, internas e externas, tem dificultado o processo de tomada de

decisão dentro das empresas, como também a medição de desempenho das empresas. Para Milán et al. (2020) o BI tem se mostrado uma ferramenta eficaz para resolver esse cenário mencionado, proporcionando rapidez ao acesso da informação por meio de dashboards, assim como seu uso pode proporcionar melhores resultados nas decisões corporativas.

Shi (2013) afirma a necessidade urgente do uso de sistemas de Business Intelligence (BI) para a tomada decisão no âmbito científico. O trabalho expõe um estudo realizado em uma empresa que utiliza BI na manutenção de energia nuclear, mostrando a importância dessa ferramenta na melhoria empresarial e realizando análise de três propostas de aplicações de BI.

Com foco na melhoria do desempenho da política energética e climática da União Europeia, Pinheiro et al. (2020) implementou uma solução de Business Intelligence, por meio de criação de Dashboard, para analisar a eficiência em Edifícios Habitacionais. Por meio de análise histórica e projeções até o ano de 2035, concluiu-se uma ineficiência energética nos edifícios habitacionais. A solução também proporcionou uma redução do valor da fatura de eletricidade e emissão de gases de efeito estufa.

Para Taylor e Letham (2017) a tarefa de prever, ou de realizar previsão, é comum na ciência de dados e ajuda as empresas nos seus planejamentos, definição de metas e detecção de fenômenos, porem muitos desafios estão associados a essa tarefa. Eles apresentam uma abordagem prática para realizar previsões em escala, por meio de um modelo de regressão modular parametrizados que podem ser ajustados intuitivamente por especialistas, e apresentam uma análise de desempenho de comparações de procedimentos de previsão para um melhor ajuste manual. Taylor e Letham (2017) ratificam que ferramentas que ajuda especialistas a usarem suas experiências de forma mais eficaz resultam em previsões mais confiáveis de séries temporais.

O estudo apresentado por Santos e Muller (2006) tem como objetivo apresentar o Problema de Localização de Máxima Cobertura com algumas adaptações, sugerindo novas restrições, para suprir a necessidade de cobertura de uma grande área geográfica de aproximadamente 5,08 milhões de quilômetros quadrados que tem que ser monitorados por facilidades, que possam atender aos clientes com maior rapidez. O estudo concluiu que sob determinadas circunstâncias as

alterações propostas permitiram um maior realismo no modelo matemático utilizado em relação ao modelo do Problema de Localização de Máxima Cobertura original.

Oliveira (2020) apresentou a importância de um rápido atendimento a população por parte do Serviço de Atendimento Móvel de Urgência (SAMU), em que o tempo de resposta desse serviço é relevante para o atendimento ser bem-sucedido, e para isso é importante que as viaturas do SAMU estejam bem localizadas ao ponto de atender a maior quantidade de usuários. O trabalho tem como objetivo fornecer um conjunto de ferramentas que possa possibilitar um melhor posicionamento de viaturas do SAMU aplicando o Problema de Localização de Máxima Cobertura.

Sena e Nagwani (2015) enfatiza que a análise de serie temporais é uma das principais técnicas de previsão. O trabalho apresenta um estudo onde é aplicado o modelo ARIMA sobre a renda *per capita* na Alemanha Ocidental. A renda *per capita* é o valor recebido por pessoa após a dedução de impostos e sua previsão ajuda ao governo a avaliar as condições econômicas atual do país em comparação a outras economias do mundo, como também ajuda a avaliar a inflação. O trabalho desconsiderou os ajustes sazonais e foi obtido um resultado com um erro relativo de 0,0225. Como o model ARIMA são baseados em autocorrelações, o preço atual previsto está linearmente relacionado ao preço anterior.

Ahmed et al. (2020) utilizam o Prophet, que é baseado em um modelo de aprendizado de máquinas, para prevê series temporais de energias de painéis fotovoltaico. A previsão confiável desses dados torna-se essencial para o planejamento da capacidade, em antecedência, para gerenciar com eficiência a distribuição de energia. Os dados gerados pelos painéis são submetidos ao modelo Prophet um dia antes para ser realizada a previsão de saída de energia. Concluiu-se que os dados coletados foram bastantes confiáveis.

Darapaneni et al. (2021) realizou um estudo referente a uma campanha de vacinação, na Índia, onde tinha como objetivo prever o tempo mínimo necessário para vacinação da população para alcançar a imunidade de rebanho. Nesse estudo foi utilizado modelo SIR, que define a capacidade de disseminação da doença, e posteriormente, por meio de análise de séries temporais usando Prophet foi possível prever a quantidade de dias necessários para vacinar a população suficiente para atingir a imunidade de rebanho.

No estudo apresentado por Jain e Prasad (2020) é relatado a importância do controle da medição de alguns parâmetros referente ao tráfego de redes utilizado na área de telecomunicação para prestar um serviço de melhor qualidade, permitindo que seja feito planejamento adequado os ativos de redes. O trabalho apresenta uma solução para esse cenário de previsão de parâmetros de performance de redes utilizando series temporais com o Prophet e o algoritmo XGBoost. É relatado pelos autores que o Prophet é um algoritmo capaz de realizar previsão utilizando grande quantidade de dados irregulares, ou anormal.

Kumar e Pande (2021) destaca que técnicas de previsão de dados são utilizadas em vários âmbitos, a exemplo de vendas, bancos, saúde, mercado de ações etc, para resolver os diversos problemas relacionados, pois prever ajuda na tomada de decisão. O trabalho destaca algumas ferramentas disponíveis com esse propósito, utilizando modelo de regressão e modelo exponencial logístico. É examinado alguns modelos de previsão, tais como o modelo aditivo, o modelo autorregressivo de média móvel integrada (ARIMA) e o modelo Prophet. O trabalho concluiu que o Prophet foi o modelo com melhor desempenho, pois apresentou um baixo índice de erro, melhor previsão e melhor ajuste.

Gong et al. (2020) apresentou estudo onde utilizou os algoritmos Prophet e ARIMA para prever consumo de energia elétrica em prédios, que pode variar de acordo com época do ano, economia, tipo de estabelecimento comercial, etc. O estudo apresenta uma comparação, em uma janela de tempo específica, dos dois algoritmos, de dados de series temporais reais de consumo de energia elétrica e dados gerados pelos Prophet e ARIMA. Com base na taxa da média relativa de erro com o resultado gerado pelos algoritmos, o autor concluiu que o Prophet foi o que apresentou o melhor resultado, em relação ao ARIMA.

Silva (2014) apresenta em seu trabalho uma solução para posicionamento de viaturas policiais para um atendimento a população com maior eficiência e eficácia. O estudo, por meio das localizações geográficas de crimes ocorridos na cidade de João Pessoa, identificou áreas crônicas com maior intensidades de crimes, através de uma maior concentração de ocorrências, e assim permitindo ao gestor uma melhor análise do problema e consequentemente a definição de estratégia tática e operacional mais adequada. Foram utilizadas técnicas como *k-means* e p-mediana para atingir os objetivos estabelecidos, assim como a implementação de duas meta-heurísticas, a

GRASP combinado com VND e GRASP com um modelo exato, no qual obteve resultados muito semelhantes.

Nesse cenário, destacando-se a importância dos temas supracitados, esse trabalho submete os dados de ocorrências policiais a um processo com objetivo identificar regiões através de técnicas de clusterização, criar séries temporais para prever dados de ocorrências por meio da API *Prophet* e definir a localização de patrulhas policiais para um melhor atendimento a população empregando o modelo matemático baseado no Problema de Localização de Máxima Cobertura Capacitada, dispondo todos esses dados em plataforma de Business Intelligence que irá subsidiar os gestores da área de segurança pública.

Capítulo

4

Metodologia

Com o objetivo propor uma melhor gestão dos recursos policiais no combate efetivo à criminalidade, a presente seção apresenta a metodologia utilizada para o desenvolvimento desse trabalho, que é baseada em três pilares, que juntos compõe a solução proposta, quais sejam:

1) Clusterização de dados de ocorrências policiais resultando na identificação de regiões e sua incidência criminal;

2) Previsão da quantidade de crimes por dia e semana nas regiões identificadas e bairros;

3) Alocação inteligente de quantidade e localização de viaturas policiais para melhor atender as regiões/bairro e seus respectivos crimes.

Inicialmente foram coletados dados de registros de ocorrências policiais gerados por meio do Sistema de Procedimento Policial (SPP), que se trata de um software presente nas unidades policiais que registram os crimes sofridos pela população, materializado por meio de Boletim de Ocorrências (B.O). Os dados foram coletados em formato tabular, contendo 25 (vinte e cinco) colunas e mais de 300 mil registros de ocorrências policiais que ocorreram na cidade de João Pessoa-PB no intervalo de tempo de janeiro de 2017 à dezembro de 2021, ou seja, compreendendo um período de 5 (cinco) anos. Cada registro de ocorrência representa um crime ocorrido.

Entre os dados coletados existem informações referentes ao crime perpetrado e da vítima, e principalmente do dia e local do delito. A Figura 3 ilustra como os dados foram dispostos.

	Procedimento	Ocorrên	data_registro	data_ocorrencia	Cidade	Bairro	ipo do Loc	Rua	hicial Ocor	Final Ocor	natureza	tipificacao	ualificaçã
2	101201760	BO	2017-01-02 00:	02/01/2017	JOAO PES	VALENTIN	OUTROS	RUA CELIT	10:30	10:30	FATOS AT	PERDA OU	VITIMA
3	142012017	BO	12/02/2017	12/02/2017	JOAO PES	CABO BRA	VIA/LOCA	AV CABO	08:10	08:10	CRIMES C	ART. 129 C	VITIMA
4	103701201	BO	12/02/2017	12/02/2017	JOAO PES	VALENTIN	OUTROS	RUA VICE	13:00	13:00	FATOS AT	OUTROS F	VITIMA
5	182012017	BO	12/02/2017	12/02/2017	JOAO PES	CUIA	RESIDENC	RUA PRAI	12:00	12:00	LEI 11.340,	ART. 7º, IN	VITIMA
6	790012017	BO	12/02/2017	12/02/2017	JOAO PES	JAGUARIB	OUTROS	RUA ABER	17:32	17:32	FATOS AT	PERDA OU	VITIMA
7	791012017	BO	12/02/2017	12/02/2017	JOAO PES	VALENTIN	VIA/LOCA	R.TEN. ALI	18:10	18:10	FATOS AT	OUTROS F	COMUNIC
8	156801201	BO	12/02/2017	12/02/2017	JOAO PES	JOSE AME	VIA/LOCA	RUA RADI	18:40	18:40	CRIMES C	ART. 157 C	VITIMA
9	795012017	BO	12/02/2017	12/02/2017	JOAO PES	VALENTIN	VIA/LOCA	R. JUIZ AR	21:45	21:45	CRIMES C	ART. 157 C	COMUNIC
10	187012017	BO	12/02/2017	12/02/2017	JOAO PES	MANGABE	RESIDENC	RUA	20:30	20:30	LEI 11.340,	ART. 7º, IN	VITIMA
11	157301201	BO	13/02/2017	13/02/2017	JOAO PES	VALENTIN	ONIBUS IN	NAO INFC	07:30	07:30	CRIMES C	ART. 155 C	VITIMA
12	552012017	BO	13/02/2017	04/12/2016	JOAO PES	VALENTIN	VIA/LOCA	NAO INFC	08:00	08:00	FATOS AT	PERDA OU	VITIMA
13	158301201	BO	13/02/2017	10/02/2017	JOAO PES	VALENTIN	ORGAO PU	NAO INFC	17:00	17:00	CRIMES C	ART. 157 C	VITIMA

Figura 3 - Dados originais na forma que foram coletados

De posse da base de dados de ocorrências policiais, foi possível observar algumas inconsistências e características que precisavam ser ajustadas para um melhor aproveitamento na solução proposta, a exemplo de campos vazios, formatos de dados incompatíveis, duplicação de dados, dados desnecessários, dentre outros, além da necessidade da inclusão de dois campos fundamentais que serão necessários para a aplicação da solução proposta: latitude e longitude.

Essa fase é denominada de Extração, Transformação e Business Intelligence (BI), onde foram realizadas as intervenções necessárias nos dados. Ainda nesta fase foram incluídos na base de ocorrências policiais os campos latitude e longitude. Para fins de preenchimento desses novos campos, foi utilizado a o serviço de geolocalização da *API* do Googolemaps que, por meio dos endereços presentes nos registros de ocorrências, foi possível identificar a latitude e longitude do local onde ocorreu o fato criminoso.

Com a fase de Extração, Transformação e Business Intelligence concluída, os dados estavam prontos para serem aplicados na solução composta pelos pilares definidos inicialmente.

O primeiro pilar trata-se do procedimento de clusterização. Por meio dos dados do dia e local exato onde os crimes foram cometidos e com objetivo de identificar as regiões e seus índices criminais, esses dados foram submetidos a um algoritmo de clusterização. O processo de clusterização identifica e agrupa a maior quantidade de pontos semelhantes, nesse caso representados pelas coordenadas do local do crime (latitude e longitude), e resultando na criação de regiões (*clusters*). Com essas regiões identificadas é possível visualizar as áreas com maior incidência criminal para uma maior atuação policial.

Na Figura 4 (a) é possível observar os locais de ocorrências policiais antes do processo de clusterização, onde todos os pontos estão dispostos de forma homogênea, já a Figura 4 (b) representa o resultado da clusterização dos dados e suas regiões (*clusters*) resultantes, propiciando de imediato uma melhor identificação com a delimitação formada.

Após a clusterização cada registro de ocorrência policial presente na base de dados submetida ao processo passou a ser vinculado a uma região resultante, consequentemente se fez necessário a criação de um novo campo na base de dados, denominado *cluster,* que foi preenchido com o número da respectiva região a qual a ocorrência pertencia.

O próximo passo dentro da solução proposta trata-se do segundo pilar, responsável pela predição de crimes.

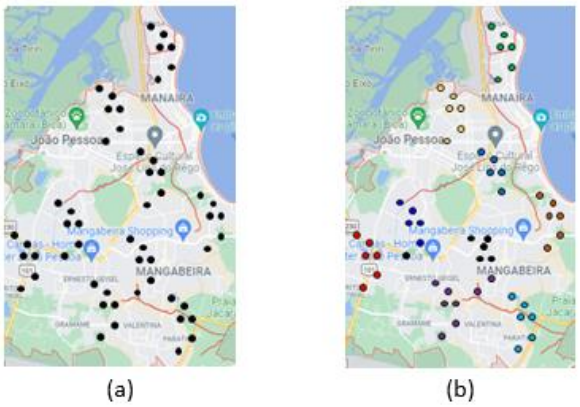

Figura 4. Fase de clusterização dos registros criminosos. (a) Localização dos registros. (b) A grupamentos dos registros por proximidade.

De posse dos dados dos dias e regiões resultante do processo de clusterização, foram construídas suas respectivas séries temporais. As séries temporais foram formadas dispondo os dados por região, dia e quantidade de crimes. É possível observar na Figura 5 as regiões e quantidade de crimes presentes em um determinado dia.

Figura 5. Quantidade de crimes registrados em um dia para cada região criminal em dia especifico.

Com as séries temporais definidas, o próximo passo foi realizar a previsão da quantidade de crimes que poderiam ocorrer nas regiões identificadas por meio dessas séries. O processo de predição torna-se primordial para uma atuação mais assertiva no sentido do emprego de recursos policiais para o combate efetivo à criminalidade, tendo em vista ser mais apropriado ser demandado um maior efetivo policial para áreas com maior quantidade de crimes. Dessa forma, por meio do algoritmo de previsão *API Prophet*, foi possivel prever a quantidade de crimes por região (*cluster*), bairro e seus respectivos dias.

A *API Prophet* recebeu como entrada séries temporais de dados históricos por meio de dois campos, quais sejam: data e quantidade, representado pela linha azul do gráfico da Figura 6. Após a realização do treinamento, o *Prophet* apresenta uma previsão de uma série temporal de dias predefinidos e quantidades futuras, conforme ilustrado na linha vermelha do mesmo gráfico.

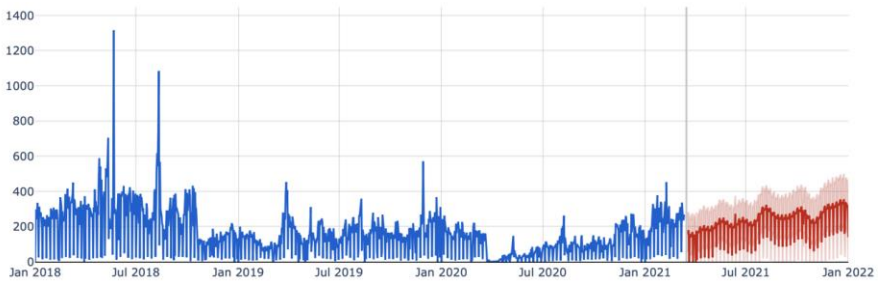

Figura 6. Previsão dos crimes em uma região baseado no seu histórico criminal.

O terceiro e último pilar da solução proposta trata-se da alocação inteligente de quantidade e localização de viaturas policiais nas regiões já definidas. Na fase anterior foi possível prever a quantidade de crimes em cada uma dessas regiões e bairros por dia. Com base nesses dados é necessário definir a quantidade de viaturas (facilidades)

necessárias e suas respectivas localização para que possam atender as ocorrências com maior rapidez e cobrir as áreas identificadas.

Com esse propósito, com as áreas e suas previsões de quantidade de ocorrências policiais já identificadas em dias específicos, estes dados foram submetidos à um modelo matemático, proposto nesse trabalho, com o objetivo de cobrir o máximo de ocorrências policiais por cada viatura, dentro de um raio de distância pré-definido, e uma capacidade máxima de ocorrências a serem atendidas. A Figura 7 ilustra uma possível solução de cobertura das viaturas proposta pelo modelo.

Figura 7. Solução do modelo proposto de localização de viaturas policiais para melhor atender as regiões

As próximas subseções descreverão em detalhes as implementações destas fases levantadas. A sessão 4.1 abordará os procedimentos aplicados para preparar e transformar os dados de forma que subsidiem esse trabalho. Nas sessões 4.2, 4.3 e 4.4 será detalhado, respectivamente, todo o processo referente a clusterização, séries temporais, previsão e o Problema de Localização de Máxima Cobertura Capacitado.

4.1 Extração, Transformação e Business Intelligence

Nessa fase denominada de Extração, Transformação e Business Intelligence (B.I) os dados recebidos foram submetidos a um tratamento preliminar, nos quais foram corrigidas inconsistências assim como realizados ajustes necessários para que esses dados ficassem prontos para as fases seguintes. Na Figura 8 é possível visualizar o fluxo que representa a fase.

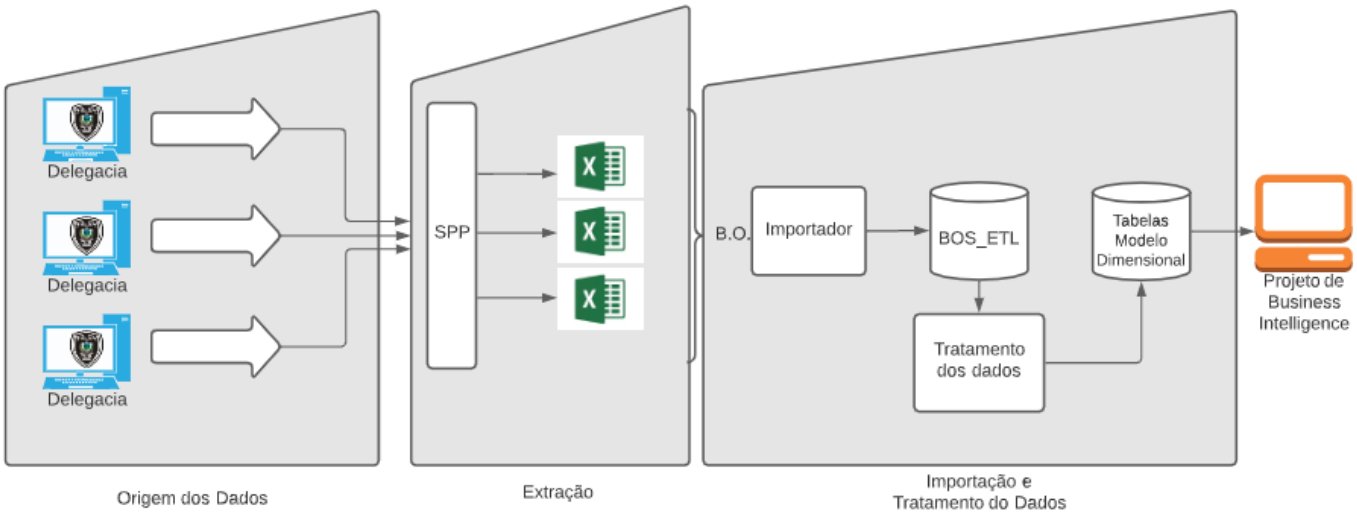

Figura 8. Fluxo referente aos processo de Extração, Transformação e Business Intelligence

Nesse cenário, com objetivo de subsidiar as fases principais que compõe esse estudo, os dados utilizados foram extraídos do software denominado Sistema de Procedimento Policial (SPP), que se trata de um sistema gerencial presente e utilizado nas unidades policiais que registra todas as ocorrências de crimes informadas pela população. O SPP possui uma modelagem de dados própria, que não é discutida porque está fora do escopo desse trabalho, dessa forma, para fins desse estudo, foram extraídos do SPP apenas alguns dados selecionados.

Os dados foram fornecidos em formato tabular, em arquivo no formato .xlsx, contendo 25 campos e mais de 300 mil registros de ocorrências de crimes ocorridos na cidade de João Pessoa-PB entre o intervalo de tempo de janeiro de 2017 a dezembro de 2021, compreendendo um período total de 5 últimos anos. Os campos que compõe os dados são: número do procedimento policial, tipo do procedimento policial, data do registro, data da ocorrência, cidade (vítima), bairro (vítima), tipo de local (fato criminoso), rua (vítima), hora inicial (fato criminoso), hora final (fato criminoso), natureza (fato criminoso), tipificação (fato criminoso), qualificação, data de nascimento (vítima), estado civil (vítima), sexo (vítima), orientação sexual (vítima), gênero (vítima), nacionalidade (vítima), naturalidade (vítima), raça/cor (vítima), profissão (vítima), logradouro (fato criminoso), cidade (fato criminoso) e bairro (fato criminoso).

O primeiro passo realizado nos dados foi uma inspeção sistemática com objetivo de identificar inconsistências e/ou ajustes que fossem necessários para

adapta-los ao modelo de solução proposta. Foram identificadas as seguintes inconsistências/ajustes:

- Dados desnecessários para o escopo do trabalho;
- Formato de dados incompatíveis;
- Redundância de dados;
- Campos vazios;
- Ausência de dados relevantes para aplicação;
- Formato e modelagem de dados ineficiente.

Após identificar as inconsistências e ajustes necessários, algumas medidas foram executadas para deixar os dados compatíveis para solução. Primeiramente foram eliminados os campos não utilizados: número do procedimento policial, tipo, data do registro, cidade (vítima), bairro (vítima), rua (vítima), qualificação, data de nascimento (vítima), estado civil (vítima). No total, foram excluídos 9 campos dos 25 campos inicialmente presentes no arquivo.

Os registros de ocorrências criminais que apresentaram dados vazios também foram desconsiderados. Tal procedimento se fez necessário porque não poderíamos definir valores aleatórios nesses casos para não distorcer a realidade dos fatos, sendo mais coerente retirar os registros da base de dados.

Dando prosseguimento as próximas etapas, com os dados restantes se fez necessário converter sua modelagem, que se apresentava de forma desestruturada e em formado .xlsx, para um formato mais adequado e para melhor manipulação nas fases posteriores. Dessa forma os dados foram transformados para uma modelagem relacional. No modelo relacional os dados são dispostos em tabelas com seus respectivos relacionamentos (entidade-relacionamento), eliminando inconsistências e principalmente a redundância de dados. Nesse propósito foram criadas seis tabelas, quais sejam: *bos, genero, local, orientacao_sexual, tempo e tipificacao*. A tabelas serão detalhadas a seguir.

A tabela *genero* é formada pelos dados distintos de gênero encontrados na base de dados. A tabela *orientacao_sexual*, *tipificacao* e *local*, da mesma forma, ou seja, foram populadas por dados distintos de orientação sexual, tipificação e bairro,

respectivamente, com uma particularidade na tabela local, que além dos dados do bairro contém os dados da cidade, porém, como esse trabalho tem como escopo apenas os dados da cidade de João Pessoa-PB, o que varia nos registros dessa tabela são apenas os dados do bairro. A tabela *tempo* foi formada contendo os dados do dia, mês, ano, nome do mês (janeiro, fevereiro, etc), dia da semana (segunda-feira, terça-feira, etc), o trimestre (1º trimestre, 2º trimestre, etc), o semestre (1º semestre e 2º semestre) e o turno (manhã, tarde, noite ou madrugada). Os dados mencionados não existiam originalmente no arquivo fornecido, e foram criados como resultado de uma transformação realizada por meio de *scripts SQL* considerando a data e horário da ocorrência. A citada transformação se fez necessário para permitir uma melhor análise e filtro das ocorrências nas próximas etapas desse trabalho.

A tabela denominada de *bos* foi populada com todos registros de boletins de ocorrências policiais. Nela estão os dados das ocorrências policiais, dentre eles, a respectivas chaves estrangeiras relacionados a tabela *bos* com as tabelas *orientacao_sexual*, *tipificacao*, *local*, *tempo* e *genero.*

Devido a necessidade de clusterização dessas ocorrências policiais, a ser apresentada posteriormente, foi necessário a definição do local de ocorrência por meio das coordenadas geográficas, ou seja, latitude e longitude. Nesse cenário, foi criado os dois novos campos latitude e longitude na tabela *bos*, para armazenar tais dados.

Com objetivo de obter as coordenadas geográficas do local do crime, a *API Google Maps* foi utilizada para esse propósito. Foi desenvolvido um *script* utilizando a linguagem *Python* que, por meio do serviço de geocodificação da *API*, fora submetido o endereço da ocorrência (logradouro, bairro e cidade), consequentemente resultando a latitude e longitude correspondente.

Após ser criado e populado o modelo relacional supracitado, realizando os ajustes necessários nos dados e criação daqueles necessários, os dados ficaram prontos, e foram transportados para um sistema de BI.

Para um melhor uso de um ambiente de BI foi criado um modelo dimensional complementar para a representação das ocorrências policiais extraídas do SPP. Para isto, foram definidas as seguintes dimensões: tempo, local, gênero, orientação sexual, profissão e tipificação. Posteriormente foi criada a tabela fato Bos, representando a ocorrência de um crime especificado por pelas dimensões definidas. As tabelas referentes as dimensões e fatos foram criadas no Sistema de Gerenciamento de

Banco de Dados (SGDB), e populadas utilizando como fonte de dados o modelo relacional anterior. A Figura 9 ilustra a modelagem dimensional proposta.

A ferramenta de B.I definida para uso nesse projeto é a *QlikSense Cloud*, na versão web, desenvolvida pela empresa *QlikTech*. Ela permite conectar-se a dados oriundos de várias fontes. É uma ferramenta que está disponível na web, baseada em scripts, para importar os dados.

Após a importação dos dados para o *Qliksense*, inicia-se o processo de criação dos *dashboards*, que são ferramentas de gestão que tem como objetivo o acompanhamento e exibição de alguns indicadores que são medidos pela gestão.

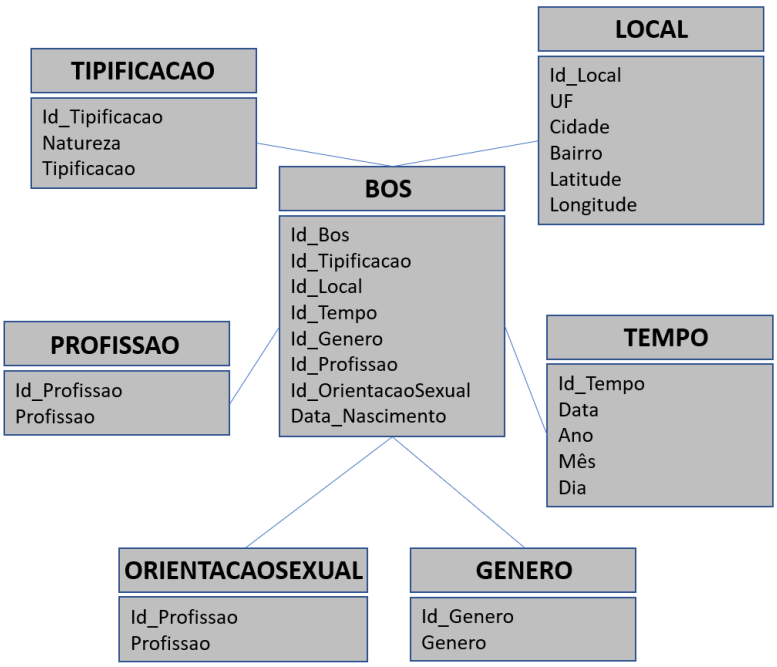

Figura 9. Modelagem dimensional estrela do Business Intelligence.

Algumas funcionalidades foram implementadas para visualização com os dados presentes no *Qliksense*, sendo possível visualizar os seguintes cenários:

- Totalização de Crimes;
- Totalização de Crimes por gênero (homem/mulher);
- Top 10 dos bairros com mais incidência criminal;
- Top 10 dos crimes com maior incidência;
- Visualização da previsão de crimes;

- Distribuição, por meio de mapa de calor, das áreas com maior incidência criminal;

É possível, de forma intuitiva, utilizando-se apenas de click do mouse, a inserção de filtros, ou a combinação entre eles, a exemplo de: filtro por bairro, filtro por cidade, filtro por gênero, filtro por ano, filtro por mês, filtro por trimestre e filtro por semestre. A Figura 10 apresenta uma visão geral do *dashboard* principal.

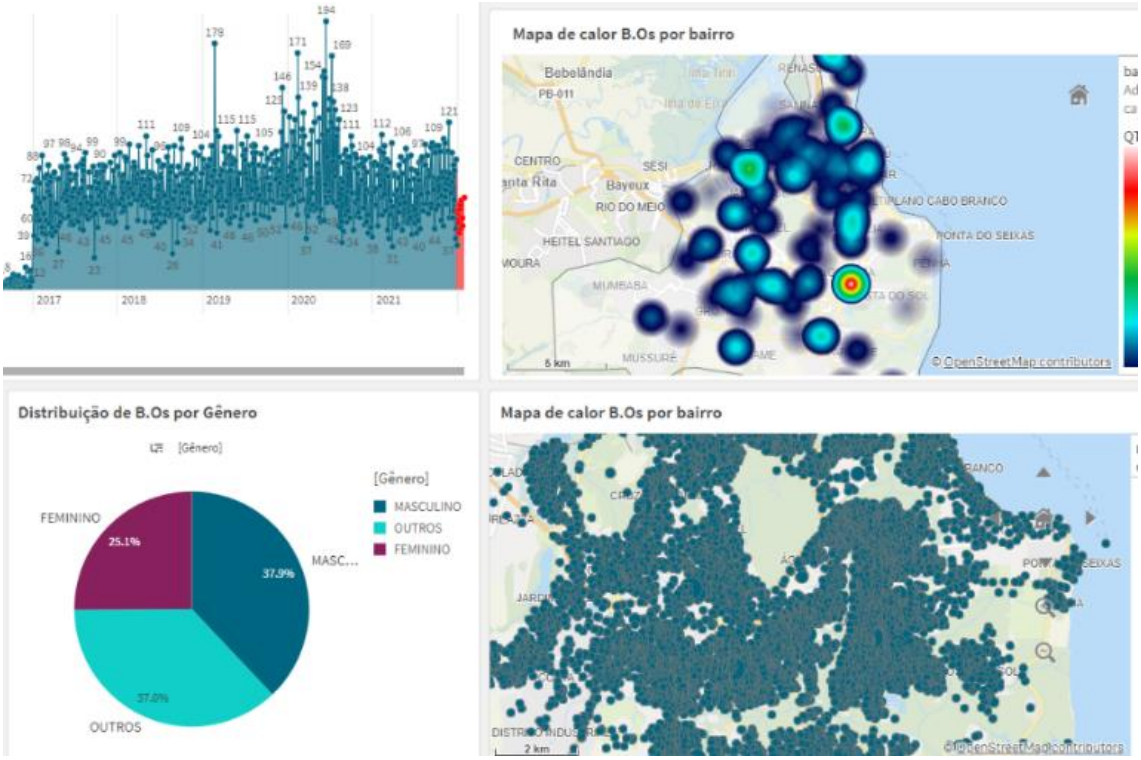

Figura 10. *Dashboard* desenvolvido com o auxílio da ferramenta *QlikSense*.

4.2 Clusterização

Nessa seção será abordado o processo de clusterização dos dados das ocorrências de crimes por meio das coordenadas do local do cometimento do fato.

A clusterização é uma técnica de agrupamento de dados pertencentes a determinas categorias que possuem um grande grau de similaridade entre si e maior dissimilaridade entre dados de outras categorias. Nesse trabalho utilizaremos o K-means para realizar a esta clusterização.

O K-means é definido por Sinaga e Yang (2020) e Li e Wu (2012) como um método de clusterização não supervisionado, com uma abordagem paramétrica e particional, onde o método agrupa os dados baseado principalmente em uma função objetivo, formando uma quantidade finita de *clusters*, porém destaca que o algoritmo K-means é amplamente afetado por inicializações onde precisa receber a quantidade de clusters a ser formado. A quantidade de clusters a ser formado pelo K-Means vai variar e depender dos dados que são submetidos ao processo de clusterização. A literatura dispõe de várias métricas que possibilitam identificar a quantidade de clusters ideal para um determinado conjunto de dados, a exemplo do *Bayesian Information Criterion* (BIC), *Silhouette Width* (SW), *Elbow Method*, dentre outros.

Com objetivo de definir a melhor quantidade de clusters a serem formados, tendo em vista a importância desse valor para um melhor resultado do algoritmo K-means, considerando os dados de ocorrências utilizados nesse trabalho, esse estudo aplicará a técnica de *Elbow Method*.

O *Elbow Method* definido por Humaira e Rasyidah (2021) como uma técnica visual para determinar a quantidade de clusters K a ser formado por um conjunto de dados específico, avaliados pela soma do erro quadrático (SSE) das distâncias euclidianas entre os pontos de um mesmo cluster e seu centroide. A variância de SSE é plotado em um gráfico, que inicialmente considera como quantidade de cluster ideal $K = 2$, onde consequentemente os primeiros clusters agruparão muitos pontos, obviamente, demonstrando um valor de SSE maior, e com a o incremento de K em 1 em cada etapa da clusterização com o algoritmo K-means, o SSE tende a cair, ao ponto que com o aumento de K, em dado momento, o valor de SSE cairá drasticamente, formando um menor ângulo no gráfico, fazendo uma analogia a um cotovelo, como pode ser observado na Figura 11. Nesse momento, identifica-se visualmente o melhor valor para K. Após a formação do menor ângulo, percebe-se que a variação do SSE é tão pequena que a adição de mais clusters torna-se insignificante, ratificando dessa forma o melhor valor para K.

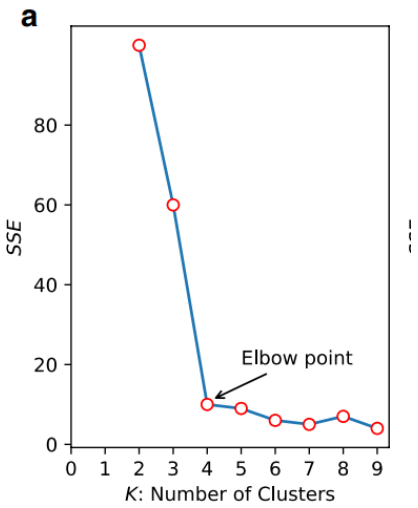

Figura 11. Gráfico exibindo o ponto de menor ângulo definido como Cotovelo.
Fonte: Humaira e Rasyidah (2021)

É importante ressaltar que Shi et al. (2021) e Kodinariya e Makwana (2013) destaca que o método Elbow tem um problema, enfatizando que por se tratar de uma técnica de análise unicamente visual, nem sempre permite distinguir de forma inequívoca, até por analistas experientes, o ponto com o menor ângulo formado, tendo em vista que em alguns casos a curva plotada ser bastante suave, conforme pode ser visto na Figura 12.

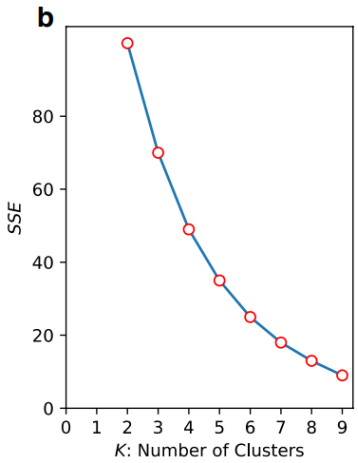

Figura 12. Gráfico exibindo a curva suave plotada pelo método Elbow.
Fonte: Humaira e Rasyidah (2021)

4.3 Séries Temporais e Previsão de Dados

Dando continuidade aos processos que compõe a solução proposta, nessa fase será abordado o processo de previsão de dados criminais, ação relevante e imprescindível diante a escassez de recursos e contingente policial para combater à criminalidade como um todo. Prever o número aproximado de crimes permite fazer um

planejamento estratégico mais apurado pelas forças de segurança. Na fase anterior foi possível identificar regiões com maiores índices criminais, ou seja, as regiões que devem ter mais a presença de patrulhas policiais com objetivo de coibir possíveis delitos, como também atender com mais rapidez os chamados da população quando o crime não pôde ser evitado.

Os dados que estão sendo utilizados nesse trabalho estão representados pelos Boletins de Ocorrências policiais, onde cada registro do banco de dados representa um crime ocorrido e registrado pela polícia.

Nessa seção serão utilizados os registros de ocorrências para a criação de séries temporais. As séries temporais tem como objetivo apresentar os dados de quantidade de crimes ocorridos por dia, permitindo visualizar o comportamento do fenômeno analisado, podendo identificar tendências, sazonalidades, periodicidades e descrever alguns comportamentos. Logo, após a criação das séries temporais prosseguiremos com a previsão de quantidade de crimes que podem ocorrer nas regiões.

O primeiro passo para a criação das séries temporais foi extrair os dados da base de ocorrências criminais agrupando-os considerando a região (*cluster*) que ele pertence e o dia, isso porque cada registro no banco representa um crime, portanto deve-se agrupar todos os crimes de uma determina região por dia, totalizando-os. Como exemplo, a Figura 13 ilustra um gráfico referente a série temporal de crimes diários da região 1 no intervalo de tempo de janeiro de 2017 a dezembro de 2021.

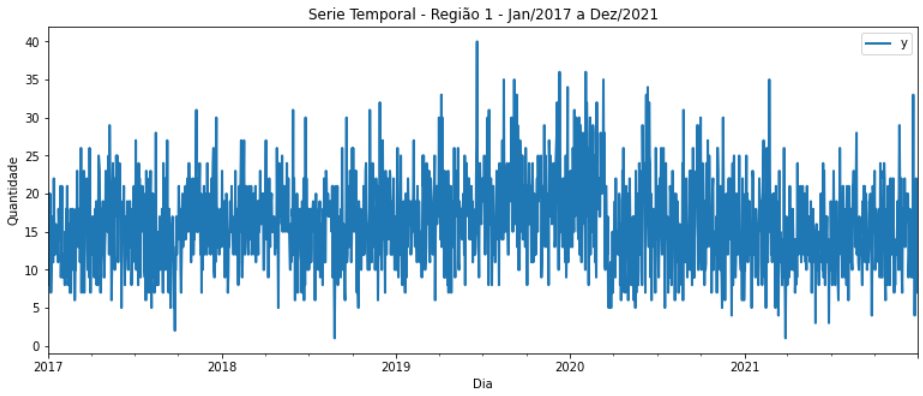

Figura 13. Gráfico da série temporal por dia do período de jan/2017 a dez/2021 da região 1.

O mesmo processo foi realizado para todas as outras regiões identificadas, onde foram geradas séries temporais diárias e semanais.

O próximo passo é o de previsão de quantidade de crimes nessas regiões. Para realizar a previsão foi utilizada a *API Prophet*. Com a *API Prophet* é possível realizar a previsão utilizando as séries temporais (dados históricos).

O Prophet trabalha com ajustes dos dados de entrada, identificando um modelo que melhor descreva as componentes da série temporal fornecida. Em seguida é fornecido um intervalo de tempo futuro, não contido na série original, sendo seus crimes preditos pelo modelo ajustado.

4.4 Problema de Localização de Máxima Cobertura Capacitado

Dando continuidade à proposta de solução apresentada, o modelo matemático discutido nessa seção visa fornecer a localização de patrulhas policiais em locais específicos considerando algumas restrições. Inicialmente se tem um conjunto de clientes I e uma quantidade de ocorrências a_i, além de um conjunto de facilidades K (patrulhas policiais) em possíveis locais J que possam atender essas ocorrências visando uma maior agilidade. Considerando s_k como a capacidade de atendimento de cada facilidade k e uma distância D como sendo o raio máximo pré-estabelecido que uma facilidade pode se encontrar de um ponto de demanda de clientes, busca-se o conjunto de locais onde essas facilidades devem ser postas, que é definido por N_i. O conjunto N_i é formado por todos os locais que $d_{ij} \leq D$, onde d_{ij} representa a função de distância do cliente i a localidade j, resultando em todas as localidades que se encontram dentro do limite máximo de distância D de cada demanda a_i.

Duas variáveis de decisão são empregadas no modelo, x_{jk} e y_{ik}. A variável x_{jk}, quando tiver seu valor igual a 1, indicará que uma facilidade k foi instalada em uma localidade j, caso contrário, seu valor será igual a 0. Já a variável y_{ik} representa a quantidade de demandas de um cliente i atendidas pela facilidade k.

Dessa forma a solução estabelece locais para posicionar as facilidades considerando a uma distância máxima pré-definida do cliente, assim como, limitar a quantidade de demanda a ser atendida por cada facilidade. Nesse contexto,

apresentamos o seguinte modelo proposto e aplicado para suprir uma melhor otimização.

$$Maximizar \quad \sum_{i \in I} \sum_{k \in K} y_{ik} \tag{1}$$

$Sujeito\ a$:

$$\sum_{i \in I} y_{ik} \leq s_k, \qquad \forall k \in K, \tag{2}$$

$$\sum_{k \in K} y_{ik} \leq a_i, \qquad \forall i \in I, \tag{3}$$

$$y_{ik} \leq a_i \cdot (1 - x_{jk}), \qquad \forall i \in I, \forall k \in K, \forall j \notin N_i, \tag{4}$$

$$s_k \cdot \sum_{\forall j \in N_i} x_{jk} \geq y_{ik}, \qquad \forall i \in I, \forall k \in K, \tag{5}$$

$$\sum_{j \in J} x_{jk} \leq 1, \qquad \forall k \in K. \tag{6}$$

A função objetivo (1) procura maximizar o número total de demandas (ocorrências policias) atendidas por cada facilidade, por conseguinte, deve-se considerar as seguintes restrições.

A restrição (2) define que cada facilidade só pode atender s_k demandas dos clientes. A restrição (3) estabelece que cada cliente só pode ter a_i demandas atendidas pelas facilidades. A restrição (4) garante que os clientes só podem ser atendidos por facilidades atribuídas para a localidades cujo raio de cobertura máximo D seja atendida. A restrição (5) assevera que é obrigatório uma facilidade em uma localidade caso algum cliente seja atendido, como também que dada uma facilidade naquela localidade o número de demandas a_i que essa facilidade atende é menor ou igual a s_k. Por fim, a restrição (6) garante que cada facilidade será atribuída para no máximo uma localidade.

Resolvemos este modelo para uma dada instância através do auxílio de resolvedores matemáticos.

Capítulo

5

Resultados

O presente capítulo tem como objetivo a apresentação dos resultados computacionais obtidos no desenvolvimento do presente trabalho, com intuito de validar os três pilares que compõe esta solução, sendo organizados nos seguintes tópicos: clusterização de regiões, previsão de crimes e alocação de viaturas utilizando o modelo proposto.

As fases de clusterização e de previsão de crimes das regiões foram implementadas na linguagem *Python*, com o auxílio da heurística K-means, disponível na *API Scikit Learn (*Pedregosa et al., 2011*),* para as clusterizações e da *API Prophet* para realizar as previsões. Para implementar o modelo de alocação de viaturas proposto foi utilizada a linguagem de programação C++ com o compilador GNU GCC versão 9.4.0 (opção de compilação -O2) e auxílio do resolvedor matemático CPLEX na versão 20.1. Todos os experimentos foram executados em um computador equipado com processador AMD Ryzen 5700U com 8GB de RAM e executando o sistema operacional Ubuntu 20.04, limitados a 1 thread.

5.1 Clusterização de Regiões

Conforme descrito na seção 4.2 foi utilizado o algoritmo K-means para o processo de clusterização dos dados das ocorrências policiais, ou seja, para agrupar pontos com maior similaridade, porém, como entrada, o K-means precisa da quantidade de cluster K a ser considerado no processo. A quantidade de cluster K foi definido utilizando o método Elbow, mas como já discutido nesse trabalho o método Elbow tem seus desafios quando a curva formada é muito suave.

Para superar a dificuldade de identificar o ponto ideal (cotovelo) que define o melhor K pelo método Elbow, tendo em vista ser um método exclusivamente visual, esse trabalho utilizou o método proposto por Shi et al. (2021), que propõe um método quantitativo.

O primeiro passo do método apresentado por Shi et al. (2021) é calcular a soma do erro quadrado (SSE, sigla em inglês) para cada valor de K. Com a lista dos valores de SSE para cada K, e considerando cada valor como um ponto, calcula-se o ângulo α formado pelas distâncias de três pontos adjacentes ($P_i, P_j\ e\ P_k$) através da equação $\arccos = \frac{a^2+b^2+c^2}{2ab}$, onde a é distância euclidiana entre os pontos $P_i\ e\ P_j$, b a distância euclidiana entre os pontos $P_j\ e\ P_k$ e c sendo a distância euclidiana entre $P_k\ e\ P_i$. Todos os ângulos calculados são armazenados e no final é identificado o menor ângulo formado, sendo este o escolhido como melhor K para o conjunto de dados.

Foi submetido o conjunto de dados de ocorrências policiais e foi possível identificar o valor de K ideal igual a 3, como pode ser visto na Figura 14 que mostra os valores de SSE para cada valor de K. O Ponto de menor ângulo está marca com um x vermelho.

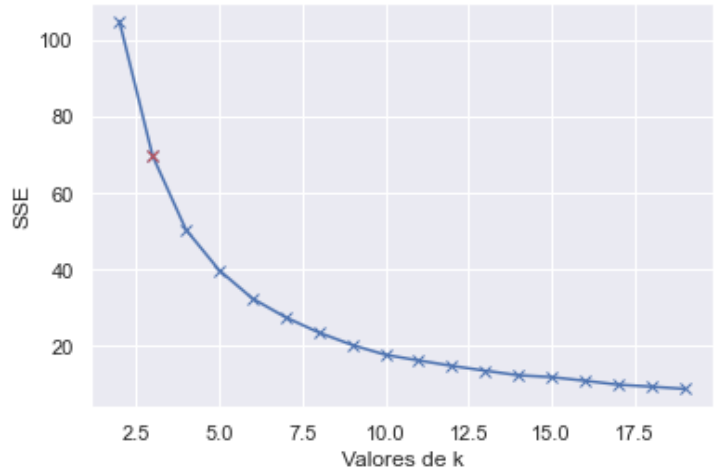

Figura 14. Gráfico Elbow indicando o melhor K igual a 3.

Considerando o melhor valor para $K = 3$, foi possível observar que o raio médio entre os pontos de cada cluster em relação ao seu centróide é de aproximadamente 9.7 km, sendo um raio de cobertura proibitivo para se atribuir a uma viatura. Sendo assim, foi estabelecido um raio médio de 4 km para cada cluster, sendo

este valor atingido por clusters com pelo menos $K = 12$, como podemos observar na Figura 15.

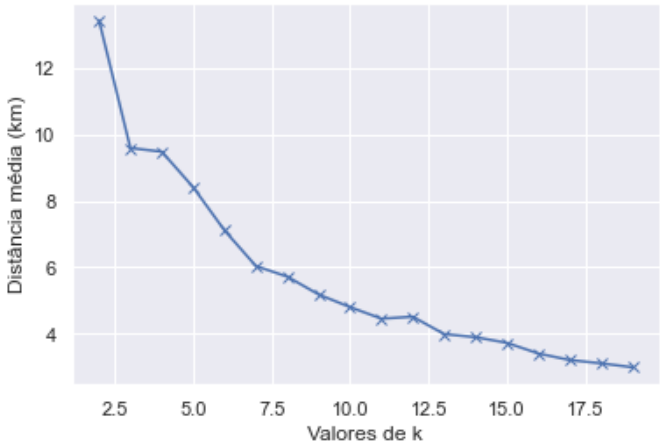

Figura 15 – Relação do raio médio entre os pontos e os centroides dos clusters.

Considerando a situação acima descrita, o conjunto de dados de ocorrências policiais foram submetidos novamente ao método Elbow, porém, de forma que considerasse a distância média entre os pontos e seus respectivos centroides de no máximo 4.0km de média, resultando no gráfico da Figura 16. Com o raio médio de 4.0km, o melhor ângulo formado pelo método Elbow indica um $K = 13$, ou seja, que o ideal para o cenário proposto é a formação de 13 clusters.

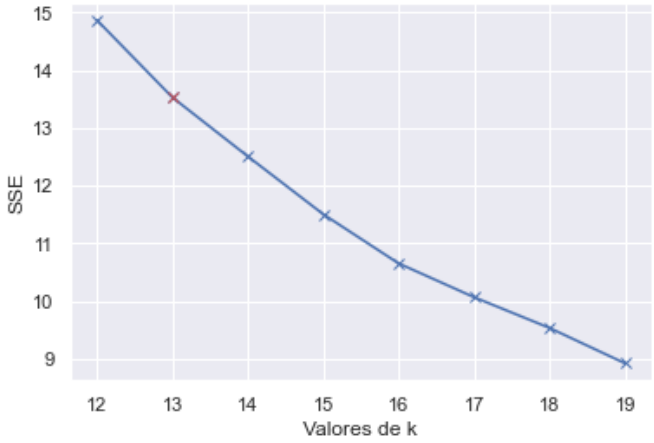

Figura 16. Gráfico de Elbow considerando o raio médio máximo de 4.0Km.

Os clusters formados podem ser observados na Figura 17.

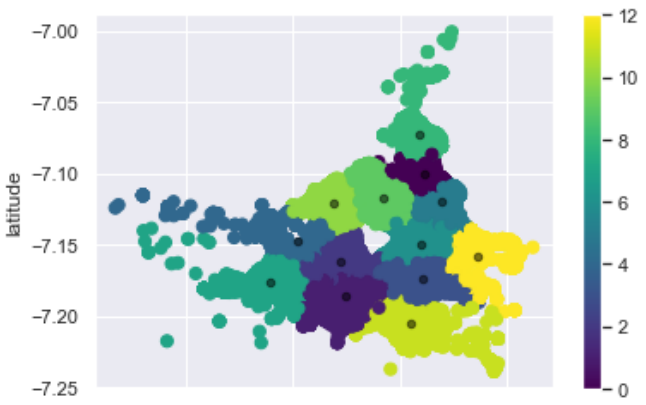

Figura 17. Clusterização dos dados considerando o número de cluster $K = 13$.

5.2 Previsão de Crimes

O processo de predição já foi explanado na seção 4.3 desse trabalho. Na presente seção serão comparados os resultados obtidos nas previsões da *API Prophet* sobre as séries temporais agrupadas por regiões, produto da fase de clusterização anterior e as séries temporais agrupadas por bairro, como são considerados nas tomadas de decisão da equipe de planejamento atual da polícia.

Inicialmente vale-se ressaltar que a métrica MSE (Erro Médio Quadrado) não foi utilizado como parâmetro para medição da eficiência do processo preditivo, tendo em vista que a referida métrica não é adequada para avaliar a previsão de diferentes séries temporais, e sim, para avaliação de desempenho do processo de predição de uma mesma série temporal utilizando-se diferentes modelos preditivos, nesse caso, indicando qual o melhor modelo apresentou o menor MSE sobre uma mesma série temporal. Para os fins de medição de desempenho do processo de previsão desse trabalho foi utilizado a métrica MAPE (Erro Percentual Médio Absoluto).

O MAPE é uma métrica de simples entendimento, definida pela diferença média percentual entre os valores previsto e real de uma série temporal, ou seja, se o valor do MAPE foi 10% significa um erro de 10% no processo de previsão do valor previsto em relação ao valor real.

Considerando y a série temporal com os crimes que realmente ocorreram em uma localidade avaliada durante o ano de 2021 e o $\hat{y}$ a série temporal predita pelo *Prophet,* temos que o MAPE é dado pela seguinte expressão:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}.100\%$$

O processo de previsão realizado nesse trabalho utilizou-se uma base de dados de ocorrências policiais que compreende no período de janeiro de 2017 a dezembro de 2021, totalizando 88 mil ocorrências criminais, distribuídas em 97 (noventa e sete) bairros e 13 (treze) regiões na cidade de João Pessoa-PB, apresentando a previsão mensal de crimes nessas regiões e bairros.

Para fins de previsão, os dados foram divididos em duas partes:

- Janeiro de 2017 a dezembro 2020 representa a base de dados de treinamento do modelo *API Prophet*;
- Janeiro de 2021 a dezembro de 2021 representa a base de dados de teste a ser utilizada para medir o MAPE com a série temporal prevista.

Os dados são agrupados em séries temporais referentes a cada localidade. A Tabela 1 ilustra a série temporal da base de teste referente a região 2.

Tabela 1. Série temporal da base de teste referente a região 2.

Mês	Qtd. Crime
2021-01-31	201
2021-02-28	221
2021-03-31	180
2021-04-30	182
2021-05-31	188
2021-06-30	205
2021-07-31	225
2021-08-31	254
2021-09-30	230
2021-10-31	263
2021-11-30	258
2021-12-31	200

Para cada localidade, sua base de treinamento é usada no treinamento de modelo preditor, uma série temporal é prevista pelo modelo treinado. A Tabela 2, por exemplo, ilustra estes dados previstos para região 2.

Tabela 2. Série Temporal prevista pelo modelo do *Prophet* para a região 2.

Mês	Qtd. Crime
2021-01-31	154,50
2021-02-28	161,95
2021-03-31	207,61
2021-04-30	156,86
2021-05-31	167,70
2021-06-30	195,81
2021-07-31	201,79
2021-08-31	166,92
2021-09-30	167,10
2021-10-31	189,18
2021-11-30	193,70
2021-12-31	197,57

De posse das duas séries temporais de cada localidade serão calculados os MAPEs referente aos dados de todas as localidades (regiões e bairros).

Como pode ser observado na Tabela 3, após o processamento do MAPE da região 2, foi obtido o valor de 18%, ou seja, a margem de erro de previsão apresentado pelo modelo quando comparado ao valor real que compreende todos os crimes ocorridos na região 2 é de 18%. Quanto menor o valor, mais preciso é a previsão.

Tabela 3. MAPE computado para a região 2.

Mês	Real	Previsto	(|Real – Previsto|) / Real
2021-01-31	201	154,50	0,23
2021-02-28	221	161,95	0,26
2021-03-31	180	207,61	0,15
2021-04-30	182	156,86	0,13
2021-05-31	188	167,70	0,11
2021-06-30	205	195,81	0,05
2021-07-31	225	201,79	0,10
2021-08-31	254	166,92	0,34
2021-09-30	230	167,10	0,27
2021-10-31	263	189,18	0,28
2021-11-30	258	193,70	0,25
2021-12-31	200	197,57	0,01
Média((|Real – Previsto|) / Real) * 100			**18%**

Na Tabela 4 são apresentados os valores de MAPE calculados de todas as regiões (coluna MAPE), já a coluna quantidade é referente a quantidade de crimes reais que ocorreram nas respectivas regiões contida em sua base de teste por fim a coluna erro absoluto traz a multiplicação do valor de MAPE por quantidade.

Tabela 4. MAPE ponderado das 13 regiões criminais.

Região	MAPE	quantidade	erro absoluto
1	38%	1483	563,54
2	18%	2607	469,26
3	28%	1665	466,22
4	20%	2706	541,20
5	14%	1467	205,38
6	25%	1636	409,00
7	19%	1602	304,38
8	14%	1006	140,84
9	35%	1125	393,75
10	25%	2243	560,75
11	20%	3005	601,00
12	12%	1490	178,80
13	31%	555	172,05
	Total	**22590**	**5006,15**

O MAPE geral das regiões obtido com dados mensais foi de 22%, sendo computado pela razão entre o erro absoluto total de todas as regiões pelo total de crimes que realmente ocorreram ($\frac{5006,15}{22590}$ * 100 = 22%).

O mesmo procedimento foi aplicado aos dados agrupados por bairros, com todos os crimes e separando os crimes nos tipos de crimes contra o Patrimônio e crimes contra a Pessoa. Os resultados comparativos estão contidos na Tabela 5.

Tabela 5. Comparativo do MAPE em cada tipo de agrupamento dos dados.

Tipo de Crime	Agrupamento	MAPE
Todos	região	22%
	bairro	36%
Patrimônio	região	25,90%
	bairro	39,67%
Pessoa	região	36,69%
	bairro	68,29%

Em todos os cenários apresentados, considerando todos os crimes, ou separando por tipo, é possível identificar que ao utilizamos o agrupamento por região, o erro de previsão é no mínimo 14% menor comparado ao agrupamento por bairro.

5.3 Alocação de Viaturas utilizando o modelo matemático

Nessa seção serão apresentados os dados oriundos do processamento do modelo matemático proposto nesse trabalho e descrito na Seção 4.4. O referido modelo tem como objetivo realizar a alocação de viaturas policiais em locais onde possa realizar o melhor atendimento considerando algumas restrições. Comparamos o resultado da cobertura obtida na atribuição das viaturas em centroides definidos na fase de clusterização por região, com as soluções obtidas pela atribuição de viaturas em pontos centrais dos bairros, simulando a clusterização do sistema atual da Polícia.

Para fins de processamento do modelo, alguns dados de entrada foram necessários, tais como: matriz de distância entre os centroides (latitude e longitude) dos bairros e regiões, quantidade de crimes previstos em cada região e bairro nos sete primeiros dias da semana do mês de janeiro de 2022, quantidade de viaturas disponível para patrulhamento em cada dia, capacidade de atendimento individual de cada viatura que fixamos em 5 crimes diários ($\boldsymbol{s_k} = 5$) e por fim, o raio de cobertura das viaturas fixado em 4 km ($\boldsymbol{D} = 4$).

As Tabelas 6 e 7 são formadas das seguintes colunas: coluna dia representa o dia do mês de janeiro de 2022 que foi realizada a projeção. As colunas $\sum \boldsymbol{a_b}$ e $\sum \boldsymbol{a_r}$ representam o somatório da quantidade de crimes previstos nos bairros e regiões, respectivamente. A coluna $|\boldsymbol{K}|$ representa o número de viaturas que podem ser alocadas na solução. A coluna $\boldsymbol{Z}$ apresenta a quantidade de crimes coberto pela solução, coluna $\boldsymbol{T}(\boldsymbol{s})$ representa o tempo computacional em segundos e $\boldsymbol{C}(\%) = \frac{Z}{\sum a}$*100 como o percentual de cobertura de crimes. Por fim, o $\boldsymbol{\Delta}(\%)$ representa a diferença percentual de cobertura entre os valores obtidos da cobertura de regiões e bairros.

Tabela 6. Resultado de Alocação de Viaturas por dia.

Instância				Bairros			Regiões			
dia	$\sum a_b$	$\sum a_r$	$\|K\|$	Z	T (s)	C (%)	Z	T (s)	C (%)	Δ (%)
			10	26	0,2	51	43	0	76,8	25,8
			11	28	0,2	54,9	46	0,1	82,1	27,2
1	51	56	12	29	0,2	56,9	49	0,1	87,5	30,6
			13	30	0,2	58,8	52	0	92,9	34,1
			14	31	0,2	60,8	54	0,1	96,4	35,6
			10	24	0,1	54,6	39	0,1	83	28,4
			11	25	0,1	56,8	41	0,1	87,2	30,4
2	44	47	12	26	0,2	59,1	43	0,1	91,5	32,4
			13	27	0,2	61,4	45	0,1	95,8	34,4
			14	28	0,1	63,6	46	0,1	97,9	34,3
			10	30	0,2	44,8	48	0,1	68,6	23,8
			11	32	0,2	47,8	52	0,1	74,3	26,5
3	67	70	12	34	0,2	50,8	56	0	80	29,2
			13	36	0,3	53,7	60	0,1	85,7	32
			14	38	0,2	56,7	63	0,1	90	33,3
			10	30	0,2	47,6	48	0,1	70,6	23
			11	32	0,2	50,8	51	0,1	75	24,2
4	63	68	12	34	0,1	54	54	0,1	79,4	25,4
			13	36	0,2	57,1	58	0,1	85,3	28,2
			14	38	0,4	60,3	61	0,1	89,7	29,4
			10	31	0,2	49,2	48	0,1	68,6	19,4
			11	33	0,2	52,4	52	0,1	74,3	21,9
5	63	70	12	35	0,2	55,6	56	0,1	80	24,4
			13	37	0,2	58,7	60	0,1	85,7	27
			14	39	0,3	61,9	63	0,1	90	28,1
			10	31	0,2	50	47	0,1	68,1	18,1
			11	33	0,2	53,2	51	0,1	73,9	20,7
6	62	69	12	35	0,2	56,5	56	0,1	81,2	24,7
			13	36	0,2	58,1	59	0,1	85,5	27,4
			14	37	0,2	59,7	62	0,1	89,9	30,2
			10	32	0,2	49,2	47	0,1	68,1	18,9
			11	34	0,2	52,3	51	0,1	73,9	21,6
7	65	69	12	36	0,2	55,4	56	0	81,2	25,8
			13	38	0,2	58,5	59	0,1	85,5	27
			14	39	0,2	60	62	0,1	89,9	29,9

Na Tabela 6, a instância do dia 01 possui um total previsto para os bairros de 51 crimes e para as regões de 56 crimes. São cinco soluções para este dia variando o número de viaturas $|\boldsymbol{K}|$. Utilizando 10 viaturas no cenário de bairros, o modelo consegue apresentar um resultado que cobre 51% do total de crimes (26 dos 51 previstos), e gastou 0,2 segundos para fornecer essa solução, porém utilizando o mesmo número de viaturas, quando no cenário de regiões, o modelo apresenta uma solução que cobre 76,8% do total de crimes (43 dos 56 previstos), totalizando um aumento de 25,8% de cobertura dos crimes. Com o aumento de números de viaturas a serem utilizadas na solução, observa-se que há um aumento na cobertura de crimes quando utilizado o agrupamento de regiões obtendo até 96,4% da cobertura criminal.

Tabela 7. Resultado de Alocação de Viaturas por Semana.

Instância				Bairros			Regiões			
Sem	$\sum a_b$	$\sum a_r$	$\|K\|$	Z	T (s)	C (%)	Z	T (s)	C (%)	Δ (%)
			10	169	0,6	46,7	281	0,1	77,4	30,7
			11	179	0,5	49,5	300	0,1	82,7	33,2
1	362	363	12	189	0,6	52,2	319	0,1	87,9	35,7
			13	198	0,9	54,7	337	0,1	92,8	38,1
			14	206	1,1	56,9	349	0,1	96,1	39,2
			10	180	0,5	47,2	292	0,1	76,4	29,2
			11	192	0,6	50,4	312	0,1	81,7	31,3
2	381	382	12	202	0,7	53	332	0,1	86,9	33,9
			13	212	0,9	55,6	351	0,1	91,9	36,3
			14	220	1	57,7	365	0,1	95,6	37,9
			10	201	0,5	48,2	314	0,1	74,9	26,7
			11	212	0,7	50,8	337	0,1	80,4	29,6
3	417	419	12	223	0,8	53,5	359	0,1	85,7	32,2
			13	232	1	55,6	379	0,1	90,5	34,9
			14	241	1	57,8	398	0,1	95	37,2
			10	209	1,2	47,2	322	0,1	72,5	25,3
			11	221	1,8	49,9	348	0,1	78,4	28,5
4	443	444	12	231	2,2	52,2	370	0,1	83,3	31,1
			13	241	6,7	54,4	391	0,1	88,1	33,7
			14	250	2,6	56,4	411	0,1	92,6	36,2
			10	208	1,2	46,1	324	0	71,5	25,4
			11	220	1,3	48,8	349	0,1	77	28,2
5	451	453	12	230	2,9	51	371	0,1	81,9	30,9
			13	239	6	53	391	0,1	86,3	33,3
			14	248	5	55	409	0,2	90,3	35,3

Já a Tabela 7 são apresentados os resultados do modelo considerando as demandas previstas por semanas, totalizando 5 semanas no mês de janeiro de 2022. Como os dados são agrupados semanalmente, o valor anterior $\boldsymbol{s_k = 5}$ que considerava que o número máximo de atendimento de ocorrências por viatura é de cinco crimes por dia, foi multiplicado por 7, adotando assim um $\boldsymbol{s_k} = 35$.

É possível perceber na Tabela 7 que a qualidade das soluções geradas para o agrupamento por regiões continua sendo muito superior a soluções dos bairros, sendo todos os valores de $\Delta\,(\%) > 25$ e alcançando um $\Delta\,(\%) = 39{,}7$ no dia 1 com $|K| = 14$.

Capítulo

6

Conclusão

O presente trabalho apresentou uma solução que otimiza o processo de alocação de viaturas policiais tendo em vista o número previsto de crimes em áreas, tais como bairros e regiões, para uma melhor gestão operacional e estratégica por parte do gestões das unidades de policiamento operacional.

A solução proposta foi embasada em três pilares, sendo eles: clusterização de áreas baseadas nas ocorrências policiais, previsão de crimes nos respectivos bairros e regiões, e por fim na prescrição de um modelo matemático que sugere a alocação de viaturas policiais em locais que seja possível um atendimento mais eficaz quando na ocorrência de um fato criminoso, considerando a capacidade de atendimento de crimes por cada viatura e distância máxima pré-definida.

No processo de clusterização foi utilizado o algoritmo K-means, no qual foram submetidos as ocorrências policiais e suas respectivas localizações (latitude e longitude). O algoritmo K-means tem como entrada a quantidade de *cluster* a ser considerado no processo de clusterização. Por meio do método *Elbow* (Cotovelo) e considerando o raio de distância máximo de 4 quilômetros entre os pontos, foi possível estabelecer a quantidade de regiões (*clusters*) ideal para o conjunto de dados submetido, apresentando como resultados 13 regiões.

Dando sequência a construção da solução foi realizada a fase de previsão de crimes, de forma separada, onde foi realizada a previsão considerando os 103 bairros e a previsão das 13 regiões. Após o processo de previsão, foi calculado o MAPE das regiões e bairros, com objetivo de mensurar a eficiência da previsão dos dados. Foi possível observar que, a perspectiva proposta por essa pesquisa, onde dividiu a área geográfica em regiões por meio da clusterização, é mais eficiente em relação a

respectiva da divisão por bairros. Em todos os cenários apresentados obtendo um ganho mínimo de 14% na previsão dos crimes.

O último pilar da solução tratou do processo de alocação de viaturas policiais, considerando algumas restrições. Como resultado dessa fase é notório a diferença de performance quando a solução considera a perspectiva de região em detrimento de bairro, chegando a ter um ganho de cobertura de 35,6% a mais quando utilizado as regiões na previsão diária, e de 39,2% a mais quando na previsão semanal. É possível observar que enquanto o agrupamento em regiões tem soluções que apresentam mais de 90% de cobertura dos crimes previstos, o agrupamento por bairros possui soluções cuja cobertura não passam de 63.6% dos crimes previstos.

Por fim, é possível concluir que o presente trabalho apresentou como resultados valores bastantes satisfatório e de grande contribuição para ciência e pesquisa, já que em todos os cenários analisados é possível observar que a perspectiva de divisão da área geográfica por região apresentou dados mais performático em relação aos dados resultantes da divisão por bairro, possibilitando que sejam utilizados como e direcionamentos para futuros trabalhos, ou até mesmo, para evolução deste.

Capítulo

7

Referências

Ahmed S., Akter S., Islam K., Rahman M., Shawon M. H., “Forecasting PV Panel Output Using Prophet Time Series Machine Learning Model,” 2020 IEEE REGION 10 CONFERENCE (TENCON) Osaka, Japan, November 16-19, 2020.

Braghittoni, R. Business Intelligence: Implementar do jeito certo e a custo zero. Casa do Código. 2015.

Campello, R.J.G.B. e Hruschka, E.R., “A fuzzy extension of the silhouette width criterion for cluster analysis”, Fuzzy Set and System, Vol.(157), Issue 21, 1 November 2006, p. 2858-2575. Elsevier.

Carvalho, V. A. de C., 2011. Disponível em:< https://www.scielo.br/j/rk/a/bnjfd8BgmpTSXSSSyXQ3qbj/?lang=pt>. Acesso em: 25 de julho. 2021.

Chartier, S. e Morissette, L. “The k-means clustering techinque: General considerations and implementation in Mathematica.”. Tutorials in Quantitative Methods for Psychology 2013, Vol. 9(1), p. 15-24.

CHURCH, R.; REVELLE, C. The maximal covering location problem. In: SPRINGER. Papers of the Regional Science Association. [S.l.], 1974. v. 32, n. 1, p. 101–118.

Cruz, M. D., Ochi, L. S. “O problema de clusterização automática: um novo método utilizando o ILS,” 10th Brazilian Congress on Computacional Intelligence (CBIC’2011)

Darapaneni et al. (2021) “Forecasting Vaccination Drive In India for Herd Immunity using SIR and Prophet Model” 2021 IEEE World AI IoT Congress (AIIoT), 2021.

Dias, C. R. Algoritmo Evolutivo para o Problema de Clusterização de Grafos Orientados: Desenvolvimento e Análise Experimental. Dissertação de Mestrado. Universidade Federal Fluminense, 2004.

Dias, C. R., Ochi, L. S., Soares, S. S. F., Clusterização em Mineração de Dados. Encontro Regional de Informática RJ/ES, 2004, Vitória. Anais do ERI 2004 RJ/ES, Vitória: ERI, 2004. p.1-6

Drummond, L. M. A, Ochi, L. S, Rosário, S. S. “Um algoritmo de Construção e Busca Local para o Problema de Clusterização de Bases de Dados,” Sociedade Brasileira de Matemática Aplicada e Computacional, Vol. 7. nº 1 (2006), 109-118.

Ehlers, R.S. Análise de Séries Temporais. Quarta Edição Publicada em 2007.

Faceli, K.; Lorena, A. C.; Gama, J.; Carvalho, A. C. P. de L. F. Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina. Rio de Janeiro: LTC, 2011.

Ferreira, R. J. P. e Ferreira, H. L.,”Decision support system for location of back-up transformers based on a multi-attribute p-median model”, 2012 IEEE International Conference on Systems, Man, and Cybernetics, 2012.

Ferro, A. L. Inteligência de segurança pública e análise criminal. Revista Brasileira de Inteligência. Brasilia: Abin, v. 2, n. 2, abr. 2006.

Galvão, R. D., Chiyoshi, F. Y., Espejo, L. G. A. and Rivas, M. P. A., 2003. Disponível em:<https://www.scielo.br/j/pope/a/xDj87ZWcysCqBsjqfCpw9sx/?lang=pt>. Acesso em: 10 de junho. 2022.

Garcia, C. A. F. M. Análise de Séries Temporais com Recurso a Técnicas de Bases de Dados. Dissertação de Mestrado. Universidade do Porto – Faculdade de Engenharia, 2000.

Gong et al. (2020) “Trend Analysis of Building Power Consumption Based on Prophet Algorithm” 2020 Asis Energy and Electrical Engineering Symposium, 2020.

Horner, D., Resolução do Problema das P-Medianas Não Capacitado. Dissertação de Mestrado. Universidade Federal de Santa Catarina, 2009.

Humaira H. e Rasyidah R., "Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm," ResearchGate, 2021, doi: DOI: 10.4108/eai.24-1-2018.2292388.

Isler, C. A., Bonassa, A. C. e Cunha, C. B., “Algoritmo genético para resolução do problema de p-medianas capacitado associado à distribuição de peças automotivas”. Revista TRANSPORTES, v. 20, n. 2, 2012.

Jain, G. e Prasad, R. R., “Machine learning, Prophet and XGBoost algorithm: Analysis of Traffic Forecasting in Telecom Networks with time series data” 2020 8th Internationa Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Amity University, Noida, India. June 4-5, 2020.

Khan et al. (2020), "Analysis of Business Intelligence Impact on Organizational Performance," 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), 2020, pp. 1-4, doi: 10.1109/ICDABI51230.2020.9325610.

Kodinariya T. M. e Makwana P. R, "Review on determining number of Cluster in K-Means Clustering," International Journal of Advance Research in Computer Scince and Management Studies, 2013.

Kumar J. B. e Pande S., "Time Series Forecasting Model for Supermarket Sales using FB-Prophet," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 547-554, doi: 10.1109/ICCMC51019.2021.9418033.

Leal, E. O. Técnicas operacionais de inteligência e ações de busca na produção de provas em investigação e processo criminal: admissibilidade e limites. 2016. 87 p. Trabalho de Conclusão de Curso (Especialização em Inteligência de Estado e Inteligência em Segurança Pública) – Centro Universitário Newton Paiva, Associação Internacional para Estudos de Segurança e Inteligência, Belo Horizonte, 2016. Disponível em: http://www.mpsp.mp.br/portal/page/portal/documentacao_e_divulgacao/doc_biblioteca/bibli_servicos_produtos/BibliotecaDigital/Publicacoes_MP/Todas_publicacoes/Evandro-Ornelas-Leal-TCC.pdf. Acesso em: 15 julho. 2021.

Li Y. e Wu H., "A Clustering Method Based on K-Means Algorithm," 2012 International Conference on Solid State Devices and Materials Science, 2012, pp. 1104-1109.

Lorena, L. A. N. e Senne, E. L. F., “Abordagens de Geração de Colunas para um Problema de P-Medianas Capacitado”, 2005.

Morettin, P. A. e Toloi C. M. C. Análise de séries temporais. São Paulo – Edgard Blucher, 2004.

Morettin, P. A. e Toloi C. M. C. Modelos para previsão de séries temporais. Rio de Janeiro – Instituto de Matemática Pura e Aplicada, 1981.

Milán et al. (2020), "Success factors and benefits of using business intelligence for corporate performance management," 2020 9th International Conference On Software Process Improvement (CIMPS), 2020, pp. 19-27, doi: 10.1109/CIMPS52057.2020.9390108.

Oliveira, C. P. de. Modelos de Otimização Aplicados ao Problema de Máxima Cobertura: Estudo de Caso do Samu-BH. Dissertação de Mestrado. Centro Federal de Educação Tecnológica de Minas Gerais, 2020.

Pal, A. e Prakash, P., Pratical Times Series Analysis. Packt Publishing Ltd – B3 2PB, UK, 2017.

Pedregosa et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Pinheiro et al. (2020), "Business Intelligence Applied in Buildings Energy Efficiency," 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), 2020, pp. 1-6, doi: 10.23919/CISTI49556.2020.9141019.

Pontin, V. M., Garcia, R. A., Neto, P. B. e Ribeiro, G. M., Análise de Modelos Matemáticos para Problema Probabilístico de Localização-Alocação de Máxima Cobertura. Cadernos do Ime – Série Estatísticas. Universidade Federal do Rio de Janeiro, 2010.

Primak, F. V. da S. Decisões com B.I (Business Intelligence). Rio de Janeiro: Editora Ciência Moderna Ltda., 2008.

Prophet:< https://facebook.github.io/prophet/docs/quick_start.html#python-api>. Acesso em: 09 de junho. 2022.

Ramesh, Sharda,. Business Intelligence e Análise de Dados para Gestão do Negócio. Disponível em: Minha Biblioteca, (4th edição). Grupo A, 2019.

Ribeiro R., Oliveira A. e Pedrosa I., "Analysis of the Impact of Business Intelligence in Public Administration," 2021 16th Iberian Conference on Information Systems and Technologies (CISTI), 2021, pp. 1-5, doi: 10.23919/CISTI52073.2021.9476489.

Santos, R. P. e Muller, C., “Problema de localização de máxima cobertura aplicado à localização de esquadrões de aeronaves de interceptação na região Amazônica,” XXXVIII Simpoósio Brasileiro de Pesquisa Operacional, Goiânia-GO, 2006.

Santos, W. A., Nunes, E. R., Dias, B. R. Lucena e Pizzolato, N. D., Aplicação de modelo das p-medianas para a localização de unidades estratégicas de saúde da família ribeirinhas: um estudo de caso em uma localidade amazônica. Brazilian Journal of Development, 2020.

Sena, D. e Nagwani, N. K., “Application of Times Series Based Prediction Model to Forecast Per Capita Displosable Income” 2015 IEEE International Advance Computing Conference (IACC), 2015.

Shi et al. (2021)., "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," EURASIP Journal on Wireless Communications and Networking, 2021.

Shi Z., "Research and Application of Business Intelligence in the Nuclear Power Maintenance," 2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2013, pp. 619-622, doi: 10.1109/IIH-MSP.2013.159.

SILVA, Valtania Ferreira da. Problema de alocação de viaturas policiais: estudo de caso na cidade de João Pessoa-PB. 2014. 128 f. Dissertação (Mestrado em Engenharia de Produção) - Universidade Federal da Paraíba, João Pessoa, 2014.

Sinaga K. P. e Yang M., "Unsupervised K-Means Clustering Algorithm", IEEE Access, Abr. 2020

Söküt Açar T. e Ayman Öz N., "The determination of optimal cluster number by Silhouette index at clustering of the European Union member countries and candidate Turkey by waste indicators", Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, vol. 26, no. 3, pp. 481-487, Jun. 2020

Taylor S. J. e Letham B., "Forecasting at Scale". PeerJ Preprints 5:e3190v2 https://doi.org/10.7287/peerj.preprints.3190v2.

Williams, F. Criminology. In: Bailey, W. G. The Encyclopedia of Police Science. 2. ed. New York, London: Garland. 1995. p. 181.