Otto Tavares Nascimento

Aplicação de Métodos Não Supervisionados - Estudo Empírico com os Dados de Segurança Pública do Estado do Rio de Janeiro.

Rio de Janeiro
2016

Otto Tavares Nascimento

Aplicação de Métodos Não Supervisionados - Estudo Empírico com os Dados de Segurança Pública do Estado do Rio de Janeiro.

Dissertação apresentada à Escola de Matemática Aplicada da Fundação Getulio Vargas, para a obtenção de Título de Mestre em Modelagem Matemática.

Fundação Getulio Vargas – FGV
Escola de Matemática Aplicada
Programa de Pós-Graduação

Orientador: Moacyr Alvim Horta Barbosa da Silva
Coorientador: Cesar Zucco Junior

Rio de Janeiro
2016

Ficha catalográfica elaborada pela Biblioteca Mario Henrique Simonsen/FGV

Nascimento, Otto Tavares
Aplicação de métodos não supervisionados : estudo empírico com os dados de segurança pública do estado do Rio de Janeiro / Otto Tavares Nascimento. - 2016.
76 f.

Dissertação (mestrado) - Fundação Getulio Vargas, Escola de Matemática Aplicada.
Orientador: Moacyr Alvim Horta Barbosa da Silva.
Coorientador: César Zucco Junior

Inclui bibliografia.

1. Matemática. 2. Segurança pública - Rio de Janeiro (RJ). 3. Sociologia - Aspectos econômicos. I. Silva, Moacyr Alvim Horta Barbosa da. II. Zucco, Junior, César. III. Fundação Getulio Vargas. Escola de Matemática Aplicada. IV. Título.

CDD – 510

FGV

OTTO TAVARES NASCIMENTO

APLICAÇÃO DE MÉTODOS NÃO SUPERVISIONADOS – ESTUDO EMPÍRICO COM OS DADOS DE SEGURANÇA PÚBLICA DO ESTADO DO RIO DE JANEIRO.

Dissertação apresentada ao Curso de Mestrado em Modelagem Matemática da Informação da Escola de Matemática Aplicada da Fundação Getulio Vargas para obtenção do grau de Mestre em Modelagem Matemática da Informação.

Data da defesa: 20/12/2016.

ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA

Moacyr Alvim Horta Barbosa da Silva
Orientador (a)

Cesar Zucco Junior
Co-Orientador (a)

Sabrina Camargo

Sabrina Camargo

Joana CM Monteiro

Joana da Costa Martins Monteiro

Este trabalho é dedicado aos meus pais, familiares e amigos que acreditam mais em mim do que eu mesmo.

AGRADECIMENTOS

Agradeço primeiramente a minha mãe Maria das Graças, ao meu pai Alcendino Moreira e ao meu irmão Hugo Tavares, por me formarem como humano. O segundo agradecimento vai para meu padrinho Paulo Carneiro, para a amiga Jocilene Nazaré, para o amigo Fabricio Carneiro e meu amigo Leonardo Ghirlinzoni, por nossa insistência em sermos da mesma família, mesmo não possuindo laços sanguíneos. E o último agradecimento principal vai para Julia Guerra pela parceria nas tormentas dessa batalha.

Agradeço especialmente os meus professores, em principal Moacyr Alvim e Cesar Zucco, e os colegas da EMAp, com destaque para Mateus Alvarenga e para Bruno Lucian. Outro agradecimento especial é dedicado à equipe do ISP com destaque à Joana Monteiro e à Bárbara Caballero pelos aconselhamentos fundamentais para esse trabalho e para minha formação.

Agradeço meus amigos tanto de Volta Redonda quanto os do Rio, por terem escolhido fazer parte desta caminhada tanto no trabalho duro, quanto no lazer. Esse anos serão inesquecíveis.

Por fim, agradeço à equipe da FGV, com um abraço especial para o Ronaldo, a Cristiane, a Luziel e a Beralda, por fazerem dos dias na FGV mais iluminados, mesmo com a presença de tanto concreto.

"No princípio era o verbo. (Bíblia Sagrada, João 1, 1)

"The caterpillar is a prisoner to the streets that conceived it
Its only job is to eat or consume everything around it,
in order to protect itself from this mad city
While consuming its environment the caterpillar begins to notice ways to survive
One thing it noticed is how much the world shuns him, but praises the butterfly
The butterfly represents the talent, the thoughtfulness,
and the beauty within the caterpillar
But having a harsh outlook on life the caterpillar sees the butterfly as weak and figures
out a way to pimp it to his own benefits
Already surrounded by this mad city the caterpillar goes to work on the cocoon which
institutionalizes him
He can no longer see past his own thoughts
He's trapped
When trapped inside these walls certain ideas take roots, such as going home, and
bringing back new concepts to this mad city
The result?
Wings begin to emerge, breaking the cycle of feeling stagnant
Finally free, the butterfly sheds light on situations that the caterpillar never considered,
ending the internal struggle
Although the butterfly and caterpillar are completely different, they are one and the
same. (Mortal Man, To pimp a butterfly, Kendrick Lamar)

RESUMO

Este trabalho é uma abordagem multidisciplinar, o qual aplica-se a metodologia de matemática aplicada, em específico, aprendizagem não supervisionada, a dados de segurança pública. Busca-se identificar a semelhança entre batalhões da polícia, utilizando métodos de clusterização de modo a otimizar numericamente o critério de avaliação de McClain. Além da otimização, aborda-se intuitivamente o modelo de clusterização hierárquica, para posteriormente extrair ordem no padrão criminal dos clusters e, finalmente, aplicar o modelo de classificação OLogit, utilizando variáveis características desses clusters. Encontramos evidência de clusterização dos dados e significância na utilização de dados socioeconômicos e de policiamento na ordenação dos clusters. Resumindo, quanto maior o efetivo policial por habitante e o IDH de renda mínima em determinado batalhão maior a probabilidade de se estar em um cluster de menor incidência criminal.

Palavras-chaves: Aprendizagem não supervisionada, segurança pública, clusters, similaridade, aprendizagem dos dados, índice de McCLain.

ABSTRACT

This multidisciplinary work use an applied math methodology, especially unsupervised learning, in public security data. We seek to find the similiarity beetwen policies battalions, using clustering methods, while otimizing numerically the McCLain index. Besides that, we extract learning from data, using OLogit models in cluster's order with feature variables. We find data clustering evidence and extract significance of socioeconomic and policing data in cluster's order. In summary, a higher police force per inhabitant and a higher minimum income HDI in a given batallion results in a greater probability of being in a cluster of lower criminal incidence.

Key-words: Unsupervised learning, public security, clusters, similiarity, learning from data, McClain index.

LISTA DE ILUSTRAÇÕES

Figura 1 – Roubo Transeunte x Tempo 23
Figura 2 – Índice de criminalidade x Tempo 24
Figura 3 – Modelo de Entidades e Relacionamento do Banco de Dados 36
Figura 4 – Classificação 1 45
Figura 5 – Classificação 2 46
Figura 6 – Método: Complete 50
Figura 7 – Método: Single 51
Figura 8 – Método: Complete 54
Figura 9 – Método: Complete 55
Figura 10 – Mapas das RISP e dos Clusters Ordenados 56
Figura 11 – Índice de criminalidade x Clusters 57
Figura 12 – Índice de criminalidade x Clusters 57
Figura 13 – Índice de criminalidade x IDH 2011 58
Figura 14 – SIM x IDH 2015 59
Figura 15 – BIC x Clusters 72

LISTA DE TABELAS

Tabela 1 – Estatísticas Descritivas dos dados de Interesse para todos os batalhões 37
Tabela 2 – Matriz de similaridade 48
Tabela 3 – Resultado Numérico do problema de otimização do Índice de McClain para o método *complete* 52
Tabela 4 – Resultado Numérico do problema de otimização do Índice de McClain para o método *single* 53
Tabela 5 – Aplicação do modelo OLogit no ordenamento dos clusters. 59
Tabela 6 – Aplicação do modelo OLogit no ordenamento dos clusters 60
Tabela 7 – Aplicação do modelo OLogit no ordenamento dos clusters 61
Tabela 8 – Aplicação do modelo OLogit no ordenamento dos clusters versão alternativa 62
Tabela 9 – Classificação para o algoritmo EM 73
Tabela 10 – Cluster 1 75
Tabela 11 – Cluster 2 75
Tabela 12 – Cluster 3 75
Tabela 13 – Cluster 4 75
Tabela 14 – Cluster 5 76
Tabela 15 – Cluster 6 76

SUMÁRIO

1 INTRODUÇÃO 21

2 REFERENCIAL TEÓRICO 27
2.1 Métodos Estatísticos Não Supervisionados 27
2.2 Otimização do Índice de McClain 28
2.3 Clusters 29
2.4 Índice de McClain 32
2.5 Contribuições para Literatura de Crime 33

3 BASE DE DADOS 35

4 METODOLOGIA - ESTRATÉGIA EMPÍRICA 39
4.1 Criação de um índice de criminalidade baseado nos crimes definidos pelo SIM 40
4.2 Definição da métrica de semelhança 41
4.3 Criação dos Clusters 41
4.4 Otimização numérica 41
4.5 Escolha Intuitiva 43
4.6 Classificação dos Clusters - OLogit 43

5 RESULTADOS 45
5.1 Matriz de similaridade 45
5.2 Dendograma 49
5.3 Índice de McCLain 51
5.4 Clusters 54
5.5 OLogit 56

6 CONCLUSÃO 65
6.1 Considerações Finais 65
6.2 Trabalhos Futuros 65

REFERÊNCIAS 69

1 INTRODUÇÃO

O Brasil possui posição insatisfatória no *Global Peace Index*[1], ocupando apenas a 105ª posição em um ranking com 163 países. Embora o Rio de Janeiro figure entre os estados brasileiros que registraram queda nos indicadores de letalidade violenta [2], no período 2004-2014, o estado apresenta uma retomada no crescimento dos índices de criminalidade nos últimos dois anos. Além disso, o Rio possui histórico [3] de fortes organizações criminosas como o Comando Vermelho (CV), o Amigos dos Amigos (ADA), o Terceiro Comando Puro (TCP) e atuação de milícias.

O último responsável por liderar a Secretaria de Segurança (SESEG) do Rio de Janeiro, José Mariano Beltrame, passou dez anos no cargo. Historicamente, ser secretário da SESEG é extremamente desafiador, pois o cargo cobra protagonismo em atritos com organizações criminosas, fator gerador de alta rotatividade. Podemos destacar na gestão de Beltrame a criação das UPP, um dos maiores programas de pacificação urbana em atividade no mundo, além da implantação do Sistema Integrado de Metas (SIM), que foi adotado pela PMERJ (Policia Militar do Estado do Rio de Janeiro) a partir do ano de 2009 [4]. O SIM gera bonificação aos policiais militares sujeita à melhoria dos indicadores criminais[5]de seus respectivos batalhões.

Após o decreto do SIM, temos evidência de quais os crimes são julgados como prioridade, pelos agentes de segurança pública do estado. Sendo eles: **Roubo de Rua -** roubo a transeunte, roubo em coletivo, roubo de celular; **Letalidade Violenta -** homicídio doloso, latrocínio, lesão corporal seguida de morte, autos de resistência; **Roubo de Veículos**.

Este trabalho visa criar um indicador próprio de criminalidade do estado, utilizando não apenas os crimes da meta do SIM, mas também os registros de **furto** e de **roubo de carga**. Acrescenta-se os registros de furto por seu perfil semelhante aos roubos de rua, porém sendo um crime de natureza menos violenta, o que pode contribuir para diferenciar, em termos de níveis de violência, regiões com padrão semelhante de roubo de rua. Já os registros de roubo de carga são acrescentados para diferenciar batalhões que possam apresentar facilidade do desenvolvimento de rotas de roubo de carga, como por exemplo, batalhões que operam próximos a rodovias e vias de acesso das cidades.

---

[1] Metodologia de cálculo do índice em: *Global Peace Index 2016 Report* - Institute for Economics and Peace.

[2] Atlas da Violência de 2016 - IPEA.

[3] Ver: *O dono do morro: Um homem e a batalha pelo Rio* - Glenny [2016]; *Todo dia é segunda-feira* - Beltrame [2014]; *Quatrocentos contra um - A história do Comando Vermelho* - Lima [2010]; *Falcão - Meninos do Tráfico* - Athayde, MV Bill [2006]; *Abusado - O dono do Morro Dona Marta* - Barcelos [2003]; *Notícias de uma Guerra Particular* - Sales, Lund [1999].

[4] Nota N41931 da ALERJ - Lei que institui a criação do SIM.

[5] A SESEG cálcula o índice do SIM e define quais são as metas a se bater semestralmente.

A partir da criação desse indicador de criminalidade, propõe-se modelos que possam auxiliar políticas de prevenção e melhor interpretação das tendências criminais no estado do Rio de Janeiro. Com isso, podemos destacar as perguntas: Há semelhança na tendência e no nível dos crimes dos batalhões do estado do Rio de Janeiro? Podemos encontrar métodos não supervisionados que captem essas semelhanças? Quais são as características que explicariam distintos níveis de criminalidade? É possível contribuir para o desenho do Sistema Integrado de Metas, enquanto apresenta-se um indicador próprio de incidência criminal? Além disso, será que através de uma análise de padrões criminais somos capazes de auxiliar na função[6] desempenhada pelas Regiões Integradas de Segurança Pública?

Ao olharmos para as séries de tempo, é possível verificar que a tendência criminal do estado possui comportamento variado em diferentes batalhões analisados em separado, como podemos ver na figura1 que expõe a tendência de roubo a transeunte para todos os batalhões do estado em sombreado, com destaque para os batalhões 20 e 23.

[6] 'As Regiões Integradas de Segurança Pública - RISP objetivam a articulação territorial regional, no nível tático, da PCERJ com a PMERJ. A adequação geográfica entre as circunscrições territoriais de atuação das Polícias, no contexto das RISP, se consolida em termos práticos ao nível dos Departamentos de Polícia de Área - DPA da PCERJ e dos Comandos de Policiamento de Área – CPA da PMERJ. Os Diretores dos DPA e os Comandantes dos CPA, além das atribuições internas inerentes às suas respectivas instituições, também são responsáveis pelo estabelecimento de estratégias de integração e cooperação regionais; pela instituição de um fórum permanente de análise, compartilhamento de informações e ações conjuntas; pela adequação dos recursos humanos e logísticos às necessidades regionais; pelo acompanhamento e avaliação das ações realizadas; assim como pela promoção de uma rotina de reuniões e monitoramento do cumprimento das metas operacionais e administrativas pertinentes à sua região.' Retirado de `http://www.isp.rj.gov.br/Conteudo.asp?ident=38`.

Tendência de Roubo a Transeunte para os batalhões 20 e 23. As cores representam a proporção de Roubos de Rua no estado.

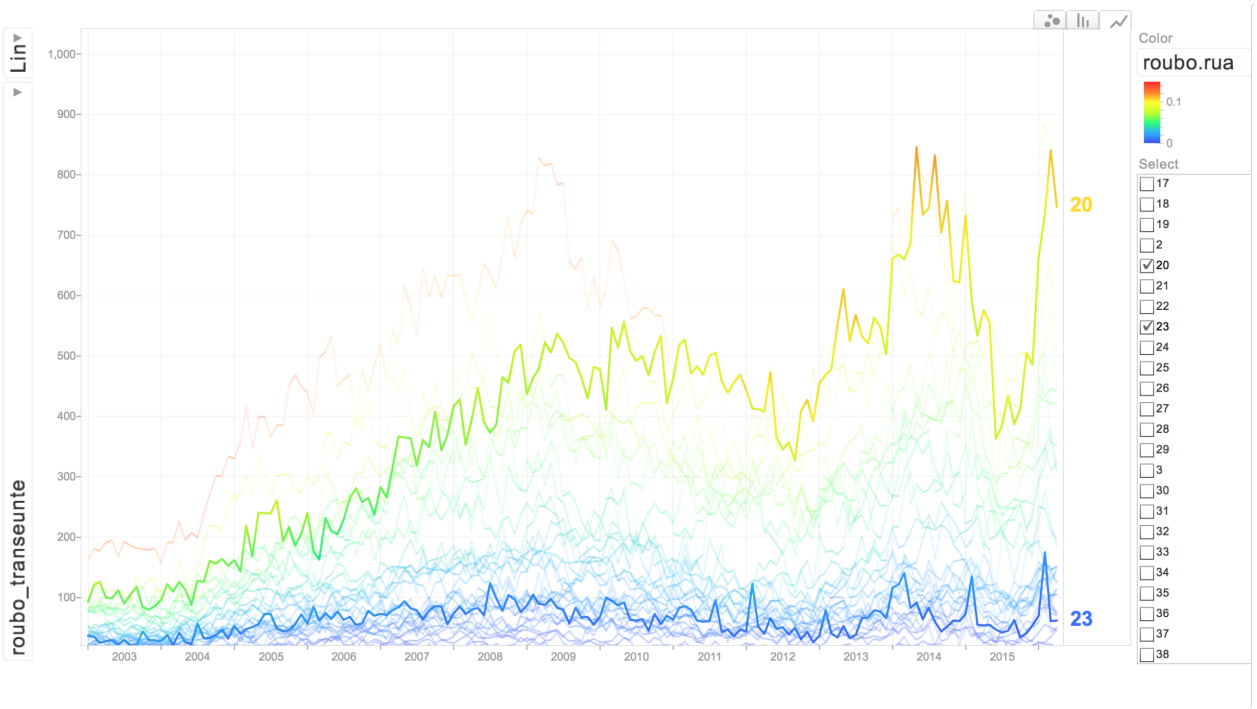

Figura 1 – Roubo Transeunte x Tempo

É importante destacar o caráter desigual do território fluminense, fator determinante no comportamento criminal. Enquanto o batalhão 23 possui em sua jurisdição bairros nobres da Zona Sul do Rio de Janeiro, com valores de IDH semelhantes ao da Suíça [7], o batalhão 20 compreende bairros na região da Baixada Fluminense, que possuem IDH ligeiramente superior ao brasileiro[8].

O gráfico2 abaixo mostra a evolução do índice de criminalidade construído, cuja fórmula X_{it} será discutida em mais detalhes no capítulo 4 desse projeto. As imagens tornam mais claras as particularidades do padrão criminal de cada batalhão e como algumas séries são mais similares entre si, comparativamente às demais.

[7] A Suíça em 2010 ocupava a 13ª posição na lista de IDH mundial, que possui 169 países. O valor do índice foi de 0.874.

[8] O Brasil em 2010 ocupava a 73ª posição na lista de IDH mundial, que possui 169 países. O valor do índice foi de 0.699.

Índice de Criminalidade para os batalhões 20 e 23

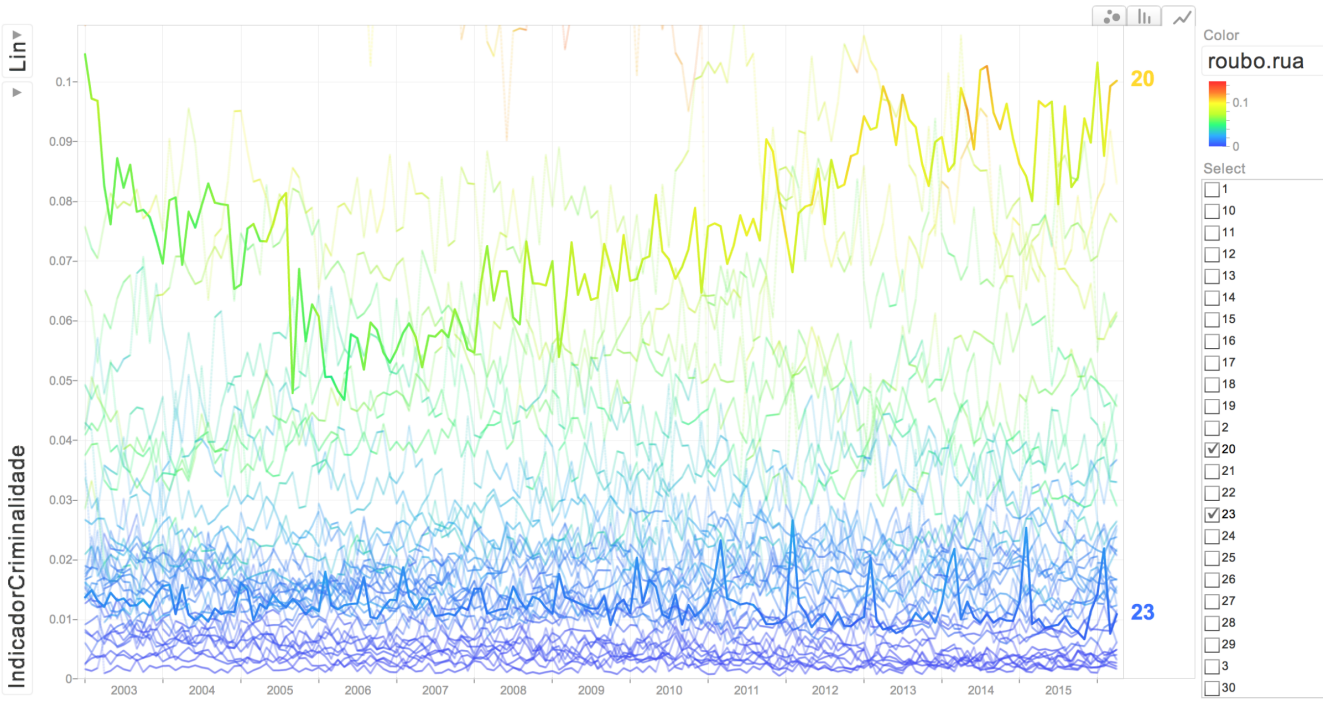

Figura 2 – Índice de criminalidade x Tempo

O que move esse trabalho é a oportunidade de estar em contato com um problema real, ainda pouco explorado pela academia de matemática aplicada [9] e com alta possibilidade de impactos sociais positivos. Além disso, o problema se mostra sofisticado do ponto de vista teórico, por demandar utilização de técnicas ainda não muito exploradas pela literatura dominante em criminologia e economia política. Isso, pois a obtenção de dados se mostra uma tarefa difícil ao envolver informações sigilosas entendidas como sensíveis à corporação. Logo, além da possibilidade de contribuição acadêmica, o projeto busca impacto prático, justamente por ser realizado em parceria com o Instituto de Segurança Pública (ISP)[10].

O problema investigado nesse trabalho pode ser dividido em dois campos de interesse da Matemática Aplicada. O primeiro é o de métodos estatísticos não supervisionados[11],

[9] "Há uma extensa literatura de matemática aplicada desenvolvida a partir de problemas oriundos das ciências biológicas e físicas. Nosso entendimento dos problemas das ciências sociais, de um ponto de vista matemático, é menos desenvolvido mas também apresenta alguns problemas muito interessantes, especialmente para jovens pesquisadores."Apresentação do curso *Mathematics of Crime* ministrado por Andrea Bertozzi na UCLA.

[10] 'O Instituto de Segurança Pública (ISP) é uma autarquia vinculada diretamente à Secretaria de Estado de Segurança do Rio de Janeiro. Sua missão é produzir informações e disseminar pesquisas e análises com vistas a influenciar e subsidiar a implementação de políticas públicas de segurança e assegurar a participação social na construção dessas políticas.' Extraído do site do ISP `http://www.isp.rj.gov.br/Conteudo.asp?ident=1`

[11] Escolhe-se o algoritmo - aqui o de criação de clusters -, apresenta-se evidência de que os dados são passíveis de serem subdivididos, enquanto o grau de similaridade é priorizado.

o qual parte-se de uma amostra de dados e busca-se, através da definição de uma função de densidade, criar classificações. O segundo é o de métodos de aprendizagem, área que busca extrair, como o nome diz, aprendizagem de um determinado conjunto de dados[12].

De posse da separação dos batalhões em clusters[13], pode-se buscar o que fundamenta essa semelhança entre batalhões e, mais ainda, prever quando terá boa (má) performance no tempo, em relação aos demais. Ao trocar de cluster, um batalhão pode estar melhorando (piorando) sua classificação em termos do indicador de criminalidade, introduzido por este trabalho, por passar a fazer parte de um novo grupo com melhor (pior) performance.

Resumimos então o exercício aqui proposto em quatro passos principais:

- Passo 1: Criação da matriz de similaridade, levando em consideração tendência e nível dos crimes por batalhão;
- Passo 2: Geração dos clusters por *clusterização hierárquica*;
- Passo 3: Análise de medidas de qualidade dos clusters criados, de modo a escolher o número ótimo de clusters, utilizando o índice de McClain;
- Passo 4: Construção de um modelo de classificação a partir dos clusters gerados nos passos anteriores. O objetivo é compreender a relação entre variáveis características selecionadas e a probabilidade na mudança de cluster.

A presente dissertação está dividida em seis capítulos. O primeiro, introduz e contextualiza o tema a métodos de matemática aplicada que possam ser úteis à análise da dinâmica criminal do estado do Rio de Janeiro. O segundo, expõe o problema teórico que buscamos resolver e posteriomente discute, em separado, cada modelo necessário para análise. Já o terceiro apresenta a construção do banco de dados, relacionando as informações do ISP,do Atlas do Desenvolvimento[14] e as fornecidos pela EMAp[15]. O quarto capítulo indica a metodologia utilizada, desde a estruturação do problema até a resolução do modelo por meio da estratégia empírica adotada. O quinto discute os resultados do exercício e, por fim, o capítulo seguinte apresenta a conclusão do estudo.

[12] Temos o problema de rodar um modelo paramétrico. Aqui, pode-se buscar melhor acurácia, mesmo que com paramentos ilegíveis do ponto de vista analítico, mas se opta por não ter a acurácia como objetivo principal, mas sim um modelo útil do ponto de vista de interpretação.

[13] Criação de cluster é um técnica de amostragem quando agrupamentos "naturais"mas relativamente heterogêneos são evidentes em uma população estatística.

[14] Ver descrição na seção de Base de dados.

[15] Escola de Matemática Aplicada da FGV.

2 REFERENCIAL TEÓRICO

Esse trabalho utiliza modelagem estatística e modelagem de dados como principais fontes teóricas. Busca-se, através da criação, avaliação e classificação de cluster em séries de tempo, extrair aprendizagem dos padrões criminais do estado do Rio de Janeiro.

Para consolidarmos a literatura encontrada neste projeto, houve uma exploração de uma série de modelos de aprendizagem supervisionados, estudados exaustivamente na área de econometria e aprendizagem por máquinas. No entanto, por incapacidade de encaixar os modelos à realidade enfrentada, percebeu-se a necessidade de utilização de métodos não supervisionados, com o objetivo de entender os padrões de crimes do estado, baseado nos graus de similaridade dos batalhões. Na área de métodos não supervisionados há uma grande variedade de modelos, sobretudo de criação de clusters. Através da criação de rótulos para os clusters, houve a facilitação do uso de um modelo classificatório que gerasse o entendimento de quais variáveis características apresentariam evidência em termos de probabilidade da permanência de um determinado batalhão em um cluster vis à vis os outros encontrados na primeira parte do exercício. Os passos do exercício são melhor explicados na parte metodológica deste projeto.

2.1 MÉTODOS ESTATÍSTICOS NÃO SUPERVISIONADOS

Em muitos estudos empíricos[1] há a predominância do uso de métodos supervisionados. O que caracteriza um método supervisionado é a presença de variáveis resposta $Y = (Y_1, ..., Y_m)$ para um dado grupo de variáveis características $X^T = (X_1, ..., X_p)$. Dito de outra forma, possui-se o conjunto p-uplas $x_i^T = (x_{i1}, ..., x_{ip})$ para cada indivíduo $i \in (1, .., N)$ a fim de explicar o conjunto de valores y_i, com a entrada do modelo de aprendizagem, o conjunto populacional $(x_1, y_1), ..., (x_N, y_N)$. Utilizam-se os dados amostrais a fim de criar um modelo preditivo, que tenha o objetivo de se aproximar da esperança condicional $E[Y_i|X_1, ..., X_p]$. Leva-se então em consideração uma função perda $L(y, \hat{y})$. Um exemplo, muito utilizado em diferentes áreas de aplicação, para a função perda seria $L(y, \hat{y}) = (y - \hat{y})^2$.

Matematicamente, supondo que (X, Y) são variáveis aleatórias representadas por uma densidade de probabilidade $Pr(X, Y)$, a aprendizagem supervisionada seria representada formalmente como um problema de estimativa da densidade condicional $Pr(Y|X)$. Busca-se, a partir desse passo, a estimativa do parâmetro θ que minimize o erro esperado, aqui

[1] Regressões lineares, modelos SVM e redes neurais podem ser exemplos de modelos abordados em estudos empíricos das áreas físicas, biológicas, sociais, dentre outras, onde busca-se o melhor entendimento de uma variável objetivo através de variáveis características.

representado por $\mu(x)$, para cada valor de x, como segue:

$$\mu(x) = argmin_{\theta} E_{Y|X} L(Y, \theta)$$

.

Em termos de estimativa de função de densidade de probabilidade, tem-se:

$$Pr(X, Y) = Pr(Y|X) \ . \ Pr(X).$$

Aqui, $Pr(X)$ é a densidade conjunta marginal apenas dos valores de X. No caso de aprendizagem supervisionada, tipicamente, não há uma direta preocupação com a densidade marginal dos valores de X, fazendo o foco da análise ser transferido às propriedades da função de densidade condicional $Pr(Y|X)$. Dessa forma, como a variável Y é geralmente de baixa dimensão [2] e o foco da análise está na estimativa de $\mu(x)$, enquanto minimiza-se a função perda, a análise se torna simplificada. Modelos dessa natureza possuem inúmeras abordagens e aplicações consolidadas em diferentes áreas de pesquisa como Ciência Política, Biologia, Administração e Economia.

Porém, para o problema real enfrentado nesse projeto, não há variável resposta, o que muda teoricamente a forma de pensar o problema, e nos leva à aplicação de métodos estatísticos não supervisionados. Sendo assim, seja um conjunto de variáveis aleatórias oriundas de um vetor X com N observações $(x_1, x_2, ..., x_N)$ com distribuição conjunta $Pr(X)$. Agora sem a presença de um conjunto de variáveis que supervisionam o modelo e garantem a presença de uma distribuição de erros, busca-se a inferência das propriedades da densidade conjunta $Pr(X)$.

A seguir apresentamos como será realizada a estimativa de $Pr(X)$ bem como sua análise.

2.2 OTIMIZAÇÃO DO ÍNDICE DE MCCLAIN

O método não supervisionado escolhido para esse projeto foi o de criar clusters das séries de tempo dos crimes [3] dos batalhões da PMERJ. A estratégia de criação dos clusters pode ser vista no problema de otimização a seguir, onde se busca os números ótimos q de clusters e do parâmetro k, que minimizem o índice de McClain. Esse índice foi escolhido, para avaliar a qualidade das subdivisões obtidas após a clusterização. Através da definição de uma medida de qualidade para os clusters, definimos um critério objetivo que nos guiará na construção dessas subdivisões. O parâmetro k é responsável por calibrar a intensidade dada à semelhança em nível das séries criminais dos batalhões, no cálculo

[2] A maioria dos casos de métodos supervisionados a variável objetivo possui dimensão um. Regressões lineares são um exemplo.

[3] Roubos de Rua, letalidade, roubo de veículos, furtos e roubos de cargas.

de quão similares são os batalhões. Dessa forma, segue o problema de otimização:

$$\begin{aligned}
&\min_{k,q} McClain(M,k,q) \\
&\text{s.a} \\
&2 < q < n-1,\ q \in B = (1,\ldots,n)\ , \\
&k \geq 0, k \in\ \mathbb{R}
\end{aligned} \tag{2.1}$$

$$\begin{aligned}
McClain &= \frac{\hat{S}_w}{\hat{S}_b} = \frac{\frac{S_w}{N_w}}{\frac{S_b}{N_b}} \\
S_w &= \sum_{w=1}^{q} \sum_{\substack{i,j \in C_w \\ i<j}} f(M_{T,i}, M_{T,j}) \\
S_b &= \sum_{m=1}^{q-1} \sum_{l=m+1}^{q} \sum_{\substack{i \in C_m \\ j \in C_l}} f(M_{T,i}, M_{T,j})
\end{aligned} \tag{2.2}$$

$$\begin{aligned}
&M = [m_{t,j}] = \phi_k[CORT(X_{j,t}, X_{B_{-j},t})]\ .\ d(X_{j,t}, X_{B_{-j},t}), \\
&t \in T,\ \forall j \in B \\
&\phi_k(u) = \frac{2}{1+exp(ku)}, k \geq 0
\end{aligned} \tag{2.3}$$

Para abordar a ideia aqui apresentada de forma mais clara, opta-se por explicar cada passo do problema de otimização enunciado acima em seções separadas deste capítulo de referencial teórico.

2.3 CLUSTERS

Para proceder na utilização de algoritmos de cluster, o primeiro passo consiste na escolha da função responsável por calcular a simetria entre os elementos $(X_{T1}, X_{T2}, ..., X_{Tn})$. A distância aqui escolhida apresenta um grau de subjetividade. Fazendo um paralelo com métodos supervisionados, a escolha do critério de distância é como escolher a função perda, responsável por calcular os resíduos.

A pergunta crucial na análise de clusters é: O que entendemos como similar em relação aos dados tratados em nosso objeto de análise? No contexto deste projeto, em específico, temos um agravante na análise, que é a dimensão temporal, o que torna o problema mais sofisticado[4].

Com isso, a função de similaridade escolhida seria um índice adaptativo cobrindo a proximidade tanto nos valores em nível quanto em sua tendência [5]. Uma medida responsável por levar em consideração tanto fórmulas convencionais de distância como a correlação

[4] Medidas de Simetria utilizadas tipicamente em análises de cluster convencionais não trabalham adequadamente no tempo pois ignoram interdependência temporal entre os valores - Montero, Vilar [2014].

[5] Pacote TSclust disponível no R - Implementação.

temporal foi introduzida por Douzal Chouakria e Nagabhushan [2007]. A função de proximidade, discustida aqui, calcula o valor de proximidade entre o comportamento dinâmico de duas séries através dos coeficientes de correlação de primeira ordem temporal, definido por:

$$CORT(X_T, Y_T) = \frac{\sum_{t=1}^{T-1}(X_{t+1} - X_t)(Y_{t+1} - Y_t)}{\sqrt{\sum_{t=1}^{T-1}(X_{t+1} - X_t)^2}\sqrt{\sum_{t=1}^{T-1}(Y_{t+1} - Y_t)^2}}$$

É fácil ver que $CORT(X_T, Y_T)$ pertence ao intervalo $[-1, 1]$. Essa componente é responsável por exprimir a similaridade dinâmica das séries. Uma interpretação despretenciosa dos valores extremos facilita o entendimento da função dessa componente. Quando $CORT(X_T, Y_T) = 1$ tem-se que ambas as séries apresentam comportamento dinâmico similar, isto é, em direção e taxa. Já no caso em que $CORT(X_T, Y_T) = -1$ temos crescimento em taxa similar mas em direções opostas. Por fim, no caso em que $CORT(X_T, Y_T) = 0$, pode-se dizer que não há monotonicidade entre X_T e Y_T, mais ainda, que as taxas de crescimento entre ambas são linearmente independentes do ponto de vista estocástico. Define-se desta forma o cálculo do índice de distância do elemento j a todos outros elementos em B de acordo com a função a seguir:

$$d_{CORT}(X_{j,T}, X_{B_{-j},T}) = \phi_k[CORT(X_{j,T}, X_{B_{-j},T})] \; . \; d(X_{j,T}, X_{B_{-j},T}),$$

Onde $\phi_k(.)$ é uma função monótona e adaptativa responsável por suavizar a distância de tendência dos dados em questão, cuja fórmula é:

$$\phi_k(u) = \frac{2}{1 + exp(ku)}, k \geq 0$$

como na equação (2.3).

Percebe-se que a interação entre a medida dinâmica e de nível nos leva ao nosso índice de distância entre os processos estocásticos em análise, sendo o fator k responsável por calibrar qual das componentes mais levamos em consideração, isto é, nível ou tendência.

No que diz respeito às medidas de distância de nível aqui destacam-se as mais usuais, a euclideana $d_2(X_T, Y_T) = \sum_{t=1}^{T}(Y_t - X_t)^2$ e a distância de Minkowski $d_{L_q}(X_T, Y_T) = (\sum_{t=1}^{T}(Y_t - X_t)^q)^{\frac{1}{q}}$. Neste trabalho, optamos pela distância euclideana a fim de reduzir a complexidade do problema[6].

Ao receber a matriz de simetria como entrada, os algoritmos utilizados para criar os clusters tem o objetivo de gerar segmentações nos dados. Dentre os algoritmos de criação de tais clusters, pode-se destacar três grandes famílias: Algoritmos combinatórios, Modelagem de Misturas e *Mode Seeking*.

Os *Algoritmos combinatórios* trabalham diretamente com o conjunto de dados observados, não havendo referência a nenhuma função de densidade de probabilidade. Já

[6] Caso utilizássemos a distância de Minkowski, acrescentaríamos mais uma variável para o problema de otimização, sendo o valor q utilizado por esta função distância.

a *Modelagem de Misturas* supõe que os dados são independentes e identicamente distribuídos, gerados a partir de uma função de densidade de probabilidade. Os clusters são obtidos através de uma estimativa paramétrica da mistura de diferentes funções de densidade de probabilidade de modo a obter um bom *fit* do comportamento dos dados. A estimativa do método se dá através do cálculo da Máxima Verossimilhança e sua avaliação se dá através de uma abordagem *bayseana*, levando em consideração o BIC[7] de cada modelo estimado[8].

Existe uma série de critérios que objetivam a criação de clusters, vide a citação das três *famílias tradicionais* acima, com alta variedade de modelos que abordam sua criação de diferentes formas. Neste projeto, escolhemos a família de algoritmos combinatórios, em particular o modelo de clusterização hierárquica sob o critério *complete*, representado pela função $f(.)$ na equação (2.2). A razão da escolha do modelo e do critério utilizados em $f(.)$ [9] será melhor explicada no capítulo 4, o qual apresenta a metodologia. No entanto, podemos adiantar que dois fatores principais nortearam a escolha do modelo hierárquico, sendo eles o fato de ser um algoritmo determinístico, o que garante que cada elemento pertence a necessariamente um cluster apenas, o que no caso estocástico é flexibilizado e o segundo fato está na essência da função de clusterização hierárquica, que só toma como entrada os dados e o critério de distância a ser utilizado, o que evita a necessidade de escolha a-priori do número de clusters. Além de reduzir a complexidade, reduz a arbitrariedade da análise, quando comparado a modelos que demandam o número de clusters também como entrada.

Algoritmos famosos como *k-means* e *k-medoids* [10] demandam o número inicial q de clusters como entrada do modelo, enquanto a clusterização hierárquica não demanda tal especificação. Na clusterização hierárquica, por outro lado, é necessária a especificação de distância de similaridade $d(x_a, x_b)$, $a \in A$ e $b \in B$, com A e B sendo diferentes clusters, que é aplicada a cada iteração do algoritmo [11]. O cálculo se inicia com todos os dados [12], realizando cada iteração ao utilizar o critério de distância, criando os clusters através de uma abordagem *divisiva*. Por outro lado, a abordagem pode ser *aglomerativa*, iniciando as iterações em cada indivíduo em separado [13], utilizando o critério de distância

[7] $BIC = -2.ln\hat{L} + k.ln(n)$, com $\hat{L} = p(x|\hat{\theta}, M)$ o valor da máxima verossimilhança do modelo M, onde $\hat{\theta}$ é o parâmetro que maximiza a função de densidade de probabilidade do modelo M. O valor x são os dados observados, o valor n é o total de observações de x e k é o número de parâmetros a serem estimados.

[8] Encontra-se a abordagem via *Mode Seeking* em *Hastie, Tibshirani, Friedman - The Elements of Statiscal Learning.*

[9] Critério de uma função criadora de clusters é o cálculo aplicado à matriz de semelhança a cada iteração do algoritmo.

[10] A partir da escolha do número de clusters, são lançados valores aleatórios iniciais pra serem centroides dos clusters, que são recalculados a cada iteração do modelo. Para *k-means* os centroides são a média dos valores pertencentes aos clusters enquanto o algoritmo de *k-medoids* escolhe o centroide como sendo o elemento dentro do cluster com menor distância a todos elementos pertencentes ao mesmo.

[11] Um exemplo seria a abordagem *complete*: $d(x_a, x_b) = max(a, b)$.

[12] Caso top-down.

[13] Caso bottom-up.

de modo a aglomerar os dados a cada iteração.

Percebe-se então uma espécie de hierarquia, em termos de semelhança, uma vez que a definição dos clusters são atualizadas a cada iteração. A escolha do nível hierárquico que define o momento de parada do algoritmo e consequentemente retorna os clusters, pode se dar de forma subjetiva. Uma série de índices pode ser útil nessa função[14] de modo a escolher os clusters de forma objetiva, isto é, sujeita a um critério numérico.

Um fator que faz o método hierárquico ser famoso na definição dos clusters é sua clareza em expor o processo de aprendizagem a que os dados são expostos. Uma parte dos métodos divisivos, quando vistos na abordagem *Bottom-up*, e os métodos que aglomeram os dados possuem propriedade monótona, isto é, ao aglomerar (dividir) os dados em cada nó, faz-se de modo a reduzir (aumentar) a semelhança dos elementos pertencentes aos grupos resultantes. A árvore resultante desse processo é chamada de *dendograma*. Um exemplo pode ser visto na seção de resultados na figura 6.

2.4 ÍNDICE DE MCCLAIN

Como discutido anteriormente, a definição do número de clusters pode ser feita de forma intuitiva no modelo hierárquico[15]. Porém, além da escolha intuitiva, opta-se nesse trabalho por utilizar o índice de McClain, de McClain e Rao [1975], pra balizar o número ótimo de clusters. O índice é calculado segundo a fórmula a seguir:

$$McClain = \frac{\hat{S}_w}{\hat{S}_b} = \frac{\frac{S_w}{N_w}}{\frac{S_b}{N_b}},$$

com a soma total das distâncias intra-cluster

$$S_w = \sum_{w=1}^{q} \sum_{\substack{i,j \in C_w \\ i<j}} f(x_i, x_j),$$

e a soma total das distâncias entre os clusters

$$S_b = \sum_{m=1}^{q-1} \sum_{l=m+1}^{q} \sum_{\substack{i \in C_m \\ j \in C_l}} f(x_i, x_j).$$

O objetivo de utilização do índice é ter uma medida de qualidade para os clusters extraídos da modelagem e essa qualidade se traduz em menor distância interna entre os elementos presentes em cada cluster relativa à distancia de cada cluster entre si.

Como é possível ver na equação (2.1), busca-se a otimização de tal índice para obter evidência de que os dados são passíveis de clusterização e, mais ainda, retornar a árvore ótima do algoritmo de clusterização hierárquica.

[14] Nesse projeto utiliza-se a otimização do índice de McClain, além de uma abordagem intuitiva na análise final.

[15] Define-se em qual altura da árvore será feita a parada de iterações e consequente o retorno do número de clusters.

MODELO CLASSIFICATÓRIO - OLOGIT

De posse de uma variável categórica dependente ordenada Y_i para observação $i \in (1, ..., N)$, temos o *componente estocástico* como Y_i^*, que segue uma distribuição logística padrão com parâmetro μ_i,

$$Y_i^* \sim Logit(y_i^* | \mu_i)$$

a qual submetemos ao mecanismo de classificação

$$Y_i = k \ se \ \pi_{k-1} \leq Y_i^* \leq \pi_j \ para \ k \in (1, ..., q).$$

onde $\pi_l \ com \ l \in (0, ..., q)$ são os parâmetros *threshold* com as restrições $\pi_l < \pi_m \forall \ l < m$ e assumindo os possíveis valores $\pi_0 = -\infty \ e \ \pi_k = \infty$.

Com isso, para realizarmos a inferência, tomamos como dados os parâmetros π_k e β, além das variáveis características x_i e calculamos:

$$Pr(Y \leq k) = Pr(Y^* \leq \pi_k) = \frac{exp(\pi_k - x_i\beta)}{1 + exp(\pi_k - x_i\beta)},$$

o que implica que:

$$\pi_k = \frac{exp(\pi_k - x_i\beta)}{1 + exp(\pi_k - x_i\beta)} - \frac{exp(\pi_{k-1} - x_i\beta)}{1 + exp(\pi_{k-1} - x_i\beta)}.$$

Essa modelagem será pertinente após a extração e ordenamento dos clusters, pois uma vez que obtenhamos os clusters, vamos buscar ordenamento nos mesmos, caso seja encontrado, seremos aptos a definir rótulos para esses clusters ordenados, de modo que, a variável Y^* seja o rotulo ordenado dos clusters em função da incidência criminal dos mesmos e as variáveis características x serão apresentadas na sessão de resultados.

2.5 CONTRIBUIÇÕES PARA LITERATURA DE CRIME

Na literatura de ciência da computação e tecnologia podemos encontrar em Malathi e Baboo [2011] a aplicação de aprendizagem não supervisionada para identificar estados semelhantes em termos de padrões criminais na Índia e com isso facilitar o desenvolvimento de um modelo de classificação criminal de maior acurácia. No entanto, a utilização do *k-means* possui duas ressalvas principais: a necessidade de escolha de um número de clusters como entrada o que pode limitar a descoberta de padrões a priori; a escolha de centroides aleatoriamente levando a diferentes resultados finais, podendo levar a inconsistência de análises a posteriori dos clusters obtidos. O artigo defende o DBscan como modelo mais apropriado para realizar a tarefa de criação dos clusters nos dados em que enfrenta [16]. Pode-se encontrar a aplicação de métodos não supervisionados em dados criminais também

[16] Aqui, como será visto na metodologia, opta-se por utilizar o modelo de clusters hierárquicos por aumentar nossa capacidade analítica.

em Babu et al. [2016], porém com abordagem de buscar o algoritmo eficiente do ponto de vista de complexidade computacional em sua aplicação.

Do ponto de vista analítico, destaca-se a teoria clássica de crime e punição enunciada por Gary Becker, que enfatiza o papel essencial do policiamento na prevenção criminal, justamente por aumentar a probabilidade de punir. Nessa linha, temos os estudos empíricos de Levitt [1997] e DiTella e Shargodsky [2004], que buscam identificar o impacto de mais policiamento na redução de incidência criminal. Tais estudos avaliam experimentos naturais, quando percebidos, ou estudos mais atuais buscam a realização de experimentos na linha de *Hot Spot policing* ou na análise de programas como a implantação da estratégia Kingpin em Lindo e Padilla-Romo [2015] no México bem como os estudos já citados de avaliação das UPP no estado do Rio de Janeiro. O ponto é que experimentos naturais em geral não possuem validade externa, isto é, os experimentos sob análise são muita das vezes fenômenos locais de difícil generalização a outros lugares expostos a diferentes condicionalidades.

O presente estudo posiciona-se nas discussões citadas como uma contribuição ao apresentar aplicação de método não supervisionado aos dados dos batalhoões do Rio de Janeiro com método facilmente replicável - como para as UPP por exemplo -. A também aplicação do modelo OLogit com dados de IDH e de efetivo policial a fim de classificar os clusters encontrados, mostra analiticamente a importância de variáveis de caráter socioeconômico e de força policial no entendimento de padrões criminais.

3 BASE DE DADOS

Os dados utilizados nesse trabalho foram extraídos do acervo do ISP, do Atlas do Desenvolvimento realizado pelo IPEA em parceira com o PNUD e a Fundação João Pinheiro e dos bancos de dados de projetos da EMAp. As bases estruturadas e utilizadas são: o painel de crimes e características locais por batalhão no estado; o painel de dados da alocação dos comandantes em cada batalhão no tempo; o volume de efetivos por batalhão no tempo; dados de população flutuante para os batalhões da capital e dados de IDH por batalhão para o ano de 2010.

As bases de dados aqui citadas exigem alta capacidade de processamento, por se tratarem de grandes bases e, com isso, possuem armazenamento total inviável enquanto modelos de pacotes estatísticos são implementados. Houve a necessidade de utilização de técnicas de mineração de dados para criação de um banco de dados no PostgreSQL, de modo que fosse possível reunir todas as fontes de dados citadas acima e possibilitar a utilização dos mesmos para análise.

Para entender a composição do banco de dados se faz necessária a leitura de um diagrama, presente na figura 3, que apresenta como as bases de dados se relacionam. Além do diagrama, será exposta uma descrição de cada base de dados aqui citada.

A base de crimes e características da região [1], como o nome já diz, expõe os crimes registrados em cada batalhão no tempo por mês, a partir de 2003 até 2016, além de possuir características locais dos batalhões como população domiciliar na área daquele batalhão e população do município o qual o mesmo se insere.

A base de comandantes [2] é composta por dados de alocação dos mesmos por mês do ano de 2005 até 2015. Do ponto de vista de mineração de dados, essa base se mostrou desafiadora, uma vez que não possuía uma chave clara de conexão com as outras bases do banco. A base de comandantes possui como indicadores primários o batalhão e o nome do comandante, enquanto as outras bases possuem como indicadores ou o batalhão ou o RG do policial em questão.

Faz-se necessária a utilização de algoritmos de *text mining*, aqui a distância de *Levenshtein* aplicados a *strings*, para criar um dicionário com nome do comandante e seu respectivo RG e, assim, ser possível conectar todas as bases do banco, facilitando as análises de interesse. Esse dicionário foi denominado de RG Comandante.

A base dos efetivos[3] é composta do número de efetivos pertencentes a cada batalhão em cada mês do ano de janeiro de 2008 até março de 2015.

No acervo do ISP, já se dispunham os valores de população domiciliar[4]. Como focamos

[1] No diagrama aparece com o nome de Crimes.
[2] No diagrama aparece como o nome de Comandantes.
[3] No diagrama é incorporada à iteração batalhão período, por simplificação.
[4] No diagrama representa-se tanto a população flutuante como a domiciliar como população.

Diagrama expositivo do banco de dados.

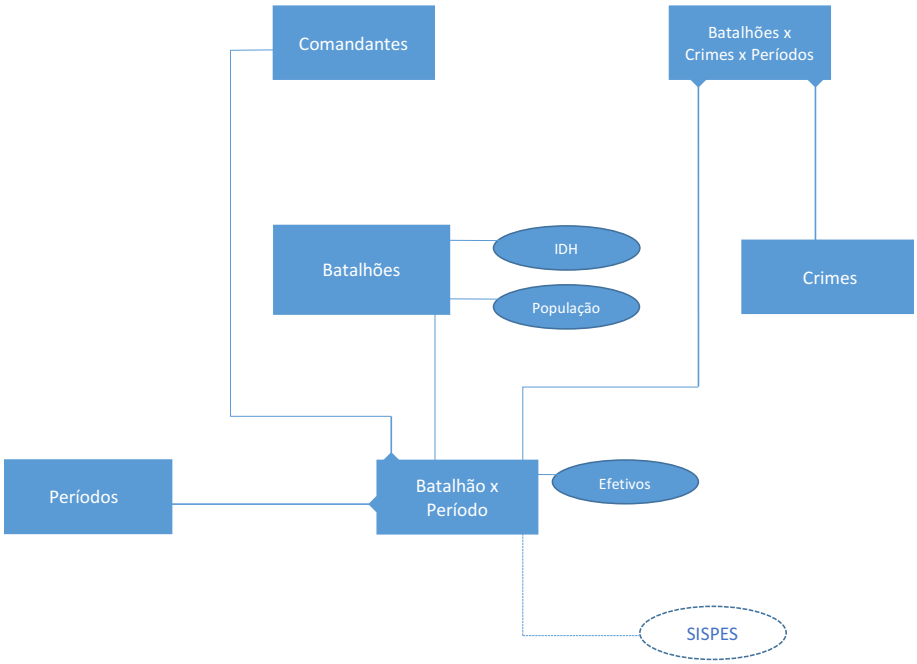

Figura 3 – Modelo de Entidades e Relacionamento do Banco de Dados

nossa análise nos crimes pertencentes ao sistema de metas da PMERJ, fez-se necessária a extração de dados de população flutuante. Utilizamos uma base de população flutuante baseada em dados de telefonia celular pertencente ao acervo de dados da EMAp. Buscamos a extração desses dados para evitar erros de mensuração populacionais, que acontecem ao utilizarmos dados de população domiciliar, uma vez que há bairros, como o Centro da capital, onde a população domiciliar é ínfima, mas a população flutuante pode ser de extrema relevância para análises locais. No entanto, essa extração só foi realizada para capital, o que impediu a utilização da mesma. Um passo futuro desse projeto será a extração da população flutuante para todo o estado.

Por fim, criamos uma matriz de dados de IDH (Índice de desenvolvimento humano) por batalhão, para completar o banco de dados. O cálculo foi possível após a criação dos dados de bairros por batalhão, a partir da última subdivisão [5], o que nos possibilita cruzar os batalhões com qualquer dados socioeconômicos[6] a nível de bairro.

Nesse projeto, em um primeiro momento, os dados de comandantes foram utilizados quase que exclusivamente, a fim de tentar extrair o impacto do trabalho dos comandantes nos indicadores criminais de batalhões do estado. Posteriormente, os dados criminais dos batalhões foram utilizados para definição e avaliação dos clusters. Por fim, foi criado uma matriz de dados que consolidasse os crimes, as características socioeconômicas, o cluster o qual se insere, o número de efetivos por 100 mil habitantes e nosso índice de criminalidade, presente na equação (4.1), para todos os batalhões, em todo período de tempo - i.e 07/2011 - 03/2015 - como pode ser visto na tabela de estatísticas descritivas 1, como segue:

Tabela 1 – Estatísticas Descritivas dos dados de Interesse para todos os batalhões

	Índice de Criminalidade	IDH Médio	IDH Médio de Educação	IDH Médio de Longevidade	IDH Médio de Renda	Efetivos por 100mil/hab.
Mínimo	0.00	0.68	0.58	0.80	0.66	55.39
Máximo	0.11	0.94	0.88	0.94	1.00	1376.17
Média	0.03	0.76	0.68	0.85	0.75	239.43
Desvio Padrão	0.02	0.07	0.08	0.04	0.08	230.20
Variância	0.00	0.00	0.01	0.00	0.01	52991.99
Mediana	0.02	0.74	0.67	0.83	0.73	154.43
N	1755.00	1755.00	1755.00	1755.00	1755.00	1755.00

É importante destacar que o banco criado para esse projeto ainda não foi explorado em sua totalidade. Algumas análises podem ser feitas futuramente a partir da atualização da modelagem exposta nesse trabalho, ou até mesmo de forma independente, pois o banco foi criado para facilitar qualquer estudo que cruze dados socioeconômicos com padrão criminal e características da polícia no estado do Rio de Janeiro.

[5] Em julho de 2011 batalhões passaram por redefinição de seu espaço geográfico, dando origem a novos batalhões, como o 41, dividindo e extinguindo batalhões antigos, como o 9, o 13 e o 1.

[6] Calcula-se os índices médio, mínimo, máximo e de variância do IDH por batalhão. Na 1 são expostos os dados de IDH médios.

4 METODOLOGIA - ESTRATÉGIA EMPÍRICA

Como exposto na introdução, o exercício desse projeto pode ser dividido em quatro passos. Nesta seção explora-se a estratégia empírica realizada para atendê-los. Assim, busca-se evidenciar a utilidade da aplicação de clusters nas séries temporais de crime, justificar as escolhas do métodos no desenvolvimento do exercício e apresentar a forma pensada para exploração dos clusters. Visa-se com esses passos responder perguntas presentes na introdução como: há semelhança na tendência e no nível dos crimes dos batalhões do estado do Rio de Janeiro? Podemos encontrar métodos não supervisionados que captem essas semelhanças?

Para aumentar nossa capacidade analítica, propõe-se uma classificação aos batalhões de modo a dividi-los em clusters, levando em consideração seus respectivos graus de semelhança de tendência e de nível, de modo a otimizar o índice de McClain. O processo de otimização é importante por mostrar evidência de clusterização dos dados. De posse da classificação, busca-se melhor compreender esses clusters de maneira socioeconômica através da aplicação de um modelo OLogit e, assim, possibilitar análises futuras como por exemplo: análise qualitativa dos batalhões; facilitar a captação do desempenho dos comandantes[1]. E, assim, responder as seguintes perguntas: Quais são as características que explicariam distintos níveis de criminalidade? É possível contribuir para o desenho do Sistema Integrado de Metas, enquanto apresenta-se um indicador próprio de incidência criminal? Além disso, será que através de uma análise de padrões criminais somos capazes de auxiliar na função desempenhada pelas Regiões Integradas de Segurança Pública?

No início do projeto, houve a preocupação de construir um banco de dados que facilitasse a relação de dados de criminalidade com variáveis socioeconômicas do estado do Rio de Janeiro, bem como variáveis características regionais e da polícia.

A partir disso, iniciamos o projeto de modo a construir uma análise do padrão criminal local, pois percebe-se uma diferença na tendência e no nível de criminalidade dos batalhões quando analisados em separado, vis à vis, aos padrões observados para o estado em agregado [2].

Em um primeiro momento, houve a busca pela utilização de um modelo linear de dados em painel, que fosse capaz de encorporar efeito fixo dos comandantes no nível do crime

[1] Uma estratégia seria verificarmos uma eventual mudança de cluster de alguns dos batalhões durante o trabalho de um dos comandantes, bem como identificarmos padrões de trabalho de determinados comandantes, baseado nos batalhões em que os mesmos desempenharam sua função de policial.

[2] Na figura 1, presente na seção de resultados, temos o caso em que os dados são agregados e há perda da dimensão temporal.

em análise [3]. Tinha-se o interesse em captar o efeito regional no nível de criminalidade, bem como uma medida de performance dos comandantes dos batalhões.

Dois problemas impediram com que tal análise fosse bem sucedida, sendo eles: nem todos comandantes exerciam o cargo em mais de um batalhão, o que impedia a extração de seu efeito no crime, em análise, com significância dos parâmetros; o outro problema seria a nossa incapacidade em definir objetivamente o que é trilhar uma carreira com méritos dentro da polícia[4].

Dessa forma, a seção metodológica discute os detalhes da implementação do já enunciado referencial teórico, apresenta o algoritmo de otimização numérica, argumenta a possibilidade de escolha intuitiva e por fim realiza a classificação dos clusters extraídos no exercício.

4.1 CRIAÇÃO DE UM ÍNDICE DE CRIMINALIDADE BASEADO NOS CRIMES DEFINIDOS PELO SIM

Para sintetizar o perfil criminal de cada batalhão, opta-se por acrescentar *furtos* e *roubos de carga* aos crimes já analisados pelo SIM: *roubo de rua, letalitade violenta, roubo de veículos*, como já introduzido. No entanto, o cálculo feito aqui se dá de forma diferente do que é feito pela SESEG. Leva-se em consideração a média das proporções do valor que cada batalhão representa no total de crimes do estado no mês em questão.

Escolhe-se esse formato, pois busca-se uma métrica que evite sub-estimativas (super-estimativas) do perfil criminal do batalhão, como pode acontecer em cálculos menos atentos, como por exemplo, os que levam em consideração a variação criminal entre períodos diferentes de tempo sem a realização dos controles devidos. O problema nesse caso está na possível crença de uma melhor (pior) sensação da incidência criminal naquele período na região de análise. Busca-se esse indicador para atender o desafio de lidar com diferentes níveis e tendências criminais no espaço amostral e assim contribuir para a atual metodologia utilizada pela SESEG no cálculo de incidência criminal.

Além disso, é relevante o cuidado com a unidade de medida, isto é, não podemos somar vítimas de crimes letais com as de crimes contra propriedade. Assim, o índice passa a ter a unidade de proporção criminal do batalhão i no estado, em um determinado mês t, como segue:

$$X_{it} = \sum_{j=1}^{r} \frac{\frac{X_{jit}}{\sum_{i=1}^{N} X_{jit}}}{R}, \ (1, \ldots, r) \ = \ lista\ de\ crimes. \tag{4.1}$$

[3] Aqui utilizamos, em separado, dados de roubo de rua, letalidade violenta, roubo de veículos para cada versão do modelo, sem a criação de um índice.

[4] Não somos capazes de definir quais os batalhões são mais desafiadores proporcionalmente aos demais do estado, nem mesmo se o policial que supostamente teve uma carreira de méritos foi premiado com um cargo mais próximo à alta gestão da segurança pública, por exemplo.

4.2 DEFINIÇÃO DA MÉTRICA DE SEMELHANÇA

De posse da medida de semelhança e da escolha do parâmetro k vistos em (2.3), tomam-se 39 [5] séries de tempo do indicador em (4.1), das quais calcula-se o grau de semelhança em suas séries de tempo batalhão a batalhão. Assim, cria-se a matriz de similaridade responsável por indicar a proximidade entre os batalhões, como segue na seção 5 de resultados.

4.3 CRIAÇÃO DOS CLUSTERS

Com a matriz de similaridade, aplica-se o algoritmo de criação de clusters hierárquicos. Deve-se indicar o método de partição a ser utilizado sobre os graus de similaridade apresentados entre os batalhões. O método escolhido em questão foi o *complete* visto anteriormente[6]. A escolha se deu de forma a atender o índice que buscamos otimizar, pois o método prioriza a alocação dos elementos em clusters de modo a garantir a distância máxima aos batalhões alocados em outros agrupamentos[7].

Através do resultado extraído do algoritmo de clusters hierárquicos, detém-se o dendograma, que nos permite escolher em qual altura da árvore realizaremos o corte, obtendo assim o número de clusters. Além da já citada solução ótima, contextualizamos intuitivamente a escolha de um número maior de agrupamentos.

É importante justificar a não utilização de algoritmos estocásticos [8]. Utilizamos séries de tempo detentoras de muito ruído, que incorporado na classificação estocástica, leva a grupos de batalhões que poderiam estar em mais de um cluster, isto é, em uma espécie de *quasi-cluster*. Em termos teóricos isso não seria um problema, uma vez que em muitos casos pode aumentar a acurácia da classificação como em Malathi e Baboo [2011]. No entanto, a capacidade analítica se mostra prejudicada[9], o que favorece a aplicação de um método com resultados determinísticos para o caso em questão.

4.4 OTIMIZAÇÃO NUMÉRICA

Aqui, apresenta-se o código em R para realização da otimização, bem como detalhes incorridos no processo.

---

[5] 39 é o número total de batalhões da PMERJ.

[6] Em nota de rodapé [11].

[7] Os métodos *ward.D*, *ward.D2*, *single*, *average*, *mcquitty*, *median*, *centroid* também foram testados através de simulação numérica, mas nenhum tão bem-sucedido como o método escolhido, como esperado.

[8] Testes foram implementados com o algoritmo EM, presente no apêndice, e também com o DBScan.

[9] Como buscamos explicar a ordenamento dos clusters e extrair utilidade prática do nosso resultado, extrair classificações que permitem ruído, reduziria confiança na análise.

A seguir temos o pacotes: TSClust utilizado para realizar o cálculo da matriz de similaridade ao tomar as séries de tempo de cada batalhão como entrada [10]; NbClust utilizado para extrair o número de clusters q de menor índice de McClain para um dado valor k [11]; plyr utilizado para processamento dos dados.

```
> library(TSClust)
> library(NbClust)
> library(plyr)
```

Toma-se a liberdade de escolha $\alpha = 100$ valores diferentes de k. É dado, de acordo com a função $\phi(.)$, presente em (2.3), que a partir de um certo valor α, a métrica de similaridade passa a dar apenas peso ao nível de semelhança da tendência, cálculo feito pela função *diss* do já citado pacote $TSClust$. Há a possibilidade de utilização de diferentes métodos, mas opta-se pelo método *complete* por sua coerência com o índice de avaliação de McClain [11].

Durante o *loop*, calculamos a matriz de simetria como em (2.3), realizamos a criação dos clusters hierárquicos através da função $hclust(.)$ [12] como em $f(.)$ na equação (2.2) e realizamos a otimização do índice de McClain através da função $NbCLust(.)$ como em (2.1).

```
> indice.NumeroClusters <- list()
> parametros.k <- seq(1, 100, by = 1)
> metodos <- c("ward.D", "ward.D2", "single", "complete", "average
  ", "mcquitty", "median", "centroid")
> for(k in 1:length(parametros.k))
> {
>    #Destaca-se que X_B_T e o conjunto de dados para todos os
  batalhoes B em todo o periodo de tempo T.
>    M <- diss(X_B_T,"CORT",k = paramteros.k[k])
>    hcortSIM <- hclust(dcortSIM, metodos[4])
>    indice.NumeroClusters[[length(indice.NumeroClusters) + 1]] <-
   NbClust(X_B_T, diss = M,distance = NULL ,min.nc = 2, max.nc =
  N, method = metodos[4], index = "mcclain")$Best.nc
> }
```

Nessa parte, extraímos os valores mínimos do índice de McClain com os respectivos valores q de clusters, para todos os valores de k. Busca-se o valor de q e k de menor índice de McClain, seguindo o objetivo de extrair a divisão de clusters com a soma das distâncias internas relativas a soma das diferenças externas de menor valor possível. Atenta-se o fato

[10] Mais detalhes em Pablo Montero, José A. Vilar [2014].
[11] Mais detalhes em Malika Charrad, Nadia Ghazzali, Véronique Boiteau, Azam Niknafs [2014].
[12] Este passo é relevante para o processo de escolha intuitiva, mas já é introduzido nessa parte, por simplificar o algoritmo.

de que a restrição apresentada em (2.1) é aplicada aqui de modo a garantir a evidência de que os dados são passíveis clusterização.

```
> Saidas.parametros.k <- ldply(lapply(indice.NumeroClusters,
  function(indice.NClusters) data.frame(NumeroCLusters = indice.
  NClusters["Number_clusters"], ValorIndice = indice.Nclusters["
  Value_Index"])), data.frame)
> k.otimo <- which(min(Saidas.parametros.k[which(2 < Saidas.
  parametros.k$'NumeroCLusters' & 38 > Saidas.parametros.k$'
  NumeroCLusters'),"ValorIndice"]) == Saida.parametros.k[,"
  ValorIndice"])
> q.otimo <- Saidas.parametros.k[k.otimo,"NumeroCLusters"]
> k.q.otimos <- c(k.otimo, q.otimo)
```

4.5 ESCOLHA INTUITIVA

Após extrair o número ótimo de clusters, busca-se a interpretação deles e percebe-se, intuitivamente, que é possível subdividir ainda mais o conjunto de dados com base no dendograma [13]. Destaca-se o cuidado em não escolher 2 clusters ou o número total de séries observadas. No primeiro caso a separação é pouco útil e no segundo é um caso de *overfitting*, como já exposto na restrição do problema de otimização. Aqui, abrimos mão da solução ótima de modo a aumentar nosso poder analítico.

4.6 CLASSIFICAÇÃO DOS CLUSTERS - OLOGIT

De posse dos clusters, pode-se extrair ordem dos mesmos, ao se levar em consideração tanto o valor médio quanto o valor da mediana de cada cluster. Após ordenar os clusters, podemos utilizar os dados de IDH por batalhão para imprimir interpretação socioeconômica dos grupos extraídos do exercício.

Assim, com os clusters ordenados, utilizamo-nos para supervisionar os dados característicos por batalhão. A partir desse passo, realizamos a estimativa dos parâmetros, com o objetivo de buscar evidência de que um maior grau de IDH e de efetivos por 100mil habitantes nos retorna uma probabilidade maior de estarmos em um cluster com menor grau de criminalidade.

[13] A correlação temporal deste exercício utiliza apenas o primeiro *lag* das séries dos indicadores criminais. Na conclusão deste trabalho apresentamos a possibilidade de alteração do *lag* na busca de melhorar a classificação.

5 RESULTADOS

Nesta seção, vamos apresentar os resultados encontrados, flexibilizando a modelagem que fizemos, para mostrar como nossas escolhas se ajustam aos dados. Além disso, destaca-se que cada passo é de extrema relevância para a determinação do resultado final.

5.1 MATRIZ DE SIMILARIDADE

Para calcular a matriz de similaridade, nosso problema apresenta o agravante de levar em consideração a dimensão temporal. Caso a dimensão de tempo não fosse incorporada, ou haveria a aplicação do algoritmo de maneira errada, ou os passos seguintes seriam extremamente prejudicados. Não tratar os dados como séries de tempo, nos levaria a uma dimensão extra para a componente temporal. O algoritmo de clusters tomaria histogramas para cada unidade de tempo com os respectivos valores criminais, o que seria um erro.

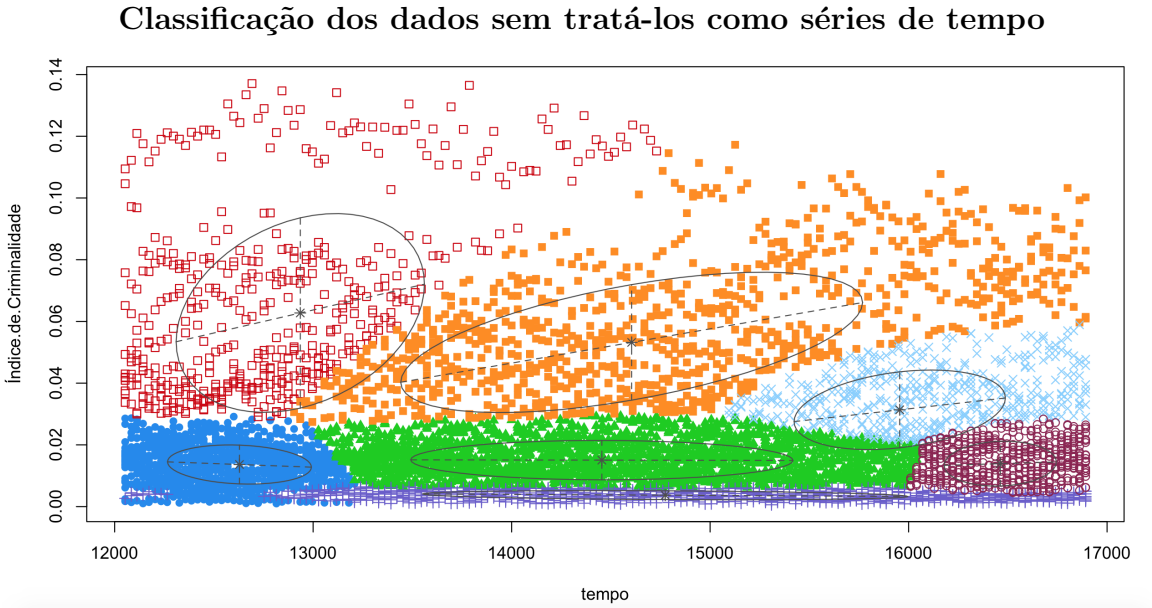

Figura 4 – Classificação 1

A figura 4 ilustra esse caso, em que se clusteriza os dados sem tratá-los como séries de tempo. Neste exercício, utilizamos o valor do índice de criminalidade de cada batalhão em todos os meses e anos do período investigado e buscamos fazer a clusterização desses pontos. É visível que há um maior número de pontos justamente por deixarem de ser uma série de tempo e passarem a ser uma 3-upla (tempo, indicador de criminalidade, batalhão). Os problemas nesse caso são que nossa função responsável por medir a similaridade não é capaz de levar em consideração a tendência de cada batalhão, logo, caso haja alguma mudança de nível, um mesmo batalhão pode estar em diferentes clusters no tempo. O

segundo problema é que a semelhança de nível em determinado mês pode ser um caso extraordinário, o que não é tratado também nesse caso.

E, caso reduzissemos a dimensão temporal, agregando os dados por batalhão, por exemplo, perderíamos a similaridade de tendência, o que prejudicaria nosso poder analítico.

Classificação dos dados sem a dimensão temporal

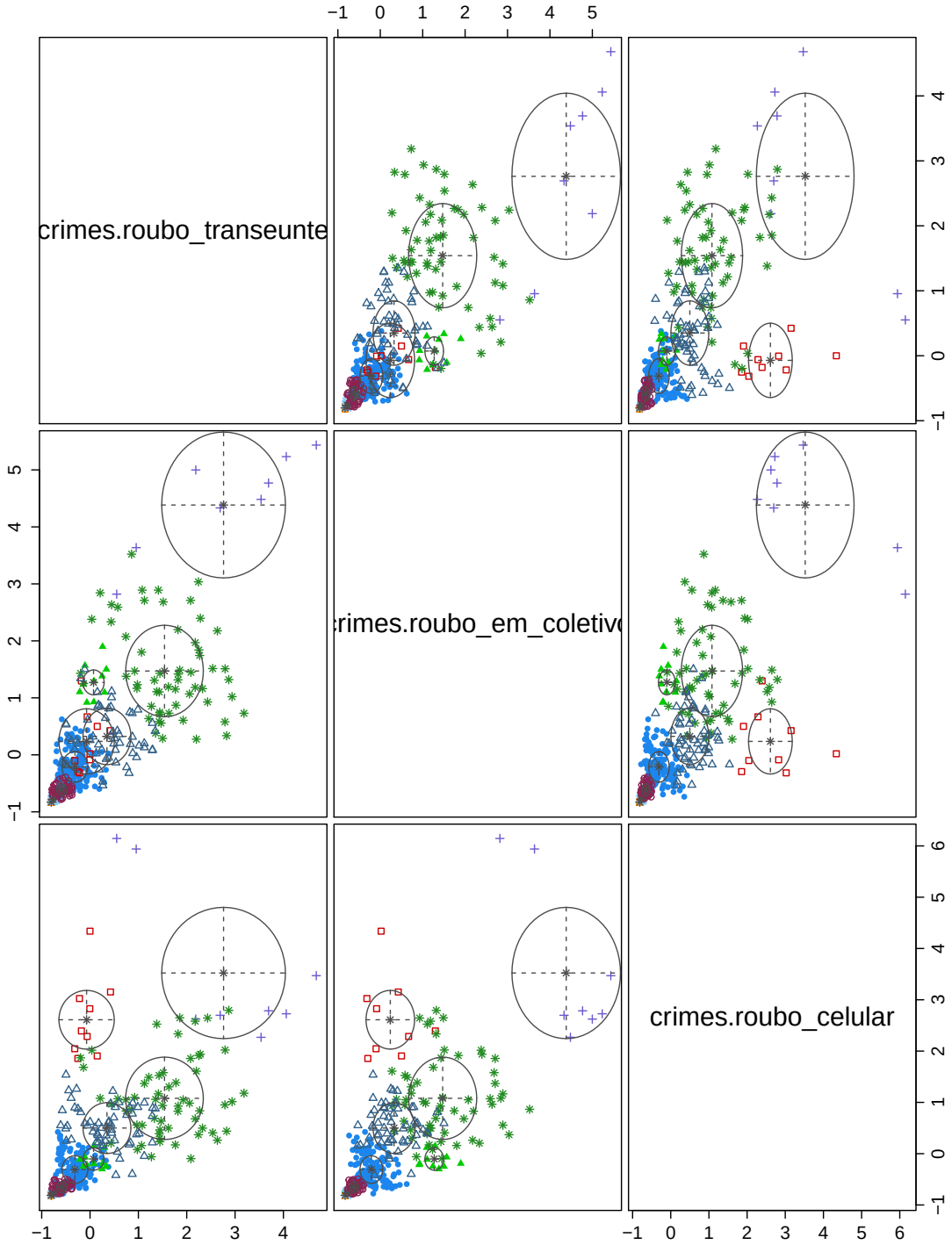

Figura 5 – Classificação 2

Ilustra-se acima na figura 5 o caso em que realizamos a redução de dimensionalidade de tempo através da soma dos crimes escolhidos por batalhão. Vale ressaltar que optamos por expor apenas os crimes: roubo à transeunte, roubo em coletivo e roubo de celular. Nesse caso, perdemos a dinâmica temporal das séries e realizamos clusterizações apenas do nível total de crimes ocorridos em cada batalhão. Os problemas aqui estão na não utilização das tendências criminais e em desconsiderar a relevância de um indicador de criminalidade. Isso, pois a construção de um indicador garante o uso da unidade de medida adequada à análise[1].

Além disso, os passos de avaliação e escolha do número ótimo de clusters não seriam mais válidos em ambos os casos expostos anteriormente. Destaca-se que a medida de similaridade introduzida no capítulo 2.1 leva em conta três dimensões de análise simultaneamente, são elas: (i) nível do indicador de crime; (ii) tendência do índice no tempo; (iii) tendência e nível juntos. O peso de cada uma dessas dimensões na construção da matriz de similaridade depende do valor k de entrada.

Com isso, tomamos o cuidado de utilizar a métrica (2.3), de modo a retornar uma matriz de similaridade que fosse fiel aos dados da análise. Os valores de similaridade estão entre 0 e 1, e quanto mais próximo de zero, mais semelhante são os batalhões (tabela 2).

[1] Somar crimes de naturezas distintas como se fossem semelhantes, por exemplo, como medir a equivalência entre um roubo de celular e um homicídio? Para não entrar nessa discussão, podemos reduzir todas as quantidades de crime a uma unidade de medida comum, como médias das proporções dos crimes de cada batalhão no estado.

Tabela 2 – Matriz de similaridade

	aisp2	aisp3	aisp4	aisp5	aisp6	aisp7	aisp8	aisp9	aisp10	aisp11	aisp12	aisp14	aisp15	aisp16	aisp17	aisp18	aisp19	aisp20	aisp21	aisp22	aisp23	aisp24	aisp25	aisp26	aisp27	aisp28	aisp29	aisp30	aisp31	aisp32	aisp33	aisp34	aisp35	aisp36	aisp37	aisp38	aisp39	aisp40	aisp41
aisp2	0.00	0.22	0.06	0.08	0.03	0.55	0.06	0.33	0.06	0.06	0.24	0.45	0.82	0.20	0.03	0.05	0.05	0.80	0.17	0.05	0.02	0.10	0.06	0.09	0.04	0.04	0.09	0.10	0.04	0.06	0.05	0.06	0.03	0.14	0.10	0.08	0.12	0.12	0.48
aisp3	0.22	0.00	0.26	0.18	0.32	0.16	0.25	0.12	0.36	0.28	0.09	0.08	0.44	0.13	0.35	0.17	0.34	0.32	0.12	0.18	0.35	0.15	0.22	0.16	0.27	0.33	0.31	0.34	0.25	0.25	0.25	0.33	0.41	0.45	0.49	0.37	0.14	0.19	0.16
aisp4	0.06	0.26	0.00	0.08	0.08	0.52	0.05	0.16	0.15	0.14	0.17	0.33	0.60	0.19	0.12	0.06	0.07	0.38	0.10	0.06	0.12	0.05	0.05	0.21	0.06	0.15	0.14	0.16	0.04	0.05	0.16	0.12	0.07	0.13	0.15	0.13	0.07	0.05	0.43
aisp5	0.08	0.18	0.08	0.00	0.15	0.46	0.14	0.18	0.25	0.19	0.10	0.24	0.59	0.08	0.26	0.19	0.23	0.63	0.09	0.12	0.03	0.12	0.06	0.31	0.14	0.19	0.32	0.33	0.14	0.09	0.16	0.20	0.18	0.28	0.17	0.16	0.10	0.08	0.36
aisp6	0.03	0.32	0.08	0.15	0.00	0.56	0.06	0.41	0.08	0.08	0.23	0.23	0.36	0.14	0.07	0.05	0.06	0.66	0.18	0.05	0.03	0.09	0.10	0.12	0.06	0.06	0.09	0.11	0.05	0.09	0.06	0.03	0.03	0.07	0.04	0.08	0.14	0.12	0.67
aisp7	0.55	0.16	0.52	0.46	0.56	0.00	0.47	0.26	0.48	0.37	0.32	0.25	0.12	0.43	0.41	0.53	0.39	0.15	0.34	0.39	0.71	0.47	0.49	0.42	0.39	0.29	0.51	0.31	0.50	0.50	0.37	0.67	0.50	0.53	0.74	0.72	0.38	0.50	0.16
aisp8	0.06	0.25	0.05	0.14	0.06	0.47	0.00	0.36	0.12	0.15	0.10	0.23	0.59	0.18	0.08	0.04	0.10	0.27	0.17	0.03	0.10	0.06	0.07	0.07	0.04	0.09	0.18	0.14	0.05	0.06	0.10	0.09	0.05	0.19	0.11	0.15	0.11	0.06	0.45
aisp9	0.33	0.12	0.16	0.18	0.41	0.26	0.36	0.00	0.51	0.49	0.19	0.12	0.44	0.18	0.39	0.23	0.46	0.26	0.14	0.42	0.36	0.22	0.29	0.55	0.43	0.34	0.31	0.37	0.28	0.27	0.51	0.38	0.35	0.58	0.34	0.51	0.25	0.23	0.08
aisp10	0.06	0.36	0.15	0.25	0.08	0.48	0.12	0.51	0.00	0.01	0.27	0.44	0.64	0.23	0.02	0.11	0.06	0.67	0.34	0.12	0.08	0.12	0.12	0.02	0.07	0.03	0.02	0.02	0.16	0.10	0.04	0.06	0.07	0.03	0.02	0.03	0.20	0.22	0.64
aisp11	0.06	0.28	0.14	0.19	0.08	0.37	0.15	0.49	0.01	0.00	0.34	0.45	0.50	0.31	0.03	0.13	0.06	0.76	0.24	0.11	0.07	0.17	0.16	0.02	0.09	0.02	0.02	0.02	0.17	0.11	0.04	0.06	0.11	0.02	0.02	0.02	0.28	0.26	0.65
aisp12	0.24	0.09	0.17	0.10	0.23	0.32	0.10	0.19	0.27	0.34	0.00	0.14	0.34	0.06	0.20	0.26	0.24	0.24	0.07	0.15	0.22	0.21	0.13	0.32	0.21	0.16	0.42	0.40	0.18	0.18	0.24	0.33	0.23	0.35	0.30	0.24	0.14	0.13	0.31
aisp14	0.45	0.08	0.33	0.24	0.23	0.25	0.23	0.12	0.44	0.45	0.14	0.00	0.24	0.11	0.37	0.22	0.32	0.29	0.18	0.25	0.36	0.27	0.22	0.31	0.27	0.49	0.51	0.44	0.26	0.23	0.46	0.22	0.32	0.41	0.21	0.48	0.26	0.22	0.25
aisp15	0.82	0.44	0.60	0.59	0.36	0.12	0.59	0.44	0.64	0.50	0.34	0.24	0.00	0.42	0.66	0.68	0.70	0.15	0.29	0.59	0.75	0.59	0.69	0.78	0.60	0.68	0.71	0.59	0.57	0.76	0.93	0.47	0.59	0.41	0.56	0.89	0.62	0.60	0.28
aisp16	0.20	0.13	0.19	0.08	0.14	0.43	0.18	0.18	0.23	0.31	0.06	0.11	0.42	0.00	0.24	0.12	0.30	0.52	0.07	0.16	0.19	0.08	0.13	0.33	0.12	0.23	0.27	0.31	0.17	0.08	0.21	0.14	0.23	0.27	0.16	0.31	0.09	0.12	0.35
aisp17	0.03	0.35	0.12	0.26	0.07	0.41	0.08	0.39	0.02	0.03	0.20	0.37	0.66	0.24	0.00	0.07	0.04	0.69	0.23	0.10	0.07	0.14	0.10	0.04	0.05	0.01	0.06	0.04	0.08	0.09	0.03	0.04	0.03	0.07	0.04	0.04	0.21	0.17	0.60
aisp18	0.05	0.17	0.06	0.19	0.05	0.53	0.04	0.23	0.11	0.13	0.26	0.22	0.68	0.12	0.07	0.00	0.05	0.58	0.21	0.04	0.09	0.05	0.06	0.06	0.03	0.09	0.09	0.12	0.05	0.03	0.07	0.07	0.04	0.11	0.14	0.16	0.09	0.04	0.42
aisp19	0.05	0.34	0.07	0.23	0.06	0.39	0.10	0.46	0.06	0.06	0.24	0.32	0.70	0.30	0.04	0.05	0.00	0.49	0.25	0.11	0.06	0.14	0.09	0.08	0.07	0.04	0.08	0.07	0.07	0.07	0.04	0.05	0.03	0.07	0.08	0.06	0.15	0.13	0.54
aisp20	0.80	0.32	0.38	0.63	0.66	0.15	0.27	0.26	0.67	0.76	0.24	0.29	0.15	0.52	0.69	0.58	0.49	0.00	0.33	0.52	0.84	0.52	0.62	0.59	0.71	0.65	0.80	0.88	0.64	0.69	0.85	0.73	0.58	0.53	0.68	0.74	0.57	0.57	0.19
aisp21	0.17	0.12	0.10	0.09	0.18	0.34	0.17	0.14	0.34	0.24	0.07	0.18	0.29	0.07	0.23	0.21	0.25	0.33	0.00	0.24	0.27	0.08	0.16	0.33	0.21	0.23	0.22	0.33	0.13	0.18	0.33	0.24	0.13	0.25	0.28	0.35	0.13	0.15	0.32
aisp22	0.05	0.18	0.06	0.12	0.05	0.39	0.03	0.42	0.12	0.11	0.15	0.25	0.59	0.16	0.10	0.04	0.11	0.52	0.24	0.00	0.09	0.10	0.05	0.10	0.03	0.10	0.18	0.17	0.04	0.04	0.07	0.08	0.07	0.14	0.14	0.14	0.12	0.04	0.58
aisp23	0.02	0.35	0.12	0.03	0.03	0.71	0.10	0.36	0.08	0.07	0.22	0.36	0.75	0.19	0.07	0.09	0.06	0.84	0.27	0.09	0.00	0.14	0.07	0.10	0.09	0.04	0.11	0.11	0.09	0.09	0.03	0.04	0.05	0.10	0.05	0.05	0.18	0.14	0.52
aisp24	0.10	0.15	0.05	0.12	0.09	0.47	0.06	0.22	0.12	0.17	0.21	0.27	0.59	0.08	0.14	0.05	0.14	0.52	0.08	0.10	0.14	0.00	0.09	0.12	0.08	0.13	0.08	0.14	0.08	0.06	0.15	0.09	0.07	0.22	0.13	0.22	0.10	0.07	0.37
aisp25	0.06	0.22	0.05	0.06	0.10	0.49	0.07	0.29	0.12	0.16	0.13	0.22	0.69	0.13	0.10	0.06	0.09	0.62	0.16	0.05	0.07	0.09	0.00	0.17	0.05	0.11	0.25	0.22	0.05	0.02	0.12	0.13	0.10	0.17	0.20	0.13	0.08	0.07	0.46
aisp26	0.09	0.16	0.21	0.31	0.12	0.42	0.07	0.55	0.02	0.02	0.32	0.31	0.78	0.33	0.04	0.06	0.08	0.59	0.33	0.10	0.10	0.12	0.17	0.00	0.12	0.04	0.01	0.01	0.18	0.16	0.02	0.07	0.10	0.02	0.02	0.02	0.25	0.12	0.54
aisp27	0.04	0.27	0.06	0.14	0.06	0.39	0.04	0.43	0.07	0.09	0.21	0.27	0.60	0.12	0.05	0.03	0.07	0.71	0.21	0.03	0.09	0.08	0.05	0.12	0.00	0.08	0.13	0.09	0.03	0.04	0.06	0.04	0.04	0.09	0.13	0.14	0.08	0.10	0.50
aisp28	0.04	0.33	0.15	0.19	0.06	0.29	0.09	0.34	0.03	0.02	0.16	0.49	0.68	0.23	0.01	0.09	0.04	0.65	0.23	0.10	0.04	0.13	0.11	0.04	0.08	0.00	0.04	0.05	0.11	0.09	0.02	0.04	0.04	0.09	0.04	0.07	0.20	0.17	0.54
aisp29	0.09	0.31	0.14	0.32	0.09	0.51	0.18	0.31	0.02	0.02	0.42	0.51	0.71	0.27	0.06	0.09	0.08	0.80	0.22	0.18	0.11	0.08	0.25	0.01	0.13	0.04	0.00	0.01	0.08	0.15	0.05	0.06	0.09	0.01	0.02	0.01	0.15	0.21	0.48
aisp30	0.10	0.34	0.16	0.33	0.11	0.31	0.14	0.37	0.02	0.02	0.40	0.44	0.59	0.31	0.04	0.12	0.07	0.88	0.33	0.17	0.11	0.14	0.22	0.01	0.09	0.05	0.01	0.00	0.15	0.17	0.05	0.05	0.08	0.01	0.03	0.01	0.13	0.21	0.38
aisp31	0.04	0.25	0.04	0.14	0.05	0.50	0.05	0.28	0.16	0.17	0.18	0.26	0.57	0.17	0.08	0.05	0.07	0.64	0.13	0.04	0.09	0.08	0.05	0.18	0.03	0.11	0.08	0.15	0.00	0.06	0.10	0.06	0.06	0.15	0.10	0.12	0.11	0.07	0.48
aisp32	0.06	0.25	0.05	0.09	0.09	0.50	0.06	0.27	0.10	0.11	0.18	0.23	0.76	0.08	0.09	0.03	0.07	0.69	0.18	0.04	0.09	0.06	0.02	0.16	0.04	0.09	0.15	0.17	0.06	0.00	0.08	0.11	0.05	0.17	0.15	0.12	0.08	0.06	0.45
aisp33	0.05	0.25	0.16	0.16	0.06	0.37	0.10	0.51	0.04	0.04	0.24	0.46	0.93	0.21	0.03	0.07	0.04	0.85	0.33	0.07	0.03	0.15	0.12	0.02	0.06	0.02	0.05	0.05	0.10	0.08	0.00	0.03	0.06	0.06	0.03	0.04	0.18	0.13	0.50
aisp34	0.06	0.33	0.12	0.20	0.03	0.67	0.09	0.38	0.06	0.06	0.33	0.22	0.47	0.14	0.04	0.07	0.05	0.73	0.24	0.08	0.04	0.09	0.13	0.07	0.04	0.04	0.06	0.05	0.06	0.11	0.03	0.00	0.03	0.05	0.03	0.09	0.13	0.13	0.52
aisp35	0.03	0.41	0.07	0.18	0.03	0.50	0.05	0.35	0.07	0.11	0.23	0.32	0.59	0.23	0.03	0.04	0.03	0.58	0.13	0.07	0.05	0.07	0.10	0.10	0.04	0.04	0.09	0.08	0.06	0.05	0.06	0.03	0.00	0.09	0.09	0.07	0.12	0.12	0.51
aisp36	0.14	0.45	0.13	0.28	0.07	0.53	0.19	0.58	0.03	0.02	0.35	0.41	0.41	0.27	0.07	0.11	0.07	0.53	0.25	0.14	0.10	0.22	0.17	0.02	0.09	0.09	0.01	0.01	0.15	0.17	0.06	0.05	0.09	0.00	0.03	0.01	0.23	0.25	0.71
aisp37	0.10	0.49	0.15	0.17	0.04	0.74	0.11	0.34	0.02	0.02	0.30	0.21	0.56	0.16	0.04	0.14	0.08	0.68	0.28	0.14	0.05	0.13	0.20	0.02	0.13	0.04	0.02	0.03	0.10	0.15	0.03	0.03	0.09	0.03	0.00	0.02	0.26	0.18	0.70
aisp38	0.08	0.37	0.13	0.16	0.08	0.72	0.15	0.51	0.03	0.02	0.24	0.48	0.89	0.31	0.04	0.16	0.06	0.74	0.35	0.14	0.05	0.22	0.13	0.02	0.14	0.07	0.01	0.01	0.12	0.12	0.04	0.09	0.07	0.01	0.02	0.00	0.25	0.25	0.47
aisp39	0.12	0.14	0.07	0.10	0.14	0.38	0.11	0.25	0.20	0.28	0.14	0.26	0.62	0.09	0.21	0.09	0.15	0.57	0.13	0.12	0.18	0.10	0.08	0.25	0.08	0.20	0.15	0.13	0.11	0.08	0.18	0.13	0.12	0.23	0.26	0.25	0.00	0.05	0.24
aisp40	0.12	0.19	0.05	0.08	0.12	0.50	0.06	0.23	0.22	0.26	0.13	0.22	0.60	0.12	0.17	0.04	0.13	0.57	0.15	0.04	0.14	0.07	0.07	0.12	0.10	0.17	0.21	0.21	0.07	0.06	0.13	0.13	0.12	0.25	0.18	0.25	0.05	0.00	0.37
aisp41	0.48	0.16	0.43	0.36	0.67	0.16	0.45	0.08	0.64	0.65	0.31	0.25	0.28	0.35	0.60	0.42	0.54	0.19	0.32	0.58	0.52	0.37	0.46	0.54	0.50	0.54	0.48	0.38	0.48	0.45	0.50	0.52	0.51	0.71	0.70	0.47	0.24	0.37	0.00

A assimetria apresentada entre os batalhões 20 e 23, como discutido e ilustrado anteriormente, já era esperada, pois eles possuem diferentes padrões criminais e socioeconômicos.

5.2 DENDOGRAMA

Ao prosseguir para a clusterização hierárquica, temos o problema de qual método escolher, pois nessa parte do exercício, o método e o critério de avaliação estão diretamente ligados. As escolhas feitas no exercício até aqui, nos levam ao dendograma da figura 6[2]. Apesar da ausência de subdivisões, é nítida a ordem monotônica dos graus de similaridade na medida em que se está a níveis mais baixos de altura da árvore.

Além disso, é importante destacar que o método *complete* prioriza que a distância entre os clusters seja a máxima possível. Caso, por consequência, os clusters obtidos apresentem elementos com distância baixa entre si, menor será o índice de McClain. Isto justifica nossa escolha de método, pois está em total acordo com o problema de otimização estruturado no capítulo 2.1. Problema esse, que busca a evidência de clusterização dos dados.

[2] O dendograma ilustrado já leva em consideração o valor de k ótimo obtido no exercício.

Dendograma para o método Complete.

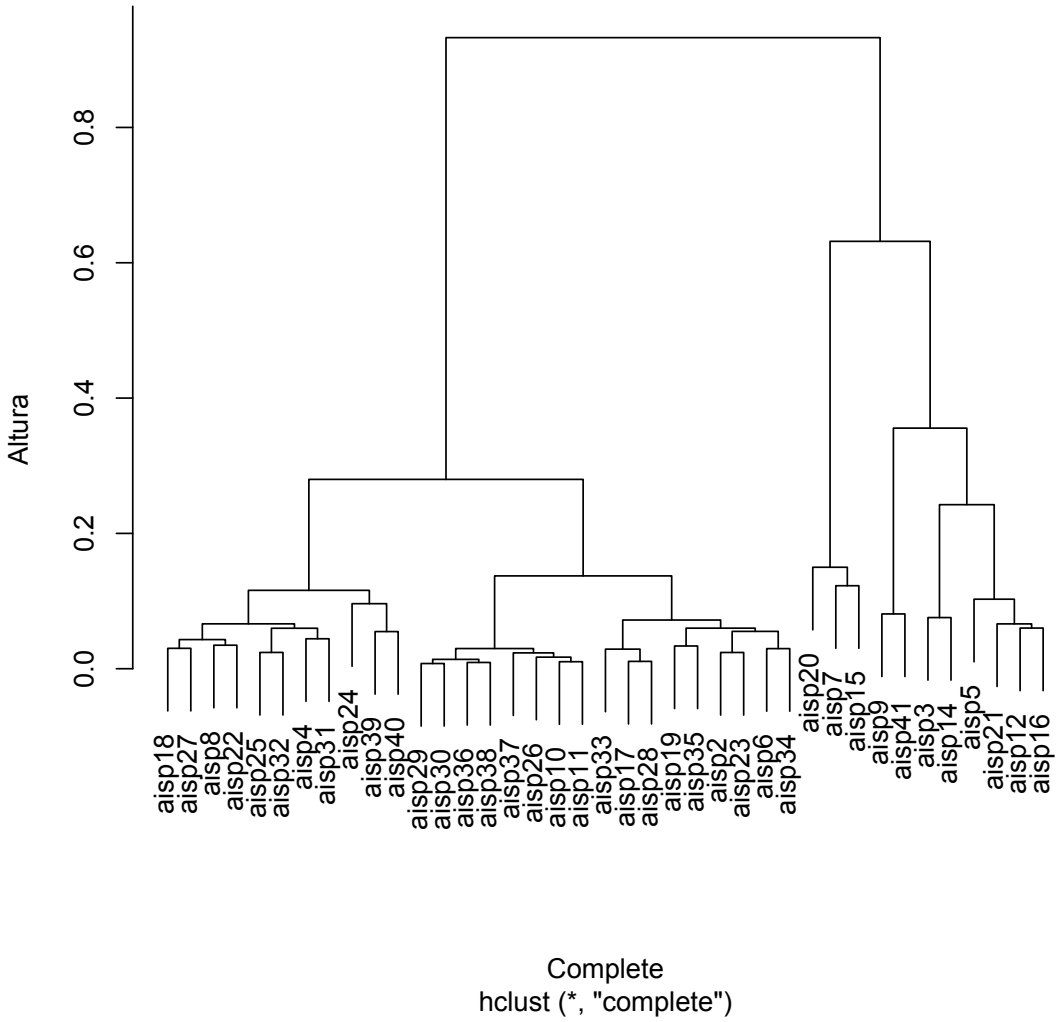

Figura 6 – Método: Complete

A escolha do método *single*[3] nos levaria a um dendograma de menor utilidade analítica, pois embora a semelhança seja preservada, a divisão em clusters se tornaria menos intuitiva. O método *single* garante a distância mínima entre os elementos dos clusters, o que, não necessariamente, resulta na distância máxima entre os clusters, gerando um viés para solução de borda (ou temos o caso de apenas dois clusters ou o caso de *overfiting*).

[3] $d(x_a, x_b) = min(a, b)$.

Dendograma para o método Single.

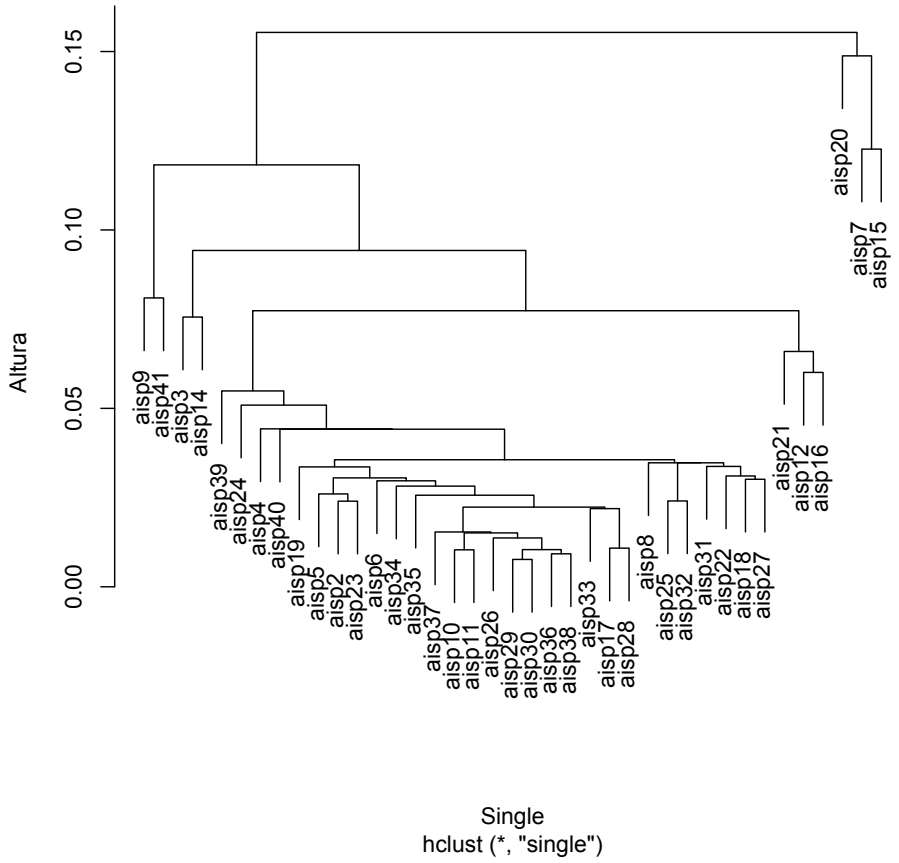

Figura 7 – Método: Single

5.3 ÍNDICE DE MCCLAIN

Essa seção apresenta a matriz de resultados numéricos do algoritmo enunciado na seção metodológica. Como podemos ver na tabela 3 [4], os valores ótimos seriam $k = 3$ e $q = 3$.

Esses resultados mostram evidência de que os dados são passíveis de serem clusterizados, pois encontramos valores de clusters que atendem a restrição apresentada na equação (2.1).

Já a tabela 4, que mostra os resultados numéricos para o método *single*, gera apenas soluções de borda, como adiantado na seção anterior.

[4] Opta-se por não mostrar todos os valores do índice até $k = 100$, pois é fácil ver que o algoritmo já convergiu quando atinge $k = 30$.

Tabela 3 – Resultado Numérico do problema de otimização do Índice de McClain para o método *complete*

	Número de Clusters	Valor do Índice
1	2.00	0.12
2	3.00	0.26
3	3.00	0.20
4	2.00	0.18
5	2.00	0.16
6	2.00	0.21
7	2.00	0.17
8	2.00	0.17
9	2.00	0.17
10	2.00	0.16
11	38.00	0.12
12	38.00	0.05
13	38.00	0.02
14	38.00	0.01
15	38.00	0.00
16	38.00	0.00
17	38.00	0.00
18	38.00	0.00
19	38.00	0.00
20	38.00	0.00
21	38.00	0.00
22	38.00	0.00
23	38.00	0.00
24	38.00	0.00
25	38.00	0.00
26	38.00	0.00
27	38.00	0.00
28	38.00	0.00
29	38.00	0.00
30	38.00	0.00

Tabela 4 – Resultado Numérico do problema de otimização do Índice de McClain para o método *single*

	Número de Clusters	Valor do Índice
1	2.00	0.0193
2	2.00	0.0196
3	2.00	0.0517
4	2.00	0.0191
5	2.00	0.0189
6	2.00	0.0187
7	2.00	0.0409
8	2.00	0.0409
9	2.00	0.0409
10	2.00	0.0188
11	2.00	0.0188
12	2.00	0.0188
13	2.00	0.0188
14	38.00	0.01
15	38.00	0.00
16	38.00	0.00
17	38.00	0.00
18	38.00	0.00
19	38.00	0.00
20	38.00	0.00
21	38.00	0.00
22	38.00	0.00
23	38.00	0.00
24	38.00	0.00
25	38.00	0.00
26	38.00	0.00
27	38.00	0.00
28	38.00	0.00
29	38.00	0.00
30	38.00	0.00

5.4 CLUSTERS

O dendograma da figura 8 destaca os três clusters obtidos a partir do exercício de otimização apresentado na seção metodológica.

Dendograma com 3 clusters extraídos do exercício de otimização.

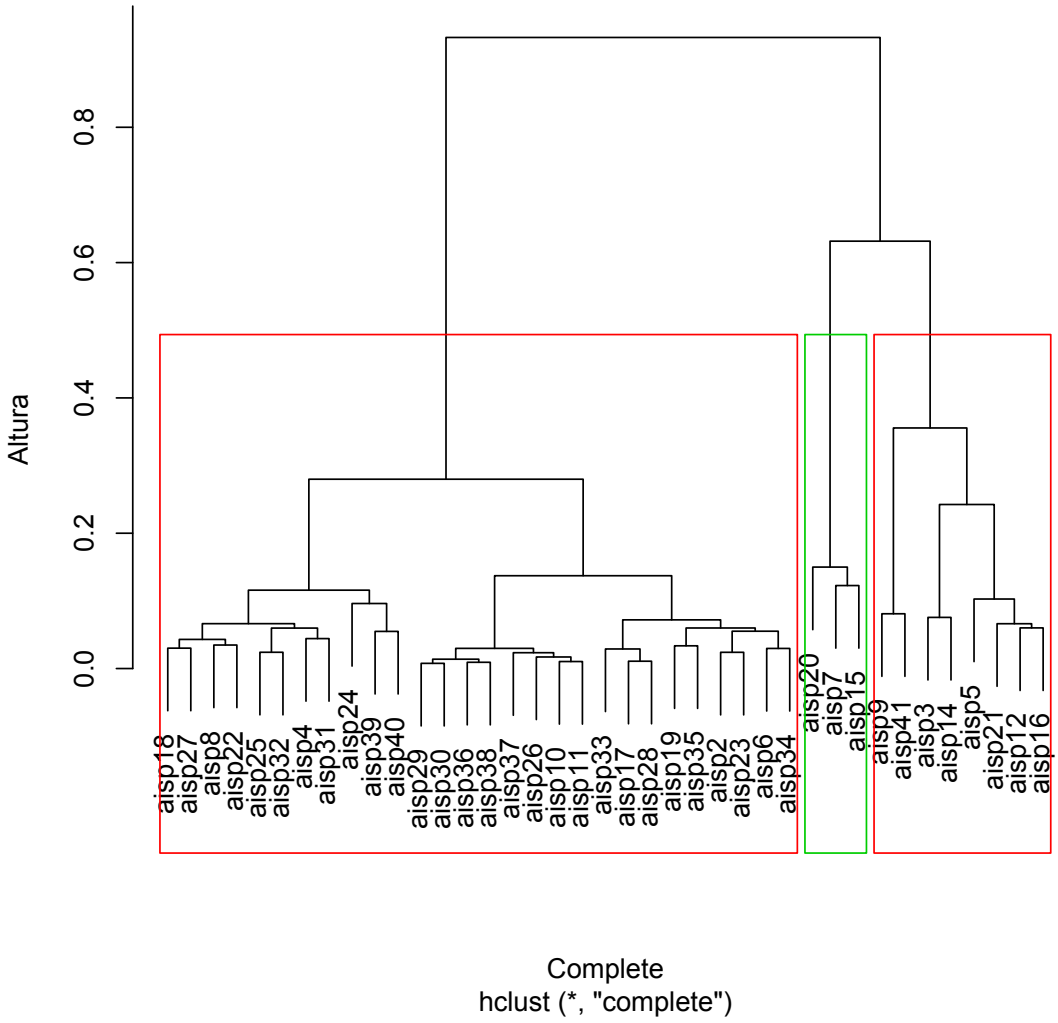

Figura 8 – Método: Complete

No entanto, ao analisarmos o perfil dos batalhões, verificamos intuitivamente a possibilidade de aumentar as subdivisões dos clusters. Ou seja, utiliza-se o mesmo dendograma, porém, adota-se um corte mais abaixo na altura da árvore. Deve-se deixar claro, contudo, que com isso abre-se mão da solução ótima.

Dentre essas informações, podemos destacar que os batalhões 9 e 41 já foram o mesmo no passado e possuem um nível de semelhança extremamente relevante. É possível identificar um grupo com batalhões alocados na zona sul da capital, região conhecida por

seu alto nível socioeconômico e potencial turístico, junto a batalhões do interior do sul do estado, região menos populosa, e consequentemente ambas com menores índices de criminalidade relativos a outros batalhões do estado. Além disso, podemos observar a subdivisão de batalhões alocados mais a norte do estado junto de outros encontrados na zona norte da capital. Por fim, é possível notar um cluster com batalhões que pertencem à Baixada Fluminense e a São Gonçalo, regiões com altos índices de criminalidade e características sociais semelhantes, como baixos indicadores socioeconômicos. Destaca-se que essa análise não altera em nada o que foi feito anteriormente. Logo, após escolher intuitivamente $q = 6$ clusters, deve-se buscar uma forma de classificar esses clusters e assim extrair aprendizagem dessa subdivisão. Os novos clusters se encontram na figura 9.

Dendograma com 3 clusters extraídos do exercício de otimização.

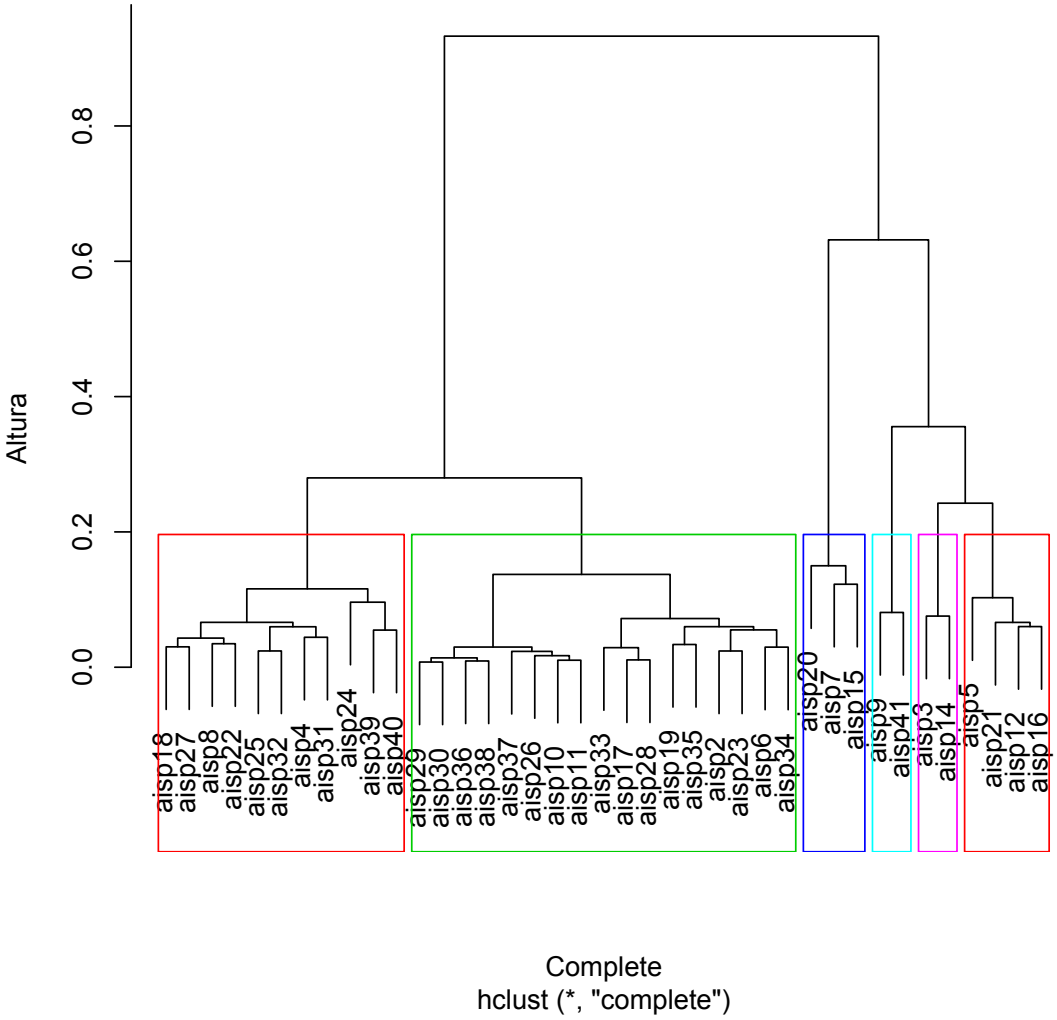

Figura 9 – Método: Complete

Após o cálculo da matriz de similaridade e da discussão acerca do modelo de cluste-

rização hierárquica, uma série de algortimos de criação de clusters podem ser utilizados nos dados. Uma discussão acerca do algoritmo *Expectation Maximization* se encontra no apêndice A.

A ilustração em mapas da figura 10 facilita a observação do que foi argumentado neste capítulo. O contraste observado no mapa indica que é possível contribuir, através de um estudo dos padrões criminais, para a utilização das RISP.

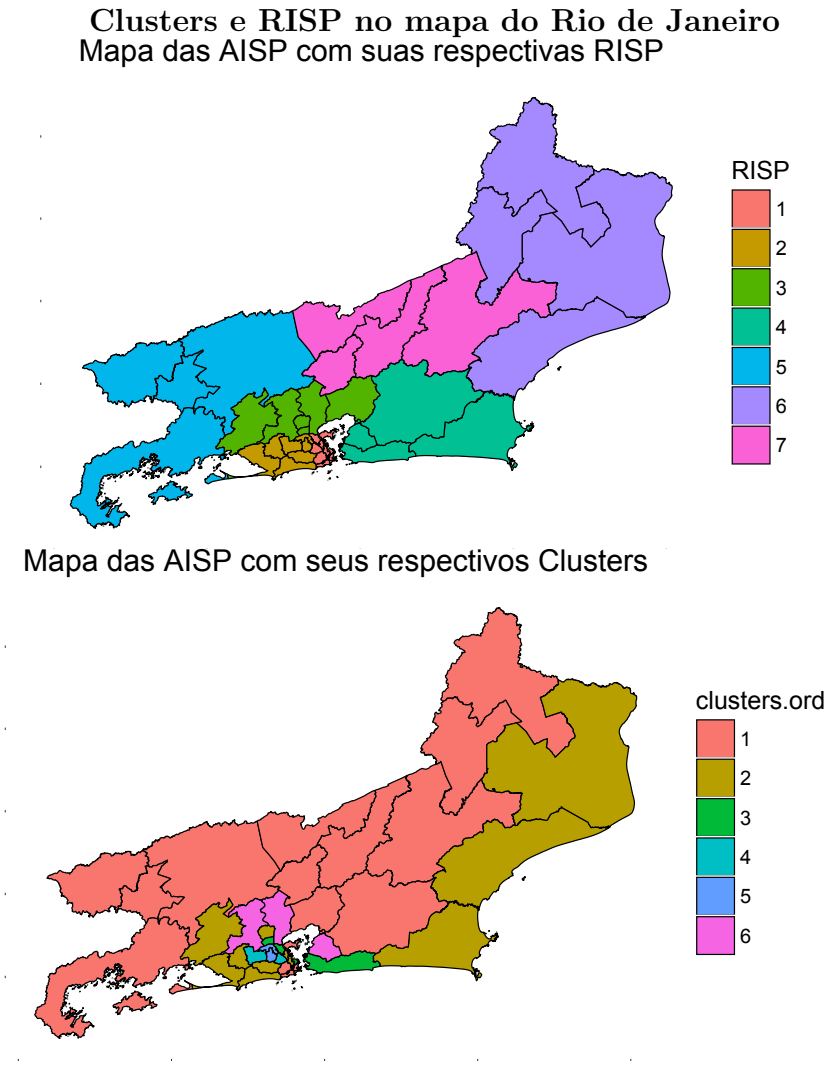

Figura 10 – Mapas das RISP e dos Clusters Ordenados

5.5 OLOGIT

Por fim, para extrair características dos grupos, o primeiro passo é verificar se os clusters possuem ordem clara. A figura 11[5] sugere que os clusters podem ser ordenados tanto em termos de mediana quanto em média. A figura seguinte (12) ilustra essa possibilidade. A

[5] Além da visualização, estatísticas descritivas são úteis para identificar a ordem em termos de mediana e média para cada cluster. Essas tabelas são apresentadas uma a uma no apêndice B.

análise que se segue considera sempre essa versão ordenada (figura 12), ou seja, clusters em posições elevadas apresentam maior incidência criminal e vice-versa.

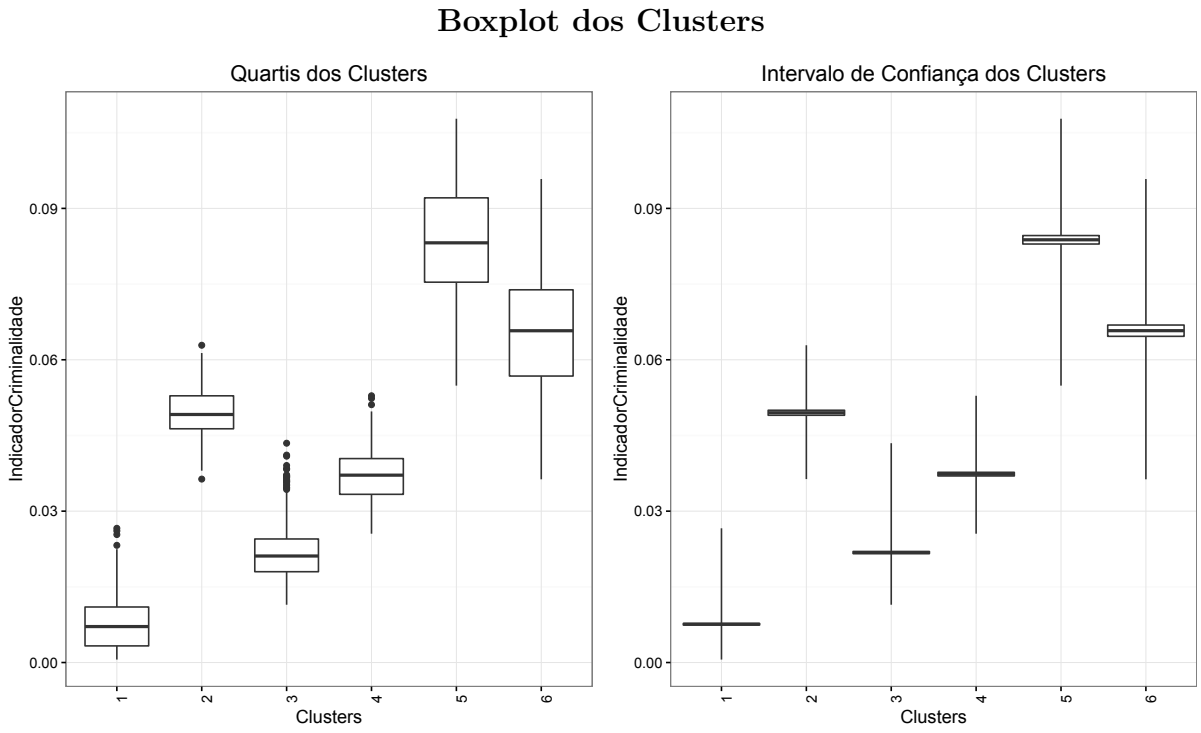

Figura 11 – Índice de criminalidade x Clusters

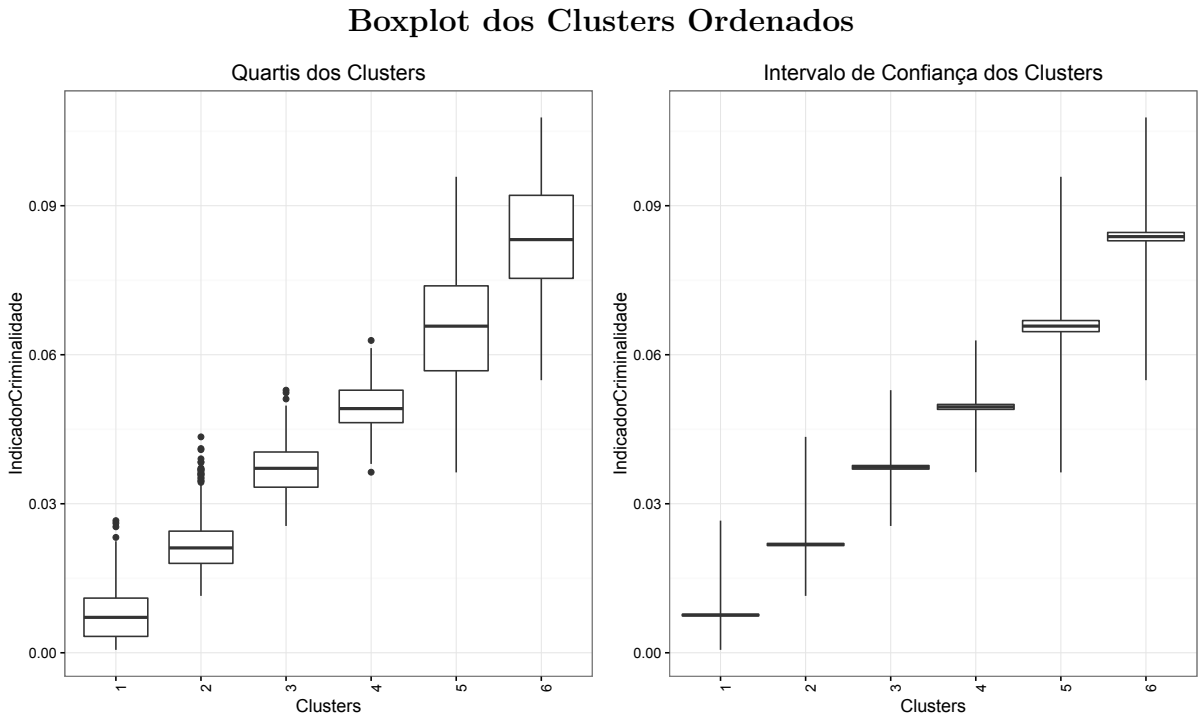

Figura 12 – Índice de criminalidade x Clusters

Uma possível variável explicativa para entender a ordem dos clusters é o Índice de Desenvolvimento Humano (IDH). Essa variável foi escolhida por capturar características de grau socioeconômico, as quais intuímos que estejam diretamente ligadas aos padrões criminais do estado. Utilizar dados de IDH em análises de padrão criminal não é uma abordagem nova. Em Beato [1998] busca-se a relação entre IDH e os padrões criminais para o estado de Minas Gerais, expondo correlações entre diferentes regiões geográficas, consequentemente diferentes níveis de IDH, do estado. Em sociologia existem linhas de pensamento, como a abordada em Cohen et al. [1979], na qual se defende que crime é oportunidade. Nesse projeto, não encontramos evidência oportunista para o crime, e sim classificações que possam dar luz para análises futuras em tal segmento.

Como exposto no capítulo 3, fez-se necessária a construção do índice por batalhão. É possível ver nas figuras 13 e 14 [6] um padrão ordenado na correlação entre os dados de IDH médios e os valores do nosso indicador de criminalidade. Além disso, devemos destacar o perfil triangular da dispersão e a reafirmação da ordem dos clusters [7].

Dispersão entre o indicador de criminalidade e IDH em julho de 2011

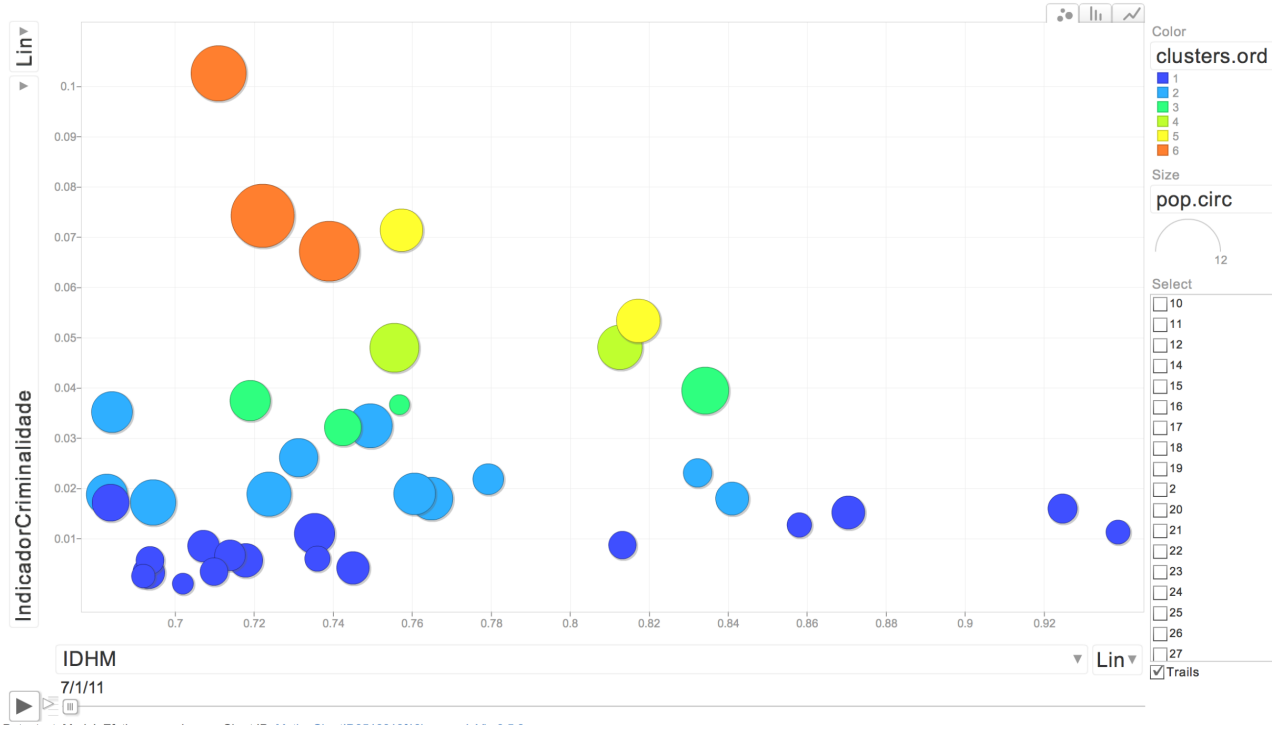

Figura 13 – Índice de criminalidade x IDH 2011

[6] Não apresentamos a dispersão para 2016, embora ela exista, pois os dados de efetivo policial disponíveis são até março de 2015. Como os modelos deste trabalho utilizam os dados de efetivo, opta-se por mostrar a dispersão do ultimo mês disponível do dado em questão.

[7] Na visualização, os clusters são representados pelas cores dos pontos.

Dispersão entre o indicador de criminalidade e IDH em março de 2015

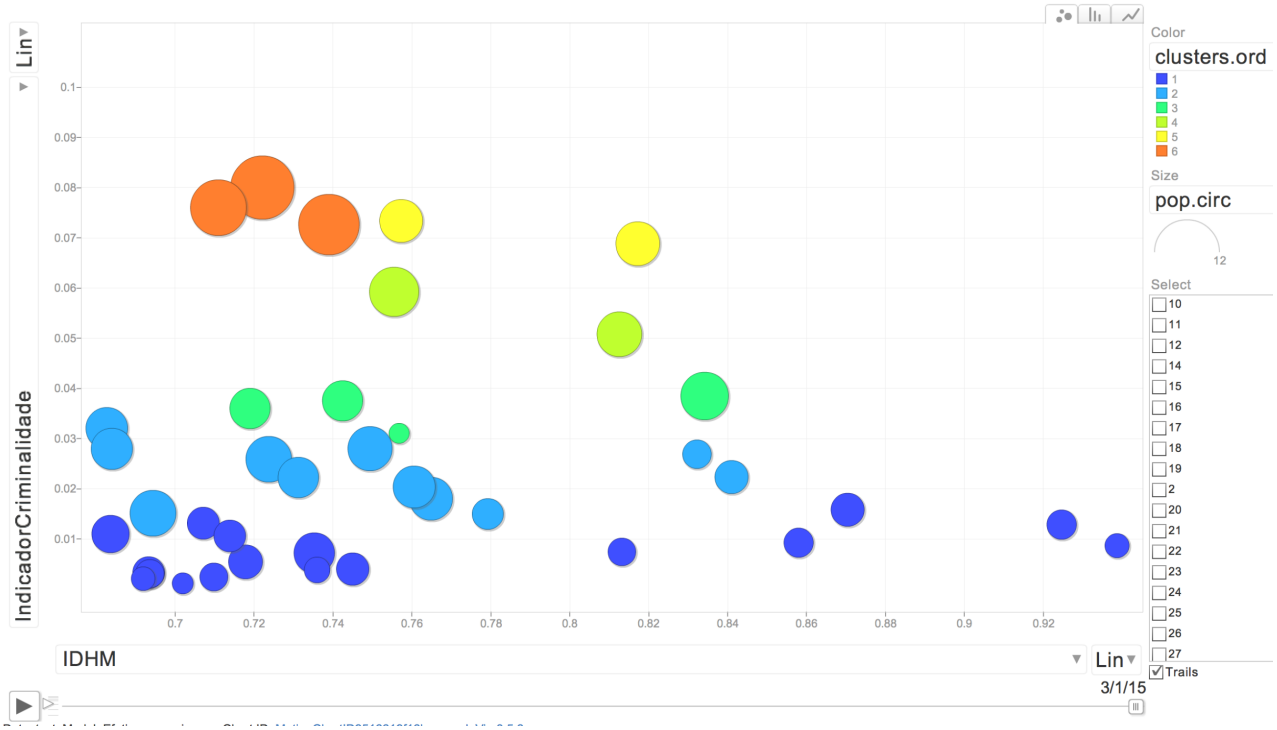

Figura 14 – SIM x IDH 2015

Utilizamos o modelo OLogit a fim de extrair o sinal do IDH na classificação dos clusters. Mais ainda, buscamos entender o efeito de uma unidade a mais de IDH na mudança de cluster e mostrar significância nos parâmetros. Essa análise, ilustrada na tabela 5, mostra a relevância do caráter socioeconômico na ordem das subdivisões encontradas.

Tabela 5 – Aplicação do modelo OLogit no ordenamento dos clusters.

	Modelo 1
IDH médio de educação	46.31^{***}
	(2.25)
IDH médio de renda	-46.09^{***}
	(2.35)
IDH médio de longevidade	-2.79
	(5.02)
AIC	4397.78
BIC	4441.54
Log Likelihood	-2190.89
Deviance	4381.78
Num. obs.	1755

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Percebe-se, que batalhões com níveis de IDH igualmente baixos possuem diferentes indicadores SIM de criminalidade. O que, somado à análise anterior, sugere que pobreza

é uma condição necessária, mas não suficiente para incidência criminal. Uma primeira hipótese seria a de que em determinados batalhões podem existir áreas reconhecidas como bolsões de pobreza. Uma solução para esse caso, seria rodar uma versão do modelo com o valor mínimo de IDH de renda como variável explicativa da ordenação dos clusters. Essa abordagem seria capaz de extrair o efeito comparativo das regiões mais pobres presentes nos batalhões, o que reduz o problema de subestimar o efeito da pobreza regional no aumento da incidência de criminalidade. Os resultados deste modelo podem ser encontrados na tabela 6.

Tabela 6 – Aplicação do modelo OLogit no ordenamento dos clusters

	Modelo 1	Modelo 2
IDH médio de educação	46.31^{***}	22.40^{***}
	(2.25)	(1.63)
IDH médio de renda	-46.09^{***}	
	(2.35)	
IDH médio de longevidade	-2.79	-41.53^{***}
	(5.02)	(3.59)
IDH mínimo de renda		-7.27^{***}
		(0.93)
AIC	4397.78	4865.78
BIC	4441.54	4909.54
Log Likelihood	-2190.89	-2424.89
Deviance	4381.78	4849.78
Num. obs.	1755	1755

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Para uma nova tentativa (tabela 7) acrescentam-se dados de efetivo policial por 100mil habitantes, para que assim, além de efeitos socioeconômicos, possamos identificar efeito de força policial na ordenação de cluster. De acordo com a teoria [8], quando feitos os controles devidos, a maior concentração policial por habitante, aumenta a capacidade de controle criminal, o que em nosso modelo faria com que batalhões com maior quantidade de efetivo total por habitante tivessem maior probabilidade de estarem em clusters com padrão criminal mais baixo.

[8] No referencial teórico cita-se Becker [1974], Levitt [1997], DiTella e Shargodsky [2004].

Tabela 7 – Aplicação do modelo OLogit no ordenamento dos clusters

	Modelo 1	Modelo 2	Modelo 3
IDH médio de educação	46.3058^{***}	22.3963^{***}	21.2821^{***}
	(2.2451)	(1.6265)	(0.6785)
IDH médio de renda	-46.0945^{***}		
	(2.3518)		
IDH médio de longevidade	-2.7875	-41.5320^{***}	-37.2524^{***}
	(5.0244)	(3.5945)	(0.2336)
IDH mínimo de renda		-7.2681^{***}	-7.5650^{***}
		(0.9330)	(0.9380)
Efetivos por 100mil/hab.			-0.0008^{***}
			(0.0002)
AIC	4397.7817	4865.7777	4856.5141
BIC	4441.5435	4909.5395	4905.7462
Log Likelihood	-2190.8908	-2424.8889	-2419.2571
Deviance	4381.7817	4849.7777	4838.5141
Num. obs.	1755	1755	1755

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

No último modelo foi possível extrair que quanto maior o IDH mínimo de um batalhão maior a probabilidade de estar em um cluster com menor incidência criminal. O mesmo é válido para o IDH médio de longevidade e para os valores de efetivos [9] por 100mil habitantes. Ao analisarmos os valores AIC, temos que o terceiro modelo tem menor probabilidade de perda de informação relativo à segunda versão, o que mostra melhor *fit* do mesmo.

Após a análise dos três modelos, percebe-se que a população é de extrema relevância para explicar o ordenamento de padrões criminais entre os clusters. Em nossa abordagem do modelo 3, já levamos em consideração a taxa de efetivo policial por 100 mil habitantes da população como variável característica. Alternativamente, vamos utilizar população como controle para um novo conjunto de modelos, sem utilizar IDH médio de educação e de longevidade por serem altamente correlacionados com dados de população, o que poderia nos levar a um problema de multicolinearidade. Os resultados dessa estimativa estão expostos na tabela 7.

[9] Possíveis controles podem aumentar o valor desse parâmetro. Na seção de passos futuros, discutimos acerca de contribuições que podem ser feitas ao modelo.

Tabela 8 – Aplicação do modelo OLogit no ordenamento dos clusters versão alternativa

	Modelo 1	Modelo 2	Modelo 3
IDH mínimo de renda	-4.1239^{***}	-2.8499^{***}	-1.4303^{***}
	(1.1608)	(0.0956)	(0.1001)
População	1.2097^{***}	1.2648^{***}	1.3860^{***}
	(0.0386)	(0.0402)	(0.0442)
Total de efetivos por batalhão		0.0013^{***}	
		(0.0001)	
Total de efetivos por batalhão sem UPP			-0.0008^{**}
			(0.0003)
Total de efetivos nas UPP			0.0017^{***}
			(0.0001)
AIC	3388.0967	3255.9555	3192.3878
BIC	3426.3883	3299.7173	3241.6198
Log Likelihood	-1687.0484	-1619.9777	-1587.1939
Deviance	3374.0967	3239.9555	3174.3878
Num. obs.	1755	1755	1755

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

A primeira versão leva em consideração um modelo com apenas as variáveis explicativas de IDH mínimo de renda e população como controle. Obtivemos os resultados esperados: (i) quanto maior a população maior a probabilidade de estar em um cluster de maior incidência criminal; (ii) quanto maior o IDH mínimo de renda maior a probabilidade de estar em um cluster de menor incidência criminal.

Na segunda versão introduzimos a variável característica de efetivo total. Aqui temos o problema de que quanto maior o efetivo policial maior a probabilidade de estar em um cluster de maior incidência criminal. Esse resultado não é satisfatório, possivelmente devido à presença de dupla causalidade, pois a teoria[10] diz que maior contingente policial leva a menor incidência de crimes.

Já a terceira versão apresenta os dados em absoluto dos efetivos alocados em batalhões sem contar UPP como variável característica. Aqui, fazemos a distinção, pois como é sabido tanto pelo desenho das UPP como por artigos que abordam o programa[11], a distribuição das UPP não foi feita de forma homogênea e aleatória. As tomadas de território ocorreram de acordo com oportunidades bem aproveitadas pelas forças do estado. Com isso, a distribuição de efetivo de UPP pode apresentar o citado problema de dupla causalidade, desafio comum em artigos que buscam o impacto de policiamento no controle de criminalidade[12], pois em geral há maior contingente policial em áreas de maior criminalidade. Espera-se que os controles e estratégias de estimação sejam eficazes para extrair o efeito de policiamento independente do tamanho do contingente.

[10] Visto em Becker [1974]
[11] Beltrame [2014], Ottoni e Ferraz [2013] e Magaloni et al [2015]
[12] Visto em Levitt [1997] e DiTella e Shargodsky [2004]

Embora a investigação do efeito causal da alocação de efetivo policial em incidência criminal nas áreas de UPP constitua um tema interessante, foge do escopo do presente estudo. Nessa terceira versão as variáveis de efetivo policial alocado nas UPP foram introduzidas como controle. A terceira versão, então, aponta que mais efetivo exceto em áreas de UPP aumenta a probabilidade de estarmos em clusters de incidência criminal mais baixa, como esperado. Além disso, novamente ao analisarmos os valores AIC, temos que o terceiro modelo tem menor probabilidade de perda de informação relativo à segunda versão, o que mostra também melhor *fit* do mesmo.

6 CONCLUSÃO

6.1 CONSIDERAÇÕES FINAIS

O presente trabalho propõe que os batalhões policiais do estado do Rio de Janeiro são passíveis de clusterização, baseada nas semelhanças de nível e de tendência de seus respectivos padrões criminais, bem como há a possibilidade de ordenamento desses clusters e aplicação de um algoritmo de classificação que explique de forma significativa o ordenamento encontrado.

Do ponto de vista prático, os resultados obtidos podem servir de alternativa a já criada RISP (Região Integrada de Segurança Pública), para análise dos padrões criminais, para divisão dos recursos e para o treinamento de efetivo dos batalhões pertencentes aos cluster. Além disso, a métrica utilizada para definir o quão violento é um batalhão[1] e, posteriormente, o possível ordenamento dos clusters obtidos, pode aprimorar o índice do SIM atual.

Uma medida de qualidade do desempenho dos batalhões e de seus comandantes é de extremo interesse ao planejador público. Muito há a se fazer nessa discussão, porém até aqui, pode-se dizer que o alto número de comandantes que não passam por mais de um batalhão, prejudica extrair suas medidas de desempenho. No que diz respeito aos batalhões, esse trabalho introduz a discussão e, através de uma abordagem de matemática aplicada, contribui para a discussão, ao propor e analisar classificações para os batalhões de forma objetiva.

Mostra-se, através de um problema de otimização numérica do índice de McClain, que as séries de tempo dos crimes são passíveis de serem clusterizados e, a partir de então, diferentes implementações são discutidas. Focamos no modelo de clusterização hierárquica sob o critério *complete* por sua capacidade analítica.

Por fim, mostra-se evidência de que o ordenamento dos clusters pode ser explicado por dados de IDH mínimo de renda de cada batalhão, por dados de IDH de longevidade médio, efetivos por 100 mil habitantes e IDH médio de educação.

6.2 TRABALHOS FUTUROS

O problema de otimização numérica foi estruturado para poder facilitar a maximização de diferentes índices de avaliação dos clusters, desde que o novo índice seja adequado a análise de interesse. Um próximo passo seria a criação de um modelo de previsão

[1] Índice de criminalidade apresentado no capítulo 4

para os clusters e, dessa forma, criarmos uma medida de avaliação [2] dos batalhões na linha de expectativa/realizado. Para definirmos o modelo de previsão, é fundamental que o problema de otimização esteja bem definido, para que a seleção dos parâmetros via aprendizagem por máquinas se dê de forma coerente.

Contribuições à classificação deste exercício são pertinentes. Aumentar o *lag* da correlação temporal incorporada na medida de similaridade pode ser interessante. É possível que haja maior aproveitamento das tendências criminais dos batalhões, ao se levar em consideração mais dados do passado. Para isso, vamos implementar o algoritmo para *lag* de 6 e 12 meses e interpretar os resultados obtidos.

Encorporar dados de população flutuante para todo o estado melhoria a estimativa dos parâmetros de nosso modelo OLogit de classificação dos clusters. Hoje possuímos dados apenas para a capital, mas a EMAp está por iniciar uma coleta de uma maior massa de dados no tempo para todo o estado. Outros dados a serem extraídos, que provavelmente contribuirão como controle para a estimativa dos parâmetros do modelo classificatório desta dissertação, serão os dados de mapeamento de atividade das facções criminosas e das milícias atuantes no Rio de Janeiro.

A definição dos clusters facilita a análise via controle sintético, bem como a identificação de fenômenos naturais não tão facilmente identificáveis. No primeiro caso, o centroide do clusters, ou alguma medida funcional que represente aquela subdivisão pode ser utilizada para representar o controle sintético. Já para a segunda, a observação dos padrões criminais de batalhões semelhantes pode facilitar a identificação de um fenômeno exclusivo a um deles, que seja responsável por melhorar ou piorar a performance do mesmo comparado ao seu semelhante. Um possível trabalho futuro será a análise exclusiva do grupo, onde estão presentes os batalhões 9 e 41, com o fim de identificar o fator que faz deles tão fortemente correlacionados em direções opostas.

Aplicar a metodologia discutida nesse projeto para a dinâmica criminal das UPP já é um projeto em curso, mas ainda sem resultados satisfatórios. Para o caso das UPP, pode ser feito um paralelo com o índice de risco criado pelo ISP para acompanhar o desempenho das unidades ao longo do tempo. Uma proposta seria a criação de um indicador de criminalidade semelhante ao da equação (4.1) e daí replicar os passos do exercício expostos nessa dissertação.

Além disso, pode ser útil para aplicação de políticas, como a de treinamento de efetivo, que hoje não possui distinsão por mais que o estado possua variados padrões criminais, como exposto no capítulo 5, o qual apresentamos os resultados.

Por fim, os clusters podem possibilitar o desenho de experimentos futuros, pois facilitam a identificação de possíveis batalhões de tratamento e seus respectivos controles. Dentre esses experimentos, uma proposta seria a de treinamento diferenciado de efetivo.

[2] Algortimos estocásticos utilizam a medida BIC usualmente. No entanto, pode-se buscar outras abordagens.

Isso poderia se dar de inúmeras formas. Aqui podemos destacar duas abordagens: a primeira seria treinar e alocar o efetivo de acordo com os clusters, de modo a levar em consideração os padrões de violência do local onde esse efetivo desempenhará sua função; o segundo formato seria a de comparação dos batalhões dentro do mesmo clusters na medida em que o efetivo diferenciado fosse alocado no tempo.

REFERÊNCIAS

[1] Yaser S. Abul-Mostafa, Malik Magdon-Ismail e Hsuan-Tien Lin *Learning from data*, AMLbook.com 2012,

[2] Jyoti Agarwal, Renuka Nagpal e Rajni Sehgal *Crime Analysis using K-means Clustering 13*, International Journal of Computer Applications 2016, Vol. 83 (Dezembro), 2013.

[3] Joshua Angrist e Jorn-Steffen Pischke *Mastering 'Metrics: The Path from Cause to Effect*, Princeton University Press 2016,

[4] A.Malathi e S.Santhosh Baboo *Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters*, Global Journal of Computer Science and Technology 2011, Vol.11 (Julho), 2011.

[5] R.Bulli Babu, G.Snehal e P.Aditya Satya Kiran *Detection of Crimes using Unsupervised Learning Techniques*, Indian Journal of Science and Technology 2016, Vol. 9 (Maio), 2016.

[6] Claudio C. Beato F. *Determinantes da criminalidade em Minas Gerais.*, Revista Brasileira de Ciências Sociais 1998, Vol. 13 (Junho), 1998.

[7] Gary S. Becker e Willian M.Landes *Crime and Punishment: An Economic Approach*, NBER, 1974.

[8] Malika Charrad, Nadia Ghazzali, Véronique Boiteau, Azam Niknafs *NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set.*, Journal of Statistical Software 2014, Vol. 61 (Outubro), 2014.

[9] Lawrence E. Cohen e Marcus Felson, *Social Change and Crime Rate Trends: A Routine Activity Approach.*, American Sociological Review 1979, Vol. 44 (Agosto): 588-608, 1979.

[10] Lawrence E. Cohen, Marcus Felson e Kenneth C. Land *Property Crime Rates in the United States: A Macrodynamic Analysis, 1947-1977; with Ex Ante Forecasts for the Mid-1980s'.*, American Journal of Sociology 1980, Vol. 86 (Julho): 90-118, 1980.

[11] Rafael Di Tella e Ernesto Schargrodsky *Do Police Reduce Crime? Estimates Using the Allocation of Police Forces After a Terrorist Attack.*, The American Economic Review 2004, Vol. 94 (Maço): 115-133, 2004.

[12] Claudio Ferraz e Bruno Ottoni *State Presence and Urban Violence: Evidence from the Pacification of Rio's Favelas.*, Working Paper, (Julho), 2013.

[13] Chris Fraley, Adrian Raftery *Model-based Methods of Classification: Using the mclust Software in Chemometrics*, Journal of Statistical Software 2007, Vol. 18 (Janeiro), 2007.

[14] Trevor Hastie, Robert Tibshirani e Jerome Friedman *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer 2008,

[15] Kosuke Imai, Gary King e Olivia Lau *Zelig: Everyone's Statistical Software.*, (Junho), 2012.

[16] Steven D. Levitt *Using Electorial Cycles in Polince Hiring to Estimate the Effect of Police on Crime.*, The American Economic Review 1997, Vol. 87 (Junho): 270-290, 1997.

[17] Jason M. Lindo e María Padilla-Romo *Kingpin Approaches to Fighting Crime and Community Violence: Evidence from Mexico's Drug War.*, IZA Discussion Paper, No. 9067 (Maio), 2015.

[18] Beatriz Magaloni, Edgar Franco e Vanessa Melo *Killing in the Slums: An Impact Evaluation of Police Reform in Rio de Janeiro.*, Working Paper Stanford, No. 556 (Dezembro), 2015.

[19] Pablo Montero, José A. Vilar *TSclust: An R Package for Time Series Clustering.*, Journal of Statistical Software 2014, Vol. 62 (Novembro), 2014.

APÊNDICE A

A abordagem estocástica através de algortimo EM [3] funciona[4] e corrobora os resultados encontrados acima. A medida em que se aumenta o número de clusters, pode-se ver na figura 15 que o poder do algoritmo aumenta segundo a medida BIC. Destaca-se a possibilidade de simulação para diferentes valores da medida k que define a função de similaridade entre as séries de tempo, como feito anteriormente. Essa abordagem pode ser mais interessante para modelos de previsão. Em um trabalho futuro poderíamos verificar se essa abordagem traria maior acurácia ao prever os clusters. O problema, do ponto de vista analítico, é que esse algoritmo permite ruído, isto é, há a possibilidade de um batalhão estar em mais de um cluster simultaneamente. Segundo a medida BIC, o algoritmo ótimo (VEI) - Volume variante, Shape estático e Orientação sendo a Identidade. Pode-se ver o resultado da classificação BIC a seguir:

[3] *Expectation Maximization* : Um modelo de misturas de Gauss estimado através de Máxima Verossimilhança.

[4] Importante destacar que a primeira fase do algoritmo se dá através da também clusterização hierárquica. No segundo passo, misturas de Gauss são utilizadas para aproximar o shape dos dados.

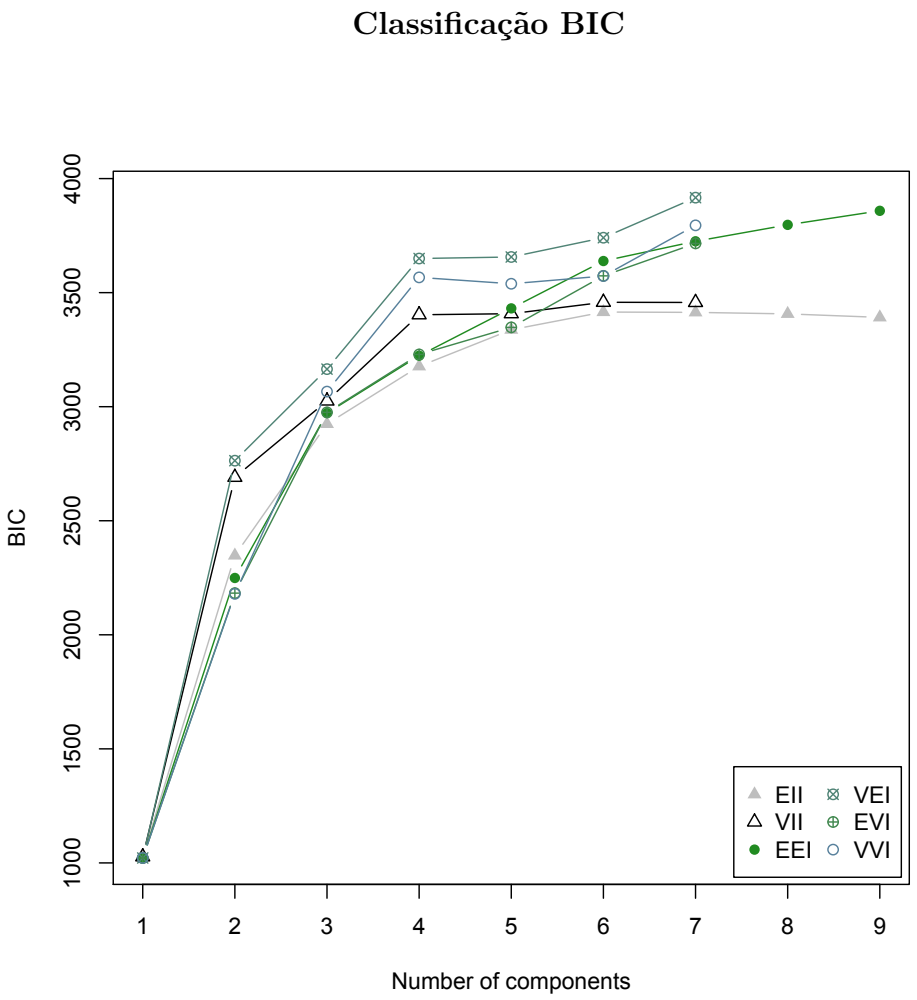

Figura 15 – BIC x Clusters

Além disso, a classificação se mostra bem próxima de nossa abordagem final, como esperado:

Tabela 9 – Classificação para o algoritmo EM

	Classificação
aisp2	1.00
aisp3	2.00
aisp4	3.00
aisp5	4.00
aisp6	1.00
aisp7	5.00
aisp8	3.00
aisp9	2.00
aisp10	6.00
aisp11	6.00
aisp12	4.00
aisp14	2.00
aisp15	7.00
aisp16	4.00
aisp17	1.00
aisp18	3.00
aisp19	1.00
aisp20	7.00
aisp21	4.00
aisp22	3.00
aisp23	1.00
aisp24	3.00
aisp25	3.00
aisp26	6.00
aisp27	3.00
aisp28	1.00
aisp29	6.00
aisp30	6.00
aisp31	3.00
aisp32	3.00
aisp33	1.00
aisp34	1.00
aisp35	1.00
aisp36	6.00
aisp37	6.00
aisp38	6.00
aisp39	3.00
aisp40	3.00
aisp41	5.00

APÊNDICE B

Neste apêndice disponibilizamos as estatísticas descritivas dos dados de cada cluster em separado como é possível ver nas tabelas de 1 a 6.

Tabela 10 – Cluster 1

	Índice de criminalidade	Roubo de Veículos	Roubo de Rua	Letalidade	Roubo de Carga	Furtos
Mínimo	0.00	0.00	0.00	0.00	0.00	0.00
Máximo	0.03	0.02	0.04	0.05	0.04	0.11
Média	0.01	0.00	0.01	0.01	0.00	0.016
Desvio Padrão	0.00	0.00	0.01	0.01	0.01	0.01
Variância	0.00	0.00	0.00	0.00	0.00	0.00
Mediana	0.01	0.00	0.00	0.01	0.00	0.01
N	986	986	986	986	986	986

Tabela 11 – Cluster 2

	Índice de criminalidade	Roubo de Veículos	Roubo de Rua	Letalidade	Roubo de Carga	Furtos
Mínimo	0.0115	0.0028	0.0044	0.0000	0.0000	0.0072
Máximo	0.0435	0.0542	0.0544	0.0922	0.0814	0.1287
Média	0.0218	0.0161	0.0182	0.0306	0.0180	0.0261
Desvio Padrão	0.0052	0.0076	0.0076	0.0197	0.0139	0.0114
Variância	0.0000	0.0001	0.0001	0.0004	0.0002	0.0001
Mediana	0.0211	0.0147	0.0174	0.0292	0.0143	0.0263
N	638	638	638	638	638	638

Tabela 12 – Cluster 3

	Índice de criminalidade	Roubo de Veículos	Roubo de Rua	Letalidade	Roubo de Carga	Furtos
Mínimo	0.02	0.00	0.02	0.00	0.00	0.01
Máximo	0.05	0.07	0.07	0.07	0.12	0.16
Média	0.04	0.04	0.04	0.02	0.04	0.04
Desvio Padrão	0.01	0.02	0.01	0.01	0.03	0.03
Variância	0.00	0.00	0.00	0.00	0.00	0.00
Mediana	0.04	0.04	0.04	0.02	0.03	0.04
N	232	232	232	232	232	232

Tabela 13 – Cluster 4

	Índice de criminalidade	Roubo de Veículos	Roubo de Rua	Letalidade	Roubo de Carga	Furtos
Mínimo	0.04	0.03	0.04	0.01	0.01	0.02
Máximo	0.06	0.09	0.08	0.08	0.11	0.05
Média	0.05	0.06	0.06	0.03	0.06	0.04
Desvio Padrão	0.01	0.01	0.01	0.01	0.02	0.01
Variância	0.00	0.00	0.00	0.00	0.00	0.00
Mediana	0.05	0.06	0.06	0.03	0.06	0.04
N	116	116	116	116	116	116

Tabela 14 – Cluster 5

	Índice de criminalidade	Roubo de Veículos	Roubo de Rua	Letalidade	Roubo de Carga	Furtos
Mínimo	0.0363	0.0513	0.0517	0.0091	0.0169	0.0187
Máximo	0.0958	0.1231	0.1124	0.0866	0.2286	0.0486
Média	0.0658	0.0920	0.0696	0.0452	0.0902	0.0319
Desvio Padrão	0.0120	0.0160	0.0103	0.0146	0.0501	0.0066
Variância	0.0001	0.0003	0.0001	0.0002	0.0025	0.0000
Mediana	0.0658	0.0941	0.0678	0.0428	0.0811	0.0313
N	116	116	116	116	116	116

Tabela 15 – Cluster 6

	Índice de criminalidade	Roubo de Veículos	Roubo de Rua	Letalidade	Roubo de Carga	Furtos
Mínimo	0.0549	0.0560	0.0487	0.0329	0.0215	0.0239
Máximo	0.1078	0.1584	0.1195	0.1568	0.1739	0.0707
Média	0.0838	0.1070	0.0896	0.0877	0.0904	0.0442
Desvio Padrão	0.0111	0.0189	0.0130	0.0227	0.0341	0.0077
Variâcia	0.0001	0.0004	0.0002	0.0005	0.0012	0.0001
Mediana	0.0832	0.1093	0.0897	0.0847	0.0860	0.0431
N	174	174	174	174	174	174