Sónia Clara Viegas Henriques

ANÁLISE ESPACIAL E TEMPORAL DE UMA BASE DE DADOS CRIMINAIS

Mestrado em Química Forense

Departamento de Química

FCTUC

Janeiro 2014

UNIVERSIDADE DE COIMBRA

Sónia Clara Viegas Henriques

ANÁLISE ESPACIAL E TEMPORAL DE UMA BASE DE DADOS CRIMINAIS

Dissertação apresentada a provas de Mestrado em Química,
Área de especialização em Química Forense

Orientador: Professor Doutor Alberto António Caria Canelas Pais

Janeiro 2014

Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Departamento de Química

Agradecimentos

Ao longo do Mestrado em Química Forense foram muitas as pessoas que cruzaram o meu caminho e que com certeza deixaram um pouco de si. Os momentos de alegria serviram para me permitir acreditar na beleza da vida, os de sofrimento, serviram para um crescimento pessoal único. O espaço limitado desta secção de agradecimentos, seguramente, não me permite agradecer como devia, a todas as pessoas que ao longo deste percurso me ajudaram, direta ou indiretamente, a cumprir os meus objetivos e a realizar esta etapa da minha formação académica. Sendo difícil transformar sentimentos em palavras, estarei eternamente grata a todas as pessoas imprescindíveis para a realização e conclusão deste projeto.

Ao Professor Doutor Alberto António Caria Canelas Pais, pela orientação, disponibilidade e paciência e pelos ensinamentos e entusiasmo que me transmitiu; para quem é feliz aquele que transfere o que sabe, e aprende o que ensina.

À Mestre Tânia Firmino G. G. Cova, agradeço a ajuda, o apoio e os ensinamentos relativos à Análise Multivariada, bem como a amizade, disponibilidade e prontidão ao longo deste projeto.

Ao Departamento de Química pela oportunidade e pelo reconhecido exemplo de competência que dá prestígio a esta Faculdade.

À Ana Rita Ferreira Matos, amiga de longa data, agradeço a partilha de bons momentos, a ajuda e o incentivo nas alturas de desânimo, que me permitiram que cada dia fosse encarado com particular motivação.

A todos os colegas e amigos, obrigado pela cumplicidade, paciência e ajuda em muitos momentos. A amizade deixa-nos caminhar à vontade lado a lado, estimula e partilha a nossa maneira de olhar a vida.

À minha família, em especial aos meus pais, ao meu irmão e aos meus avós, um enorme obrigado por acreditarem sempre em mim e naquilo que faço e por todos os ensinamentos de vida. Espero que esta etapa que agora termino, possa de alguma forma, retribuir e compensar todo o carinho, apoio e dedicação que, constantemente, me oferecem.

A todos aqueles que, não estando aqui mencionados, de alguma forma me apoiaram na realização deste mestrado.

Aos meus pais e irmão

“Só pelo conhecimento

se pode evitar a criminalidade.”

Maurice Cusson

Índice

Lista de abreviaturas

BJS – Bureau of Justice Statistics

EUA – Estados Unidos da América

FBI – Federal Bureau of Investigation

HCA – Hierarchical Cluster Analysis

IACP – International Association of Chiefs of Police

NCVS – National Crime Victimization Survey

NIBRS – Reporting Incident-Based Reporting System

PC_i – Principal Component

PCA – Principal Component Analysis

UCR – Uniform Crime Reporting

UCRRP – UCR Redevelopment Project

Resumo

Este trabalho assenta na aplicação de métodos quimiométricos no tratamento da informação multivariada relativa a estatísticas criminais.

Uma das principais contribuições destes métodos de tratamento de informação é claramente auxiliar a interpretação e racionalização dos dados que passa, em grande medida, pela deteção dos padrões subjacentes. Só quando estes são conhecidos se podem criar modelos de previsão e desenvolver políticas de prevenção e controlo.

Neste contexto a análise multivariada é incontornável quando se está perante informação proveniente de sistemas multicomponentes. Na maioria dos casos, a inspeção do perfil global dos dados é claramente mais informativa que a avaliação parâmetro a parâmetro.

Neste trabalho é efetuada a caracterização de uma variedade de dados relativos a detenções e a delitos cometidos nos Estados Unidos da América durante o período de 2005 a 2011. Os resultados mostram que a combinação dos métodos quimiométricos selecionados (análise de agrupamentos hierárquico, análise de componentes principais) com algumas medidas econométricas (curva de *Lorenz* e coeficiente de *Gini*), permite tirar conclusões sobre a distribuição e relação espácio-temporal de vários tipos de crimes. A abordagem proposta pode servir de base para futuros desenvolvimentos na criação de modelos e análise de fatores, também aqui enunciados.

Especificamente, os crimes relacionados com a droga e o álcool desempenham um papel fulcral na distinção entre os diversos estados americanos. Existe uma variação gradual das características criminais desde os estados rurais do centro Norte até à periferia mais rica e urbanizada. A correlação com outros indicadores, nomeadamente a população e o crime, foram também encontrados.

Abstract

This work is based essentially on the application of standard chemometric methods in the treatment of multivariate data from crime statistics.

A major contribution of these techniques is clearly the rationalization and interpretation of the underlying patterns related to national crime occurrences, extremely useful for crime policy-making prevention and control.

In this context, multivariate analysis is paramount when faced with information from multicomponent systems. In most cases, the inspection of the overall data profile is considerably more informative than the evaluation parameter to parameter.

In this work, the data from crimes committed in the United States during the period of 2005-2011 are evaluated. The results show that the combination of standard chemometric methods (hierarchical cluster analysis and principal component analysis) and some econometric measures (e.g. Lorenz curve and Gini coefficient), provide valuable information on spatial and time distribution of the different crime categories under scrutiny.

Specifically, crimes related to drugs and alcohol play an important role in the discrimination of the American states (from the Northern rural center ones to the periphery richer and more urbanized). Some correlation with other indicators is found, such as population and crime.

Capítulo 1

Padrões criminais

1.1 Aspetos geográficos e temporais

Nesta dissertação é realizada a análise espacial e temporal dos padrões criminais dos Estados Unidos da América (EUA), recorrendo a dados provenientes do programa *Uniform Crime Reporting* do FBI (UCR). Na realidade, a origem e evolução dos temas relacionados com os delitos e as penas estão pontuadas de fases de progresso e de retrocesso, de verdades descobertas e depois esquecidas, bem como de grandes oscilações pendulares [1]. Um dos aspetos que torna relevante a análise e o reconhecimento de padrões prende-se com a avaliação do custo do crime para a sociedade. Esta avaliação, encontra-se na base do desenvolvimento de muitos programas e indicadores sociais relacionados por exemplo, com o tratamento da dependência ou com o policiamento [2]. Diversas áreas de atividade económica podem apresentar uma sensibilidade diferente ao crime [3]. Certas comunidades vêem extremamente reduzida a sua atividade económica devido à existência, por exemplo, de crime violento.

Na literatura recente, é evidenciada uma outra preocupação associada à correspondência entre os índices de criminalidade a nível urbano e a nível nacional [4]. Alguns estudos referem a existência de um padrão claro e único, para os grandes centros urbanos e uma tendência nacional significativa. A pergunta persiste " O crime, tal como a política, é uma preocupação local, ou as taxas de criminalidade seguem um padrão nacional mais amplo?" [5].

A caracterização geográfica surge, recentemente, a par com a análise de padrões criminais [6] e permite a identificação de estruturas espaciais dinâmicas numa análise socioeconómica de crime (ou crimes) [7]. Um exemplo de uma aplicação que adveio do impacto de políticas quer a nível local, quer a nível global, foi o chamado *Project Safe Neighborhoods* [8], que motivou um estudo sobre tendências do crime violento em todas as cidades dos Estados Unidos com mais de 100 000 habitantes.

A obtenção de uma relação causal não é no entanto, uma tarefa fácil, [9] dado que, estão implícitos inúmeros fatores que influenciam o crime.

Os criminologistas sempre tentaram explicar porque é que certos tipos de crime, ou diferentes níveis de criminalidade, são encontrados de forma distinta nas várias comunidades.

Apesar da abordagem proposta ter uma perspetiva mais ampla, até ao momento as grandes comunidades são tendencialmente consideradas o foco primário da teoria e investigação criminal. Verifica-se também, uma transferência de foco para pequenas comunidades com pouco mais de quarenta habitantes, nas quais os atributos do próprio lugar e as atividades de rotina se combinam para o desenvolvimento de eventos criminais [10]. Daí que certas análises sejam feitas em segmentos de rua, tornando-se difícil obter padrões coerentes [11]. O facto de se analisarem padrões em comunidades mais reduzidas não é limitativo, uma vez que, é possível recorrer a técnicas numéricas sofisticadas [12].

Neste contexto, algumas objeções têm sido levantadas relativamente a modelos baseados em variáveis reduzidas, que impedem uma avaliação de escala [13]. Uma avaliação à escala global dos padrões do crime tem de ser conduzida antes da avaliação em zonas específicas, por mais significativas que estas sejam [14]. Por vezes, o estudo é feito diretamente sobre as amostras consideradas representativas. Estas amostras podem ser de dimensões razoáveis e sobre elas procede-se à análise de fatores determinantes, como por exemplo, o uso de drogas [15]. O facto de as amostras serem de dimensão reduzida, leva a que sejam detetadas variações espaciais de grande amplitude [16], o que conduz à necessidade da identificação de microestruturas [17]. Tem-se constatado, recorrendo a argumentos de alguma complexidade, baseados em influências temporárias e permanentes, diferenças nos estudos inter e intra comunidades [18]. Note-se que a composição da amostra também tem sido alvo de preocupação por parte dos investigadores, nomeadamente, quando se pretende analisar fatores relacionados com a composição racial ou o desemprego [19]. Incertezas que surjam na amostragem podem produzir incoerências em estudos análogos, impedindo deste modo, o estabelecimento de conclusões gerais [20].

Numa outra vertente, muitos têm sido os estudos efetuados para relacionar as características do meio envolvente e o crime. A modelação passa pela utilização de características relacionadas com o nível económico, a mobilidade residencial, a heterogeneidade racial e a estabilidade familiar [21]. Neste âmbito, a análise é complexa e assenta num modelo de desorganização social que poucas vezes é conclusivo. O conceito de desorganização social também tem sido associado a outros aspetos relacionados com a evolução económica. Verifica-se por exemplo, que o aumento do desemprego nas indústrias produtivas conduz a um aumento das taxas de *aggravated assault, larceny-theft* e *burglary* [22]. Sendo o conceito de desorganização social muito amplo, as análises que daí advenham são de grande complexidade [23].

Na identificação dos padrões criminais existem aspetos importantes a ter em conta, principalmente nos que diz respeito à escolha daqueles que promovem uma caracterização mais direta desses padrões e as suas inter-relações. Tal significa, também, que para uma análise temporal

é necessário avaliar detalhadamente as variáveis sobre as quais se faz a descrição. De outra forma, não é possível encarar a evolução temporal sem analisar em profundidade, cada um dos pontos temporais.

A análise espacial do crime e o foco atual em *hotspots* tem afastado a área de mapeamento do crime, principalmente em crimes de grande volume [24]. Ao direcionar a teoria criminal em *hotspots*, zona de grande intensidade de crime, consegue-se uma redução na taxa de crime. Centrar os agentes policiais em áreas de maior necessidade tem sido uma grande aposta, principalmente em tempos de restrição fiscal. Assim, para além da redução do crime, os *hotspots* têm-se mostrado fundamentais para a estratégia de policiamento em muitos locais [25]. No entanto, pouco esforço tem sido canalizado para a análise temporal de padrões criminais. Na tentativa de analisar a sazonalidade do crime, que se refere a flutuações periódicas anuais, os estudos são algo inconclusivos, implicando aspetos relacionados com a escolha de critérios estatísticos [26]. Para além de estudos de sazonalidade, alterações a longo prazo têm sido analisadas [27, 28], bem como relações entre diferentes períodos do dia e o crime [29].

O desenvolvimento científico e tecnológico, principalmente a nível computacional, tornou a investigação do crime mais detalhada. O recurso a tecnologias de mapeamento uniformizado das forças policiais permite a visualização de padrões espácio-temporais do crime [30]. As características do *software* de análise de crimes, nomeadamente para o estabelecimento de estratégias preventivas e de deteção têm sido alvo de constante optimização [31].

A análise espacial e temporal, proposta nesta dissertação, tem como principal objetivo a previsão de eventos criminais [32-33]. A previsão geo-temporal, têm sido desenvolvida recorrendo, por exemplo, a métodos de análise como redes neurológicas artificiais [34]. A análise espacial do crime é bastante semelhante a uma análise epidemiológica, estando os métodos utilizados em ambas as análises também relacionados [35].

1.2 Relações causais

Os fatores que se relacionam, por exemplo, com a taxa de criminalidade ocorrem em grande número e são de natureza muito variada. Uma análise recorrente das relações causais associadas ao crime, tem como base a idade, uma vez que, a associação da previsão de tendências para o crime assenta, frequentemente, em dados demográficos. Em muitas situações as detenções ocorrem primordialmente sobre adolescentes e jovens adultos [36]. Embora este tipo de informação conste na base de dados UCR, não será explicitamente usada no contexto deste trabalho. Note-se, no entanto, que a previsão com base em dados demográficos, sendo atraente, pode ser pouco

adequada. De facto, a relação idade/crime surge muitas vezes enfraquecida pois trata-se de um fator que é facilmente diluído por outros, tais como fatores sociais e económicos [37]. Quando a ligação entre idade e crime se considera robusta, várias são as tentativas para explicar a etiologia desta relação [38]. Na realidade, os resultados não são totalmente conclusivos [39] devido à forma como cada fator é medido, dificultando o estabelecimento da relação.

Também frequentes são os estudos relacionados com a origem étnica ou a composição racial, por vezes, associado a aspetos da imigração [7]. Ainda ao nível dos fatores demográficos, fatores como a densidade populacional ou o número de agentes policiais *per capita* têm sido avaliados. Se a densidade populacional tem influência no crime contra a propriedade, as relações com número de agentes policiais *per capita* estão um pouco diluídas, na medida em que parece haver alguma especificidade no tipo de crime para o estabelecimento da relação [40]. Contudo, nos últimos 30 anos, a identificação destes fatores tem levado a teorias pouco conclusivas [41].

Estudos relacionados com o género têm sido utilizados como base para a avaliação de outros indicadores, nomeadamente, fatores estruturais e de estado social (por exemplo, casamento, emprego, educação, pobreza). Note-se que *larceny-theft* é o crime contra a propriedade mais comum para ambos os géneros, seguido de *burglary* [42]. Outros estudos, que avaliam um conjunto de fatores sociodemográficos em conjunto com uma separação em género [43], mostram não ser possível identificar nos detidos, diferenças entre os géneros, no que diz respeito à idade, nível de educação e fontes de rendimento. Em alternativa, as análises incidem num só género e são efetuadas a nível nacional [44].

Apesar das limitações referidas, a análise de fatores pode ir mais adiante, incluindo aspetos de autocontrolo tornando, mais difícil ainda, a sua previsão [45]. Note-se que este mesmo conceito surge por vezes associado ao comportamento juvenil violento [46]. Na mesma linha de ideias, a relação entre o coeficiente de inteligência e envolvimento em atividades criminais ou de delinquência tem sido também investigada [47-48].

Os fatores económicos são também alvo de grande preocupação, nomeadamente na sua relação com crimes sobre a propriedade. Uma análise recorrente baseia-se nos efeitos do desemprego. A relação com o desemprego é, no entanto, um problema longe de uma solução evidente. Verifica-se que, uma análise mais rigorosa sobre esta relação exige, naturalmente, a existência de bons indicadores de desemprego e de criminalidade [49]. A influência da desigualdade dos rendimentos sobre a taxa de criminalidade tem sido também estudada, mas novamente o resultado é controverso [50]. Verifica-se que, a desigualdade está positivamente associada com a taxa de criminalidade, mas negativamente quando associada a uma evolução temporal. No mesmo sentido, verifica-se que esta desigualdade tem efeitos significativos em crimes como *burglary* e

robbery [51]. Esforços para a construção de modelos globais, descrevendo a relação entre crime, dissuasão e variáveis socio económicas têm sido também desenvolvidos [52-53]. Fatores económicos, sociais e políticos são frequentemente analisados, quando se caracteriza a evolução temporal, ao longo de um largo período de tempo, [54]. Entre outros indicadores, o índice de miséria é também usado para estabelecer uma relação com a taxa de criminalidade, sendo esta relação não surpreendentemente positiva [55].

Grande parte dos estudos relacionados com fatores pretendem estabelecer relações causais. Um dos exemplos é a avaliação das relações recíprocas entre a estabilidade residencial nas comunidades e o crime violento. Confirma-se que, certas comunidades mais estáveis têm um efeito protetor contra a violência [56].

As relações entre vários tipos de crime, por exemplo, a relação existente entre o álcool, drogas e crime violento [57], bem como a relação entre álcool e *rape* [58] têm sido debatidas na literatura.

Aproximações globais, particularmente ao abrigo da chamada ciência social evolucionária, têm sido propostas, considerando diversos fatores como o género, a poligamia e a educação, na relação com a criminalidade [59]. Relações pouco comuns de caracter ambiental, como a relação entre perfis criminais e a exposição ao chumbo [60] ou o aquecimento global e o crime são também avaliadas [61].

A existência de fatores que traduzem relações pouco conclusivas leva a que em estudos muito recentes, se tenha procedido a uma re-identificação de alguns fatores, especialmente aqueles que pesam na capacidade preditiva [62]. A realização deste tipo de estudos torna-se cada vez mais importante, na medida em que o crime é um indicador de colapso social [63].

Capítulo 2

A base de dados *Uniform Crime Reporting*

O Departamento de Justiça dos Estados Unidos da América é formado por diversas agências, com destaque para o *Federal Bureau of Investigation* (FBI) que gere o programa estatístico *Uniform Crime Reporting* (UCR) (Figura 2.1), e o *Bureau of Justice Statistics* (BJS) que coopera com a anterior através do programa *National Crime Victimization Survey* (NCVS). Ambos os programas avaliam o impacto da criminalidade no país quanto à sua extensão e natureza. Cada um destes programas fornece informações úteis sobre diversos aspetos do crime nos EUA.

Figura 2.1 – Símbolo da *Uniform Crime Reporting* [64].

Neste trabalho, são caracterizados os dados provenientes do programa UCR [64], que passamos a descrever em pormenor.

2.1 O programa UCR

O programa UCR reúne dados provenientes de relatórios mensais, registos de incidentes individuais, transmitidos diretamente para o FBI e de agências estaduais centralizadas, que reportam ao FBI. Este programa é dirigido pelo FBI [65][1], desde Setembro de 1930. Atualmente, esta agência recolhe informação sobre diversos tipos de crime: *murder and nonnegligent mausnlaughter, forcible rape, robbery, aggravated assault, burglary, larceny-theft, motor vehicle theft, e arson*. Adicionalmente, dados relativos a detenções para vinte e uma categorias de crime são comunicados por entidades de aplicação de lei.

O UCR examina minuciosamente cada relatório que recebe, analisando a sua veracidade. As oscilações nos níveis de criminalidade podem apontar para a existência de registos modificados, relatórios inconclusivos ou alterações nos limites de jurisdição. Para que possam ser detetadas eventuais falhas na contagem de uma agência, o programa confronta relatórios mensais com submissões anteriores.

Este programa apresenta o número de crimes cometidos para o país como um todo incluindo as regiões, municípios, cidades, vilas, universidades e estados. Na Figura 2.2 estão

[1] Veja-se a referência que apresenta alguns dados interessantes sobre esta organização.

representadas as quatro regiões geográficas principais dos Estados da América, amplamente reconhecidas na organização dos dados de criminalidade do país. O grande volume de informação disponível permite a realização de estudos a vários níveis, tais como, o planeamento municipal e o auxílio a estudantes na justiça criminal, legisladores, criminólogos e sociólogos em pesquisas avançadas. Trata-se de um esforço estatístico a nível nacional, que envolve a colaboração de mais de 18.000 cidades, universidades, municípios, estados, organismos tribais e agências policiais federais que relatam voluntariamente os crimes que são do seu conhecimento.

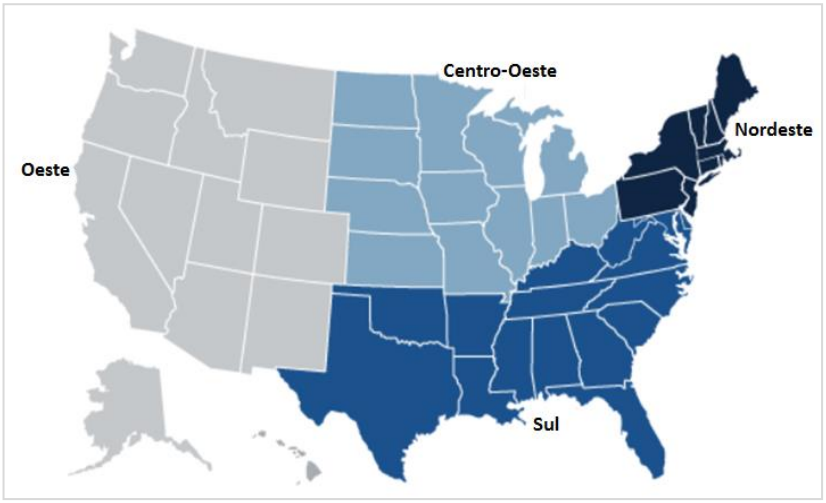


Figura 2.2 – Mapa geográfico dos EUA, no qual se encontram evidenciadas as quatro regiões geográficas usadas pelo programa UCR na seleção de dados criminais (adaptado de [65]).

Numa perspetiva meramente estatística, em 2011, as agências ativas no programa UCR representaram mais de 304 milhões de habitantes dos Estados Unidos (97.8% da população total). A cobertura foi de 98.8% da população na estimativa relativa às áreas metropolitanas, 92.3% da população nas cidades fora das áreas metropolitanas e 93.1% da população nos municípios não metropolitanos.

2.1.1 Contexto histórico

O programa UCR foi concebido em 1929 pela Associação Internacional de Chefes de Polícia (em inglês, *International Association of Chiefs of Police*, *IACP*) para atender à necessidade de um sistema de estatística criminal uniforme e confiável para o país. Após um estudo exaustivo dos códigos penais dos estados e a realização de uma avaliação da prática dos registos em uso, a comissão, nesse mesmo ano, elaborou um plano para relatar o crime que se tornou a base do programa UCR. O plano inclui definições padrão de crimes para sete principais infrações, conhecidas como crimes de tipo I, que incluem os crimes violentos de *murder and nonnegligent mauslaughter,*

forcible rape, robbery e *aggravated assault* e os crimes de tipo II, designados de crimes contra a propriedade, que incluem *burglary, larceny theft* e *motor vehicle theft*. A classificação destes tipos de crime, deve-se a uma regra de hierarquia em que os crimes de tipo I, têm uma escala de força ou ameaça de força, que varia de um nível máximo em *murder and nonnegligent mauslaughter* até um nível mínimo em *aggravated assault*. Os fundadores do UCR criaram esta classificação de crime com o objetivo de medir oscilações no volume global e na taxa de crime. A escolha deve-se à gravidade do crime, à frequência e também à suscetibilidade de serem comunicados às autoridades policiais.

Em Janeiro de 1930, quatrocentas cidades, que representam vinte milhões de habitantes distribuídos ao longo de quarenta e três estados, começaram a sua participação no programa UCR. O congresso aprovou uma autorização que permitiu ao Procurador-Geral recolher informação criminal. Por sua vez, o Procurador-Geral elegeu o FBI para proceder à recolha de todos os dados criminais disponíveis. Desde então, anualmente, os dados baseados nas classificações uniformizadas e os procedimentos para reportar crimes e detenções, passaram a ser obtidos a partir das agências de aplicação da lei a nível nacional. Por ordem do Congresso, em 1979, o crime *arson* foi acrescentado como o oitavo a fazer parte das infrações do crime tipo I.

Apesar da recolha de dados ter permanecido praticamente inalterada ao longo dos anos, na década de 1980 o programa tinha evoluído para uma aplicação mais ampla. Reconhecendo a necessidade de melhoramento da estatística, os organismos de aplicação da lei procederam a um esforço de modernização. O FBI forneceu o seu apoio formulando uma reestruturação do crime em 3 fases. Em primeiro lugar, as agências passariam a usar um sistema baseado no relato de incidentes para reportar crimes e detenções, designado por *National Incident-Based Reporting System (NIBRS)*. Em segundo, o programa nacional UCR passaria a recolher informação a dois níveis (*limited and full participation*) e em terceiro o UCR iria introduzir um programa de garantia e qualidade.

O final do desenvolvimento e a gestão por parte do FBI do programa UCR e NIBRS deu-se no final da década de 1980. Começou a receber, a partir de um número reduzido de agências, os dados provenientes deste sistema em Janeiro de 1989. À medida que as agências de aplicação de lei passaram a estar informadas relativamente às vantagens inerentes a esta forma de recolha de dados, começaram a aderir ao sistema. Considera-se que a inclusão do NIBRS fez desaparecer algumas lacunas do programa UCR tradicional, permitindo recolher informação proveniente das vítimas. Note-se que, este facto permitiu também estabelecer o paralelo entre NIBRS e NCVS [66].
Recentemente foi criado um Projeto de Requalificação UCR (em inglês, UCR Redevelopment Project, UCRRP) com vista a melhorar a eficiência, usabilidade e manutenção de processos do programa. Um dos grandes objetivos deste projeto será a redução, até ao ponto da eliminação, da troca de

materiais impressos entre as agências e o FBI. Assim, desde Julho de 2013 o programa não aceita mais pedidos em papel adotando submissões eletrónicas através do NIBRS.

2.1.2 Definição de delitos

O programa do *Uniform Crime Reporting* (UCR) divide as ofensas em dois tipos, crimes de tipo I e tipo II, como referido anteriormente. Em cada mês, os organismos de aplicação da lei, que participam no programa, submetem a informação dos crimes de tipo I dos quais tomaram conhecimento. Para os restantes crimes (tipo II) são apenas fornecidas informações acerca das detenções efetuadas.

A recolha de dados relativos aos crimes de tipo I tem o objetivo de medir o nível e o âmbito dos crimes que ocorrem no país. Os fundadores deste programa escolheram estes crimes por serem os mais graves por ocorrerem com regularidade em todos os estados e por serem suscetíveis de ser comunicados à polícia.

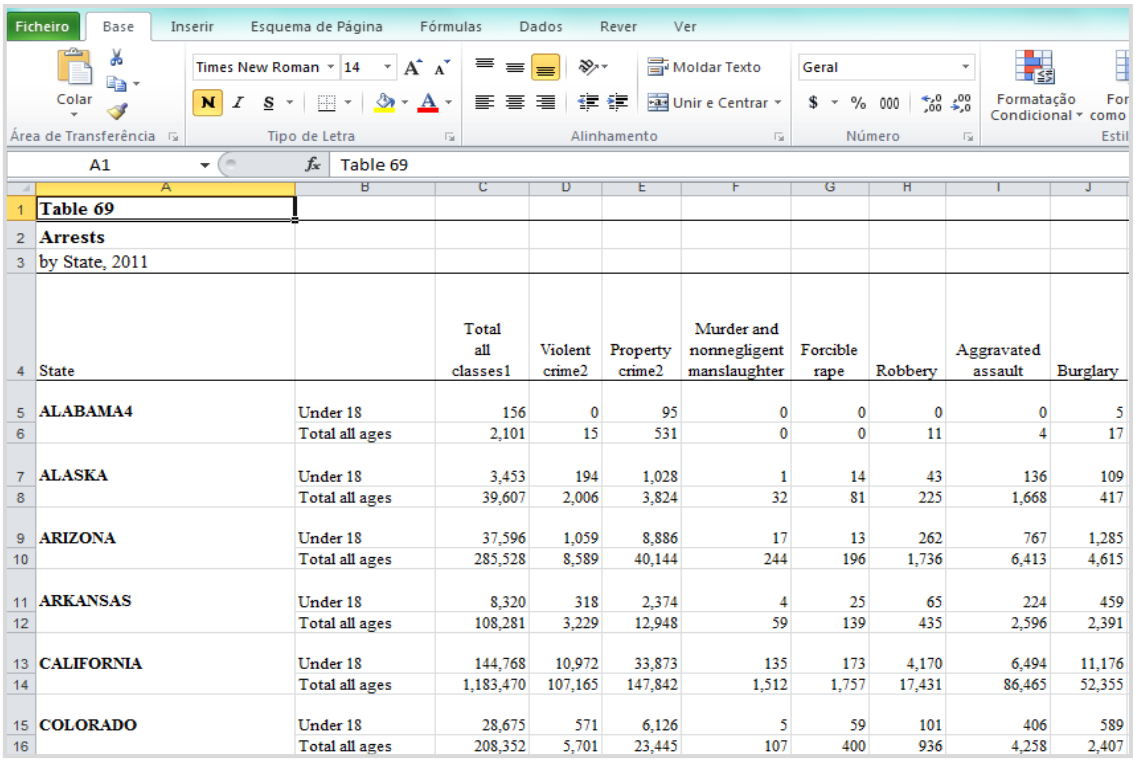
As infrações de tipo I e tipo II encontram-se descritas em detalhe no (Anexo 1).

2.1.3 Organização dos dados

O conjunto de dados analisado contém valores referentes às detenções efetuadas na população dos EUA (*Arrest, by State, 2011* [65]), nos anos de 2005, 2007, 2009 e 2011. Estas estimativas são disponibilizadas pela base de dados UCR, que contempla os valores anuais referentes a detenções efetuadas para diferentes tipos de crime.

No presente estudo foram considerados 29 tipos de crime, parcialmente descritos na Figura 2.3, como exemplo ilustrativo, e respeitando a designação original da base de dados.

Nesta Figura, os dados correspondem às detenções efetuadas em cada Estado, durante o ano de 2011, para 29 tipos de crimes. A tabela discrimina as detenções em número total e as detenções realizadas por pessoas de idade inferiores a 18 anos. Os dados representam o número de pessoas detidas, mas deve ser tida em conta a possibilidade de uma pessoa ser detida mais que uma vez ao longo do ano. Em alguns casos, o valor pode representar múltiplas detenções da mesma pessoa.



	A	B	C	D	E	F	G	H	I	J
1	**Table 69**									
2	**Arrests**									
3	by State, 2011									
4	State		Total all classes1	Violent crime2	Property crime2	Murder and nonnegligent manslaughter	Forcible rape	Robbery	Aggravated assault	Burglary
5	**ALABAMA4**	Under 18	156	0	95	0	0	0	0	5
6		Total all ages	2,101	15	531	0	0	11	4	17
7	**ALASKA**	Under 18	3,453	194	1,028	1	14	43	136	109
8		Total all ages	39,607	2,006	3,824	32	81	225	1,668	417
9	**ARIZONA**	Under 18	37,596	1,059	8,886	17	13	262	767	1,285
10		Total all ages	285,528	8,589	40,144	244	196	1,736	6,413	4,615
11	**ARKANSAS**	Under 18	8,320	318	2,374	4	25	65	224	459
12		Total all ages	108,281	3,229	12,948	59	139	435	2,596	2,391
13	**CALIFORNIA**	Under 18	144,768	10,972	33,873	135	173	4,170	6,494	11,176
14		Total all ages	1,183,470	107,165	147,842	1,512	1,757	17,431	86,465	52,355
15	**COLORADO**	Under 18	28,675	571	6,126	5	59	101	406	589
16		Total all ages	208,352	5,701	23,445	107	400	936	4,258	2,407

Figura 2.3 – Representação da folha de cálculo com os dados fornecidos através do programa *Uniform Crime Reporting* (*Table 69* na notação da base de dados original) e correspondente às detenções efetuadas, para 29 tipos de crime, em cada Estado, durante o ano de 2011.

2.2 Fiabilidade do programa

Os dados do programa UCR não estão isentos de lapsos de informação. A origem destas assimetrias está na deficiente cobertura do território, ocasionada por erros de medição [67]. Na tentativa de completar estas deficiências, foi sugerida a partilha de dados entre as duas maiores fontes de informação [68-71] UCR e NCVS.

As falhas mais significativas são ao nível de esquemas de preenchimento que por vezes, se tornam inadequados e incoerentes [72]. Estas falhas surgem ao nível do *county*, e parecem justificar o aparecimento de tendências contra intuitivas, comprometendo os estudos que daí advenham.

Na utilização da informação proveniente de uma base de dados, não é a existência de erros que é relevante, mas sim a perceção de que a persistência desses erros pode condicionar e comprometer a caracterização e validade dessa informação. Sendo assim, é difícil aferir sobre o impacto desses erros sobre os dados. No entanto, a informação considerada neste estudo encontra-se validada para uma utilização a nível estadual.

Capítulo 3

Técnicas e métodos de análise multivariada

Nesta secção são abordados alguns conceitos relativos aos métodos quimiométricos e econométricos e à importância dos mesmos no tratamento de vários tipos de problemas, de acordo com a literatura mais recente. A informação analisada é fornecida pela base de dados UCR, descrita no capítulo 2.

A UCR contém informação sobre vários tipos de crime que ocorrem anualmente nos EUA. Especificamente, são inspecionados e avaliados os padrões criminais para os quatro anos considerados (2005, 2007, 2009 e 2011). A caracterização multivariada dos estados americanos é baseada em dois métodos quimiométricos bem estabelecidos: (i) a análise de agrupamento hierárquico (HCA[2]), para definição da estrutura dos dados e (ii) a análise de componentes principais (PCA[3]), para visualização geral e seleção das variáveis mais importantes para o sistema em estudo. Os métodos são implementados com recurso à linguagem de programação R versão (2.15.2) [73], usando como interface gráfica o RStudio [74]. Como exemplo ilustrativo, na Figura 3.1 encontra-se um fragmento da página principal da interface gráfica RStudio utilizada neste trabalho.

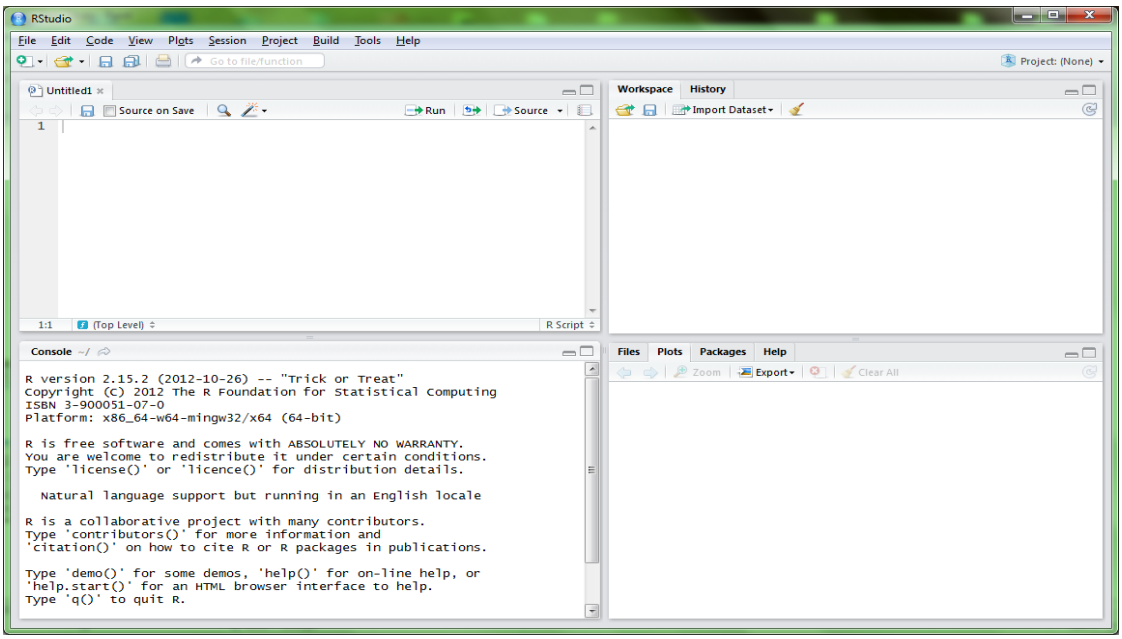


Figura 3.1 – Representação da página principal da interface gráfica do *RStudio*.

[2] em inglês, *Hierarchical Cluster Analysis*.
[3] em inglês, *Principal Component Analysis*.

O esquema proposto permite a inspeção de todo o perfil de dados, tornando-se mais informativo do que uma avaliação parâmetro a parâmetro. Os métodos escolhidos são complementares e distinguem corretamente os grupos existentes, sendo possível facilmente relacionar variáveis, isolar fatores e identificar marcas geográficas.

3.1 Considerações gerais

Atualmente, o fenómeno de erudição inerente a qualquer sociedade está relacionado com a aquisição e processamento de uma grande quantidade de informação proveniente de diferentes fontes.

A lista de métodos relacionados com a temática de explorar informação é grande e tende a crescer ainda mais. A dificuldade surge no estabelecimento de abordagens simples e compreensíveis, que estabeleçam diferenças entre os níveis de abrangência da sua aplicação. É comum encontrar métodos e técnicas desenvolvidas para aplicações específicas, na tentativa de dar resposta a questões de natureza ampla e complexa, o que, em alguns casos, leva a resultados contestáveis, que confirmam a dificuldade inerente ao tratamento da informação.

O desenvolvimento de métodos e técnicas capazes de racionalizar informação tem atraído grande interesse devido à possibilidade de localizar conhecimento útil a partir de grandes quantidades de dados.

O tratamento adequado de dados armazenados em repositórios permite o estabelecimento de correlações, padrões e tendências significativas, usando tecnologias de reconhecimento de padrões, assim como técnicas estatísticas e matemáticas. A extração da informação cujo objetivo é encontrar factos ocultos subjacentes às bases de dados, permite identificar padrões e relações latentes entre os dados e inferir regras que permitem predizer resultados futuros.

O processo de tratamento de dados consiste essencialmente em três pontos básicos: (i) exploração, (ii) construção do modelo ou definição do padrão e (iii) validação. No estudo do inter-relacionamento de variáveis, a análise revela-se normalmente complexa. Neste contexto, recorre-se particularmente à análise multivariada, que pode ser definida como um conjunto de técnicas e métodos que fazem uso de todas as variáveis na interpretação teórica dos dados [75]. Dentro desse conjunto, apenas um número reduzido de variáveis contêm os informação relevante, enquanto que as restantes pouco ou nada acrescentam à interpretação dos dados.

A crescente utilização dos métodos de análise multivariada é justificada quer pela diversidade de programas estatísticos que incluem estas metodologias, quer pela necessidade frequente de tratar e simplificar dados. Assim, é possível analisar um grande volume de dados, que

serão representados de uma forma simples, sem sacrificar a informação relevante. As ferramentas de análise multivariada permitem, assim, a ordenação e agrupamento de objetos[4] com características similares, a inspeção da interdependência dessas características, tarefas de previsão, e também, a construção e testes de hipóteses [75].

A redução de informação redundante é realizada através de critérios objetivos, permitindo a construção de gráficos bi ou tridimensionais contendo maior informação estatística, o que pode ser conseguido através de ferramentas como a análise de componentes principais (PCA).

É também possível construir agrupamentos entre objetos de acordo com a sua similaridade, utilizando todas as variáveis disponíveis, e representar estruturas numa forma bidimensional através de um dendrograma [76] num esquema de agrupamento hierárquico.

Estes métodos têm-se mostrado úteis no tratamento da informação, sendo utilizados em diversas áreas como a Química, Biologia e Medicina, Ciências Sociais, Ciências Económicas ou Engenharia. Têm sido, também, aplicados nas ciências forenses, arqueologia científica, geoquímica entre outras [77].

O presente trabalho visa a aplicação de algoritmos clássicos da Quimiometria no tratamento e interpretação da informação disponível em base de dados criminais, sob o ponto de vista espácio-temporal. Para complementar alguns resultados são utilizados métodos econométricos, como o Coeficiente de *Gini* e a Curva de *Lorenz*.

3.2 Quimiometria

A estatística univariada há muito que é aplicada a problemas químicos, mas a sua utilização tornou-se limitada. Nas últimas décadas, a análise multivariada foi introduzida no tratamento de dados químicos, aumentando a sua popularidade e dando origem à Quimiometria [78].

Por norma, a realização de medições e a caracterização de dados requerem a aplicação de procedimentos estatísticos básicos, que incluem, por exemplo, medidas centrais, medidas de dispersão e testes de significância. A criação de instrumentos analíticos mais complexos e a necessidade de lidar com grandes quantidades de dados experimentais são responsáveis pelo desenvolvimento de novas metodologias para o tratamento de resultados [79].

O termo "quimiometria" foi proposto pela primeira vez em 1972 por Bruce Kowalski e Svante Wold [80], ambos responsáveis pela criação da Sociedade Internacional de Quimiometria [81]. São considerados os fundadores desta nova disciplina com aplicação num campo interdisciplinar que combina estatística multivariada, modelação matemática, computação e química

[4] Objeto é, na prática, um ponto no espaço multivariável.

[82]. Com a aplicação de métodos de análise multivariada, propõe-se assim a racionalização, compreensão e visualização das relações existentes entre as variáveis e os objetos de estudo [83].

Os métodos quimiométricos possuem um enorme potencial no tratamento de diversos tipos de problemas. No entanto, até a segunda metade dos anos 80, o reconhecimento da sua importância na literatura existente era escassa. A partir dos anos 90 verifica-se um aumento significativo das aplicações, levando também ao aumento da capacidade dos investigadores em extrair informações dos dados [84]. Naturalmente, muito deste esforço surgiu em resposta à necessidade de desenvolver novos métodos matemáticos e estatísticos para lidar com a vasta quantidade de dados produzida pelos instrumentos analíticos modernos [84].

Na atualidade, a quimiometria permite tratar dados complexos, que requerem a utilização de técnicas estatísticas multivariadas, álgebra matricial e análise numérica [85].

A disponibilidade de computadores mais potentes e menos dispendiosos contribui também para a rápida evolução dos métodos quimiométricos, ao permitir: (i) uma análise mais flexível de grandes conjuntos de dados, (ii) o desenvolvimento de algoritmos computacionais mais eficientes e (iii) a rápida difusão do *software* quimiométricos.

A Quimiometria envolve diferentes métodos, tais como a otimização e validação de metodologias analíticas, o planeamento experimental, a estimativa de parâmetros, o processamento de sinal, a análise de fatores, a calibração multivariada, a utilização de redes neurológicas artificiais, o reconhecimento de padrões e o processamento de imagem [79].

Na literatura, são publicadas regularmente revisões detalhadas sobre a aplicação dos métodos quimiométricos em várias áreas do conhecimento (veja-se por exemplo [86]). É difícil enumerar em detalhe cada contribuição histórica inerente aos primeiros anos da quimiometria. Em geral, estes métodos têm sido aplicados com sucesso na visualização dos dados, na classificação na resolução de curvas multivariadas e na predição em química analítica, química ambiental, engenharia, investigação médica e na indústria [87-90].

Recentemente, algumas abordagens mais complexas têm sido propostas para a auxiliar em estudos de desenvolvimento como a genómica, proteómica, bioinformática e a metabolómica [91-93].

Deve-se, no entanto, ter em atenção o facto de que, devido à enorme diversidade de algoritmos e suas variantes, à complexidade crescente e à multiplicidade de conceitos e linguagens de programação usadas, é difícil implementar soluções coerentes, eficazes e inovadoras, perdendo-se alguma informação. Além disso, acreditamos que existe um conjunto de ferramentas padrão que ainda não estão totalmente desenvolvidas, compreendidas e exploradas [82]. Nas secções seguintes concentrar-nos-emos nas técnicas e métodos a que recorremos no presente trabalho.

3.3 Análise de agrupamentos

A análise de agrupamentos é uma técnica não supervisionada [5] que permite a formação de grupos a partir de um determinado conjunto de dados, nos quais os objetos partilham características semelhantes [79]. Por outras palavras, recorrendo a um critério de semelhança, permite reunir os objetos em grupos, de tal forma a que exista homogeneidade dentro do grupo e heterogeneidade entre os grupos formados.

Existem vários métodos de agrupamento [94] e a versão por nós adotada corresponde a uma abordagem não supervisionada em modo hierárquico, uma vez que a associação dos objetos é, em grande parte, independente de critérios impostos.

O processo de agrupamento, representado na Figura 3.2, envolve vários passos entre os quais a escolha de uma medida de semelhança entre os objetos e a adoção de uma técnica para a formação de grupos [95].

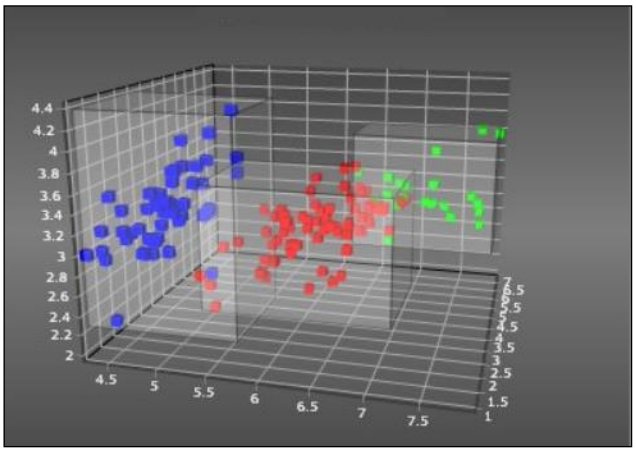


Figura 3.2 - Representação do processo de agrupamento [96].

Existe um grande número de medidas de similaridade e dissimilaridade, e a escolha deverá ser efetuada tendo em conta a natureza do problema em estudo. O passo seguinte será a adoção do método de agrupamento para a formação dos grupos. Nesta etapa, existem também vários métodos disponíveis, ficando a seleção ao critério do utilizador. Note-se que, a adoção do algoritmo requer o conhecimento das características dos diversos algoritmos à disposição, uma vez que algoritmos diferentes levam a soluções diferentes [97].

[5] Nas técnicas não-supervisionadas, não existem grupos pré-definidos; os dados são usados diretamente sem informação externa.

3.3.1 Procedimento hierárquico

A análise de agrupamento hierárquico (HCA) interliga os objetos pelas suas associações, produzindo um dendrograma onde os objetos semelhantes, segundo as variáveis escolhidas, são agrupados entre si [76]. O processo envolve uma série de agrupamentos sucessivos entre objetos. Parte-se de N grupos de apenas um objeto, que vão sendo agrupados, sucessivamente, até que se encontre apenas um grupo que incluirá a totalidade dos N objetos.

Este método conduz a uma estrutura que descreve uma hierarquia de agrupamentos sobre os dados, designada por dendrograma. Para um número inicial de *N* objetos na base de dados, ao todo ocorrem *N-1* associações. O dendrograma corresponde assim, à representação bidimensional do esquema da associação sucessiva dos objetos, atendendo à sua similaridade, até à fusão de todos os grupos num único grupo final. Este tipo de representação é especialmente útil na visualização de semelhança entre objetos representados por pontos no espaço com dimensão maior do que três, onde a representação de gráficos convencionais não é possível.

3.3.2 Semelhança e diferença

Para que d seja uma distância, e uma distância é uma medida de dissimilaridade, é necessário que as seguintes condições sejam satisfeitas, para quaisquer objetos i, j, k:

1. $\boldsymbol{d}\ (\boldsymbol{i},\boldsymbol{j}) = \boldsymbol{d}\ (\boldsymbol{j},\boldsymbol{i})\ (simétrica)$;
2. $\boldsymbol{d}\ (\boldsymbol{i},\boldsymbol{j}) > \boldsymbol{0}, se\ i\ \neq j$;
3. $\boldsymbol{d}\ (\boldsymbol{i},\boldsymbol{j}) = \boldsymbol{0}\ , se\ e\ somente\ se,\ i = j$;
4. $\boldsymbol{d}\ (\boldsymbol{i},\boldsymbol{j}) \leq \boldsymbol{d}(\boldsymbol{i},\boldsymbol{k}) + \boldsymbol{d}(\boldsymbol{z},\boldsymbol{k})(desigualdade\ triangular)$.

A propriedade (1) implica que a matriz de distâncias é simétrica em relação à diagonal. A propriedade (2) implica que todos os elementos da matriz de distâncias são positivos, a propriedade (3) implica que a sua diagonal é formada por zeros. Para que um índice de proximidade seja considerado uma métrica, este deve satisfazer, além das três propriedades anteriores, a propriedade (4) de desigualdade triangular.

O ponto de partida do processo de agrupamento corresponde normalmente à construção da matriz de distâncias, D,

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & \cdots & \cdots & d_{1n} \\ \vdots & d_{22} & & & & \vdots \\ \vdots & \vdots & \ddots & & & \vdots \\ \vdots & \vdots & & \ddots & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & \cdots & \cdots & d_{nn} \end{bmatrix} \tag{3.1}$$

que é necessário calcular e armazenar. Nesta matriz, cada elemento descreve o grau de diferença entre cada dois objetos com base nas variáveis escolhidas.

Uma das medidas de distância mais importantes é a que têm por base o coeficiente de correlação (em si uma medida de semelhança), e que pode ter, por exemplo, a forma 1-|r|, onde

$$\mathrm{r} = \frac{\mathrm{S_{xy}}}{\sqrt{\mathrm{S_{xx} * S_{yy}}}} = \frac{\sum(\mathrm{x_i} - \bar{\mathrm{x}}) .(\mathrm{y_i} - \bar{\mathrm{y}})}{\sqrt{\sum(\mathrm{x_i} - \bar{\mathrm{x}})^2 . \sum(\mathrm{y_i} - \bar{\mathrm{y}})^2}} \tag{3.2}$$

O coeficiente de correlação calcula a força e a direção de uma relação linear entre duas variáveis, e varia entre -1 e +1. Um valor próximo de +1 ou -1 indica a existência de uma forte correlação positiva e negativa respetivamente, entre as duas variáveis. Se o valor for próximo de 0 significa que não existe qualquer correlação.

Existem várias medidas que podem ser utilizadas como medidas de distâncias ou dissimilaridade entre elementos *i* e *j* da matriz de dados. A mais utilizada é a distância euclidiana, *d*, (ou o seu quadrado),

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|^2} \tag{3.3}$$

Esta distância não possui um valor limite, mas é sempre positiva ou nula. A determinação de valores de pequenas dimensões é indicativa da existência de uma forte semelhança entre as variáveis [77].

Outras distâncias de uso comum são a distância de Manhattan,

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}| \tag{3.4}$$

e a distância de Minkowski,

$$d(i,j) = \sqrt[q]{|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \cdots + |x_{in} - x_{jn}|^q} \tag{3.5}$$

onde q ≥ 1. A distância de Minkowski abrange tanto a distância euclidiana (q=2) como a distância de Manhattan (q=1).

Uma vez calculadas as distâncias, diferentes critérios podem ser utilizados para estabelecer a ligação entre objetos. A abordagem mais comum é o chamado agrupamento associativo, no qual os objetos individuais são gradualmente ligados uns aos outros em grupos, dos critérios que a seguir se descrevem [95]. A ligação é, basicamente, a estratégia para se estabelecerem distâncias objeto-grupo ou grupo-grupo.

3.3.3 Critério de ligação

Método da ligação simples (*single linkage*)

A ligação simples, cujo procedimento se encontra esquematizado na Figura 3.3**,** é um dos métodos mais simples e de rápida aplicação. A distância entre dois grupos é definida como sendo aquela entre os objetos mais próximos desses dois grupos. Assim, a semelhança será tanto maior quanto menor a distância entre os pontos. Conduz genericamente à formação de grupos de maiores dimensões do que os estabelecidos por outros critérios. Os dendrogramas resultantes deste critério são, geralmente, pouco elucidativos, dado que, a informação relativa aos objetos intermediários não é evidente.

Este método tende a formar longas cadeias[6], é sensível a *outliers*[7], pois tem tendência a incorporar os *outliers* num grupo já existente, e grupos muito próximos podem não ser identificados [95].

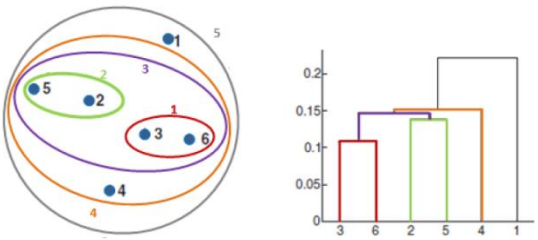


Figura 3.3 - Distância entre grupos calculada através da menor distância (*single linkage*). À esquerda a associação e à direita apresenta-se o respetivo dendrograma [76].

[6] Situação em que há um primeiro grupo de um ou mais objetos que passa a incorporar um grupo de apenas um objeto, formando uma longa cadeia, onde se torna difícil definir um nível de corte para classificar os objetos em grupos

[7] Em estatística, um *outlier*, ou valor atípico, é uma observação que apresenta uma grande discrepância relativamente a outras observações da mesma população.

Método da ligação completa (*complete linkage*)

Ao contrário do anterior, o método da ligação completa, representado na Figura 3.4, determina a distância entre dois grupos de acordo com maior distância entre um par de objetos, sendo cada objeto pertencente a um grupo distinto. Geralmente, leva a grupos compactos e discretos, sendo os seus valores de dissimilaridade relativamente grandes [95].

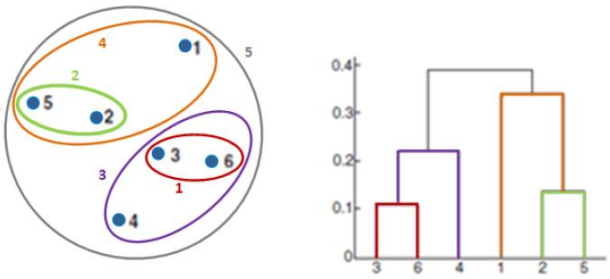


Figura 3.4 – Distância entre grupos através da associação completa (*complete linkage*). À esquerda está representado a associação e à direita o respetivo dendrograma [76].

Método da ligação média (*average linkage*)

Neste método, representado na Figura 3.5, a distância entre dois grupos é definida como a média das distâncias entre os pares de objetos em cada grupo [95].

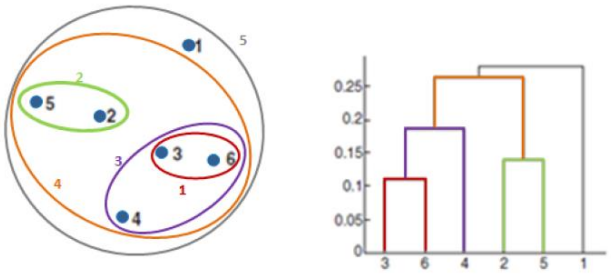


Figura 3.5 – Distância obtida através da média das distâncias entre os objetos (*average linkage*). À esquerda representado a associação e à direita o respetivo dendrograma [76].

Método de variância mínima (*Ward*)

O método da ligação de *Ward*, esquematizado na Figura 3.6, baseia-se na análise de variância, associando os objetos aos grupos nos quais promovem a menor variância intra-grupo. Este algoritmo é altamente eficiente na formação de grupos.

Inicialmente, admite que cada um dos objetos constitui um único grupo. Considerando a primeira reunião de objetos num novo grupo, a soma dos desvios dos pontos representativos dos seus elementos, em relação à média do grupo, é calculada, e dá uma indicação de homogeneidade

do grupo formado. Os grupos formados possuem uma elevada homogeneidade interna. No entanto, pode apresentar resultados insatisfatórios quando o número de elementos em cada grupo é praticamente igual, tem tendência a combinar grupos com poucos elementos e é sensível à presença de *outliers* [95].

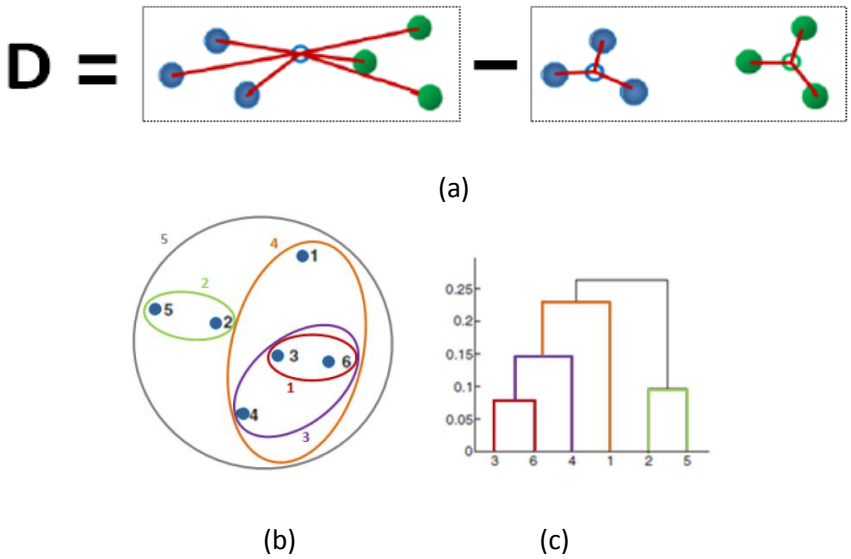


Figura 3.6 – Distância (a) obtida através do método da variância mínima (*Ward linkage*); (b) esquema do processo de associação, (c) dendrograma resultante do processo de associação [76].

3.3.4 Critérios para determinar o número de grupos - abordagem hierárquica

Determinar o número de grupos presentes num conjunto de dados é uma das tarefas mais difíceis no processo de agrupamento. No caso de não existir conhecimento prévio sobre o número de grupos em que a população em estudo é dividida, um dos métodos mais utilizados consiste na comparação gráfica do número de grupos com o respetivo coeficiente de fusão, isto é, o valor numérico (semelhança ou distância) para o qual vários objetos se unem para formar um grupo. Assim, quando a divisão de um novo grupo não introduz alterações significativas no coeficiente de fusão, considera-se essa partição como sendo a mais adequada [98]. Outro procedimento utilizado consiste na comparação dos resultados obtidos por vários métodos diferentes de agrupamento. Poder-se-á aferir o grau de convergência entre os vários métodos de agrupamento através de uma tabela de contingência, indicando o número de observações que se agrupam no mesmo conjunto, para o mesmo passo de associação. Desta forma é possível verificar a maior ou menor estabilidade das soluções encontradas e avaliar a qualidade do agrupamento efetuado. Na literatura existem

alguns textos de revisão sobre critérios para a escolha do número de grupos em procedimentos hierárquicos e não hierárquicos, com abordagens mais subjetivas que objetivas, (veja-se por exemplo [99,100]). O critério usado no presente estudo foi desenvolvido por alguns autores [101] e é complementar à estrutura do dendrograma. O algoritmo tem por base um processo de redução de *outliers*, seguido da construção de uma função descritiva, DF que permite a identificação dos grupos naturais (mais detalhes em [101]). A função $DF_{i,i+1}$, corresponde ao quadrado da distância mínima resultante de todas as etapas de ligação, em que ambos os objectos *i* e *i+1*, participam,

$$DF_{i,i+1} = d_{i,i+1}^2 \tag{3.6}$$

A função matemática dada pela $DF_{i,i+1}$, para cada par de objetos sequenciais, no vetor de associação, produz regiões com máximos localizados, correspondentes a uma elevada probabilidade de separação entre-grupos, e regiões de baixo valor indicando uma elevada probabilidade de associação intra-grupo. O quadrado da distância permite evidenciar as separações entre-grupos. Note-se que, a função apresentada não é a única possível, mas constitui uma das soluções mais simples e que pode ser usada para produzir resultados de qualidade comparáveis a outras abordagens mais sofisticadas.

3.4 Análise de componentes principais

A análise de componentes principais (PCA) é uma das técnicas mais antigas e conhecida da análise multivariada [77]. É recomendada como uma ferramenta exploratória para encontrar tendências em dados desconhecidos. Trata-se de um método não paramétrico, de extração de informações relevantes a partir de conjuntos de dados multivariados [102-103].

O principal objetivo do PCA consiste na redução da dimensionalidade de grandes matrizes de dados – as *m* variáveis originais são substituídas por um outro subconjunto de *p* variáveis não correlacionadas, de menor dimensão, designadas de componentes principais (PC`s), com a menor perda de informação possível.

Uma das vantagens desta técnica, para além da redução da dimensionalidade é o facto de que as novas variáveis, as componentes principais, não são correlacionadas e, em vez de se analisar um elevado número de variáveis originais com uma estrutura inter-relacional complexa, analisam-se apenas algumas variáveis não correlacionadas [104]. Assim, com este método, é possível efetuar uma simplificação e redução da dimensão original dos dados, modelação, deteção de *outliers*, seleção de variáveis importantes num determinado sistema, classificação e previsão [104].

De entre todas as possíveis combinações lineares, escolhe-se, em cada caso, a de variância máxima, dado que as componentes principais devem refletir, tanto quanto possível, as características dos dados, devendo explicar uma grande parte da variação associada às variáveis iniciais [104]. A representação destas variáveis segundo a primeira componente principal pode ser observada no esquema patente na Figura 3.7.

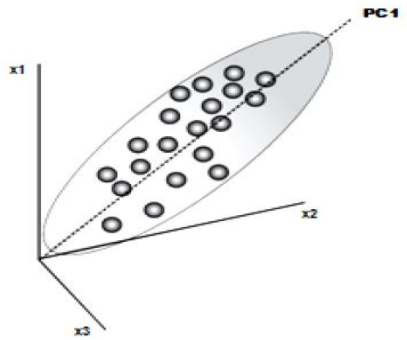


Figura 3.7 – Representação da primeira componente principal que justifica a maior variabilidade dos dados (adaptada de [105]).

As componentes principais são, portanto, combinações lineares das *p* variáveis da matriz X

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p \tag{3.7}$$

onde $j = 1, \ldots, p$ e $a_{ij}(i = 1, \ldots, p; j = 1, \ldots p)$ são constantes. As variáveis X_i resultam, quase sempre, de um processo prévio de centragem, com base na média, das variáveis originais.

Os coeficientes destas combinações lineares são determinados de modo a satisfazerem as condições seguintes:

1. $Var\ (Y_1) \geq Var\ (Y_2) \geq \cdots \geq Var\ (Y_p)$
2. Quaisquer duas componentes principais não são correlacionadas, $Corr\left(Y_i, Y_j\right) = 0, \forall_{ij}$.
3. Em qualquer componente principal a soma dos quadrados dos coeficientes que engloba é igual a 1 (para Y_i: $a_{1j}^2 + a_{2j}^2 + \cdots + a_{pj}^2 = 1$).

Das condições anteriores retiramos que Y_1 é a componente com maior variância, Y_2 é a componente principal com a segunda maior variância, sujeita à condição de ser não correlacionada com Y_1, Y_3 é a componente principal com a terceira maior variância, sujeita à condição de ser não correlacionada com Y_1 e com Y_2 (e assim por diante).

Na Equação (3.7), $a_1, a_2, \ldots, a_p$ são respetivamente, os *p* vetores próprios associados aos *p* maiores valores próprios de Σ= $(\lambda_1 > \lambda_2 > \cdots > \lambda_p)$, com $Var(Y_j) = \lambda_j$.

A covariância entre cada duas componentes principais Y_i *e* Y_j é, como se disse, imposta como nula, pois todas as componentes foram determinadas de forma a serem não correlacionadas. Tem-se então que $Cov\left(Y_j, Y_{j'}\right) = a_j^{'}\Sigma a_{j'} = a_j^{'}\lambda_j a_{j'} = \lambda_j a_j^{'} a_{j'} = 0$, que equivale a ter $a_j^{'} a_{j'} = 0$, o que indica que $a_j^{'} a_{j'} = 0$, o que indica que $a_j^{'}$ e $a_{j'}$ (com $j \neq j^{'}$) são vetores ortogonais.

Muitas vezes, as variáveis em estudo não são todas medidas na mesma unidade, na mesma escala, ou são de ordem de grandeza distinta. Surge, assim, a necessidade de estabelecer uma certa uniformização, que se consegue através da divisão de cada valor pelo desvio padrão da variável centrada correspondente. Este procedimento conduz à obtenção de variáveis com valor médio nulo e variância unitária. As variáveis em estudo passam a ter todas a mesma variância e a influência das variáveis de variância pequena tende a ser inflacionada enquanto a influência das variáveis de variância elevada tende a ser reduzida.

A matriz de covariância do conjunto destas "novas" variáveis é igual à matriz de correlação do conjunto de variáveis iniciais, dado que:

$$Cov\left(\frac{X_i}{\sigma_i}, \frac{X_j}{\sigma_j}\right) = \frac{Cov\,(X_i, X_j)}{\sigma_i \sigma_j} = Corr\left(X_i, X_j\right) \tag{3.8}$$

Assim, a análise de componentes principais de um conjunto de dados deste tipo, é efetuada utilizando a matriz de correlação, *P*. As componentes principais serão determinadas tendo em conta os valores e vetores próprios da matriz *P.* Matematicamente tudo se processa da mesma forma. No entanto, os vetores próprios de *P* não são iguais aos de Σ, e as componentes principais também não serão as mesmas.

A matriz *P* define-se como

$$P = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \tag{3.9}$$

em que

$$\rho_{ij} = Corr\,(X_i, X_j) \tag{3.10}$$

3.4.1 Redução da dimensionalidade

A redução da dimensionalidade é conseguida, considerando apenas algumas das componentes principais, isto é, as de maior variância. Dado que, as componentes principais se

podem ordenar por ordem decrescente da sua variância e que quanto maior esta for mais representativa dos dados originais será a correspondente componente principal, podemos estabelecer quais as componentes relevantes.

Assim, a soma das variâncias das componentes principais é dada por

$$\sum_{j=1}^{p} Var\left(Y_j\right) = \sum_{j=1}^{p} \lambda_j \tag{3.11}$$

Além disso, como se sabe, numa matriz simétrica (que é o caso de Σ) a soma dos seus valores próprios é igual ao traço da matriz, pelo que

$$tr\left(\Sigma\right) = \sum_{j=1}^{p} Var(X_j) \Rightarrow \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} Var\left(X_j\right) \tag{3.12}$$

de onde

$$\sum_{j=1}^{p} Var\left(Y_j\right) = \sum_{j=1}^{p} Var\left(X_j\right) \tag{3.13}$$

Isto significa que a soma das variâncias das variáveis originais é igual à soma das variâncias das componentes principais (se considerarmos todas as componentes principais explicamos toda a variabilidade). Assim, a proporção da variância total que é explicada pela *j*-ésima componente principal Y_j e que indica a importância da mesma é dada por

$$\frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} = \frac{\lambda_j}{tr\left(\Sigma\right)} \tag{3.14}$$

Existem vários critérios que podem ser usados para a escolha do número de componentes principais. Os mais conhecidos são o (i) Critério de Pearson, (ii) o Critério de Kaiser e (iii) o *Scree Plot*.

(i) Critério de Pearson (ou regra dos 80%)

Este critério é utilizado quando se recorre à matriz de covariância. O número de componentes principais é escolhido até recuperarmos mais de 80% da informação total ou variabilidade total. Por outras palavras, devem considerar-se tantas componentes principais quantas

as necessárias para que a percentagem de variância por elas explicada seja superior a 80%. Tal consiste em reter as primeiras *r* componentes principais de modo a que [106] se atinja

$$\sum_{j=1}^{r} \frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} = \frac{\sum_{j=1}^{r} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \geq 0.80 \tag{3.15}$$

(ii) Critério de Kaiser ($\lambda > 1$)

O critério de Kaiser é utilizado com a matriz de correlação, embora o critério anteriormente descrito também seja uma possibilidade nesta opção. Segundo este critério, devem ser consideradas apenas as componentes com valor próprio superior à unidade (note-se que este valor unitário é média sobre o conjunto de valores próprios) [106].

$$\bar{\lambda} = \frac{1}{p} \sum_{j=1}^{p} \lambda_j \tag{3.16}$$

(iii) *Scree plot*

Este terceiro critério permite utilizar um gráfico onde se representam os pontos de abcissa *j* e ordenada igual ao *j*-ésimo valor próprio ou à percentagem de variância explicada pela *j*-ésima componente principal, isto é pontos de coordenadas (j, λ_j) ou $(j, \lambda_j / \sum_{j=1}^{p} \lambda_j)$, onde se distinguem as contribuições das diversas componentes principais. De acordo com este critério, devem-se considerar as *r* componentes principais que mais contribuem, destacando-se de forma acentuadas das restantes [104].

3.4.2 *Scores* e *loadings*

As coordenadas dos objetos no novo sistema de referência são designadas por *score*s, enquanto o coeficiente da combinação linear descreve cada PC, isto é, os pesos das variáveis originais em cada PC, são denominados por *loadings*.

Nesta altura, sabemos que as componentes principais resultam de uma transformação sobre as variáveis em estudo (combinação linear). Podemos agora pensar em aplicar a mesma transformação aos dados, ou seja, aos vetores de observações $x_1, x_2, \cdots, x_p$ (colunas da matriz de dados *X*) das variáveis $X_1, X_2, \cdots, X_p$, respetivamente.

Obtemos uma nova matriz de dados, a matriz *Y,* com dimensão $(n \times r)$ em que o *ij*-ésimo elemento será igual ao *score* do *i*-ésimo objeto para a *j*-ésima componente principal

$$y_{ij} = a_{1j}x_{i1} + a_{2j}x_{i2} + \cdots + a_{pj}x_{ip} \tag{3.17}$$

A matriz dos *scores* dos objectos é dada por

$$Y = \begin{bmatrix} y_{11} & \cdots & y_{1r} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nr} \end{bmatrix} \tag{3.18}$$

3.4.3 Representações gráficas

As representações gráficas são um ótimo auxiliar na interpretação dos resultados do PCA. Neste trabalho, recorre-se ao *software* R versão 2.15.2 [73] e ao RStudio para efetuar estas representações.

Representação das variáveis *(loadings)*

Na representação gráfica das *loadings*, cada ponto representa uma variável e o plano é definido por dois (ou três) eixos correspondentes aos primeiros componentes principais. A cada variável é associado um ponto, cujas coordenadas são os coeficientes da transformação.

Representação dos objetos no novo sistema de eixos *(scores)*

A representação dos objetos é feita com base nos chamados *scores*, isto é, as suas coordenadas no novo sistema de eixos.

3.5 Algoritmo *convex hull*

O algoritmo *convex hull*, determina a fronteira convexa de um conjunto finito de pontos no plano, tratando-se de um dos mais antigos problemas considerados na definição de geometria computacional. Tem como uma das principais aplicações o reconhecimento de padrões [107].

O *convex hull* de um conjunto N de pontos é intuitivamente fácil de descrever. Um conjunto *N* de pontos do plano é convexo, se para quaisquer pontos de *N*, os segmentos entre estes,

estiverem totalmente contidos em *N*. Na Figura 3.8 encontram-se representados dois exemplos de conjuntos convexos, a duas (a) e a três dimensões (b).

No presente trabalho o algoritmo *convex hull* permite inspecionar a evolução temporal relativa ao número de detenções no período considerado, pela observação dos limites dos grupos formados pelos estados americanos. Deste modo, os limites dos conjuntos correspondentes a cada ano são definidos pelos estados mais afastados da origem.

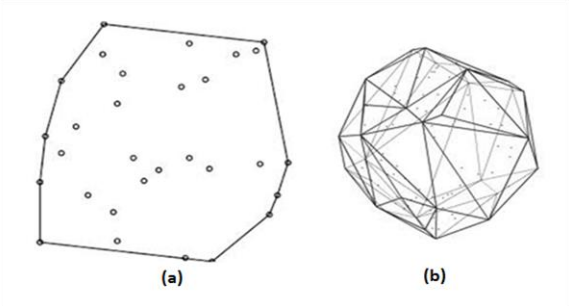


Figura 3.8 – Exemplo de dois conjuntos convexos a duas (a) e a três (b) dimensões (adaptado de [107]).

3.6 Métodos econométricos implementados

Neste trabalho, são usadas como ferramentas complementares no tratamento dos dados UCR, o coeficiente de Gini e a Curva de Lorenz.

3.6.1 Coeficiente de Gini

A caracterização de indicadores de desigualdade de uma população levanta algumas questões relativas à mensuração e quantificação da desigualdade existente numa sociedade e quais os problemas que surgem nessa mensuração. Estabelecer e compreender os indicadores de avaliação da desigualdade tem sido objeto de estudo em diversas áreas [108].

O coeficiente de Gini é um dos principais índices de desigualdade. Trata-se de um indicador desenvolvido pelo estatístico italiano Corrado Gini, publicado no documento "*Variabilitá e Mutabilitá*" em 1912 [109]. Este índice é comumente utilizado para calcular a desigualdade de uma distribuição de renda, mas pode ser também aplicado a qualquer distribuição. O Coeficiente compreende valores entre 0 (completa igualdade) e 1 (completa desigualdade) [109]. Está associado a uma medida de desigualdade calculada por meio de um *ratio*, ao invés de uma variável representativa da maioria da população, tais como, crime (ou renda) *per capita* ou ainda o produto

interno bruto. Este índice pode ser usado também para comparar as distribuições de crime entre diferentes setores da população, tais como as zonas urbanas e rurais. É um índice suficientemente simples e facilmente interpretável, principalmente quando são feitas comparações entre países. Por ser simples, permite também uma comparação da desigualdade entre populações através do tempo.

A construção do coeficiente de Gini é baseada na "Curva de Lorenz", descrita na secção seguinte.

3.6.2 Curva de Lorenz

A Curva de Lorenz (ou curva de concentração de Lorenz), representada na Figura 3.9, consiste num gráfico muito utilizado pelos economistas e que procura ilustrar a desigualdade existente na distribuição do rendimento entre as famílias numa determinada economia ou sociedade [110]. Este gráfico consiste num diagrama em que num dos eixos é colocada a variável Rendimento e no outro a População, geralmente representados por classes percentuais.

No contexto deste trabalho, esta representação é utilizada para caracterizar a distribuição do crime na população dos estados americanos. Neste diagrama é então representada uma linha representativa da percentagem de crime associado a cada estado, o que permite fazer uma leitura do tipo: "os x% dos estados com menor número de detenções retêm y% do total de crime". Quanto mais afastada da diagonal estiver esta linha, maior é a concentração do crime, ou seja, maior será a desigualdade na repartição do crime entre os estados americanos.

A curva de Lorenz pode ser complementada com o Índice de Gini, que quantifica o grau de concentração dos crimes considerados.

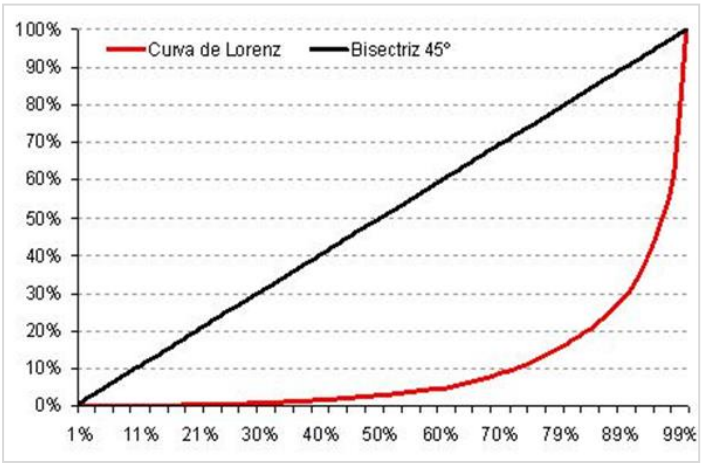


Figura 3.9 – Representação gráfica da Curva de Lorenz (adaptada de [111]).

Tanto o Coeficiente de Gini como a Curva de Lorenz são duas medidas econométricas, que neste trabalho são utilizadas num contexto diferente do habitual. As duas medidas complementares são adaptadas para distribuições de crime nos Estados Unidos, tendo em conta o número de detenções efetuadas em quatro anos (2005, 2007, 2009 e 2011).

Capítulo 4

As detenções – uma perspetiva temporal

Nesta secção é feita a caracterização global e temporal dos dados correspondentes ao número de detenções efetuadas nos EUA em 2005, 2007, 2009 e 2011, para 29 tipos de crime constantes da tabela 69 (*Arrest, by State*) do FBI UCR. Devido à ausência de informação relativa a alguns tipos de crime, *District of Columbia* e *Hawaii* não são incluídos no estudo. Os métodos quimiométricos escolhidos são a Análise de Agrupamento Hierárquico (HCA) e a Análise de Componentes Principais (PCA).

4.1 Padrões estruturais

4.1.1 Número de detenções

Numa primeira fase, a análise não será baseada em quaisquer pressupostos. Como tal, consideramos cada estado como um objeto, caracterizado pelo número total de detenções relativas a cada tipo de crime, organizados sob a forma de um vetor. O conjunto de dados é então formado por 49 estados (ao invés dos 51 constantes nas tabelas da base de dados UCR), caracterizados por um total de 29 crimes.

Após o pré-processamento do conjunto de dados inicial para eliminar os efeitos da presença de *missing values*, por substituição com o valor médio global da respetiva variável, a análise hierárquica de agrupamentos, fornece um meio visual para estimar as relações entre os dados. O método de ligação *Ward*, descrito em detalhe no capítulo 3, é considerado neste trabalho como o método padrão de ligação.

As estruturas definidas nos dendrogramas e os mapas geográficos contendo os grupos formados são observadas conjuntamente de forma a tornar mais clara e compreensível a variação da estrutura dos grupos ao longo do tempo.

Na Figura 4.1 estão representados os dendrogramas relativos a 2005 e 2007, bem com a representação dos padrões geográficos obtidos através da análise de agrupamento hierárquico. A mesma observação para os anos de 2009 e 2011 é possível na Figura 4.2.

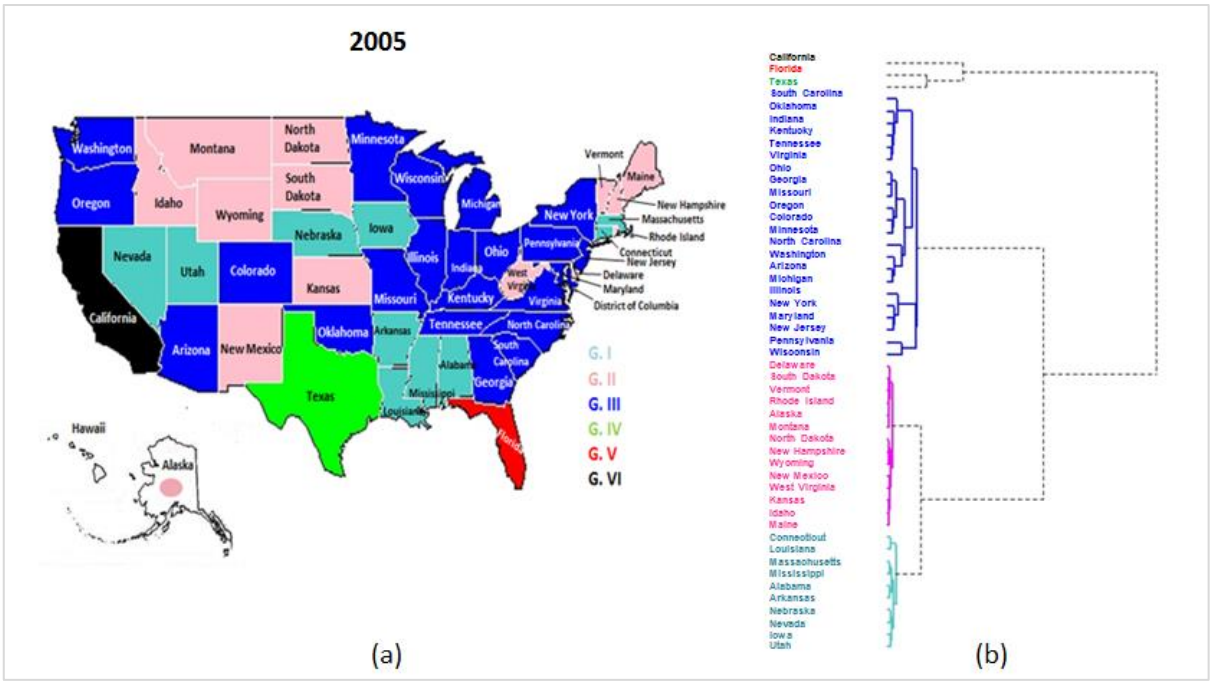
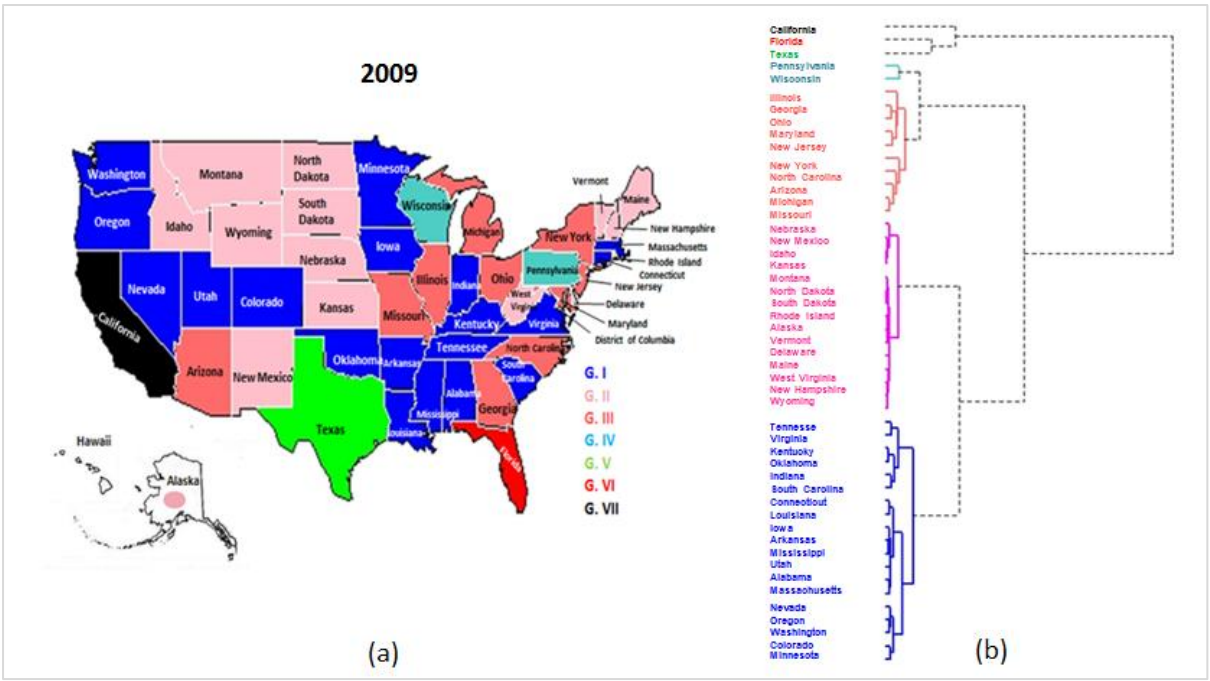


Figura 4.1 – Dendrogramas construídos por HCA (método de ligação *Ward*) sobre o conjunto de dados correspondentes a 49 estados dos EUA, de acordo com a tabela 69 do FBI UCR para o ano de 2005 (b) e 2007 (d). Para o estabelecimento dos grupos foi usada uma linha de corte definida pela função descrita na secção 3.3.4, da qual resultam 6 grupos. Na representação dos padrões geográficos obtidos por HCA para o ano de 2005 (a) 2007 (c) são usadas cores distintas para representar os grupos de estados, de acordo com as estruturas definidas em (b) e (d), respetivamente.



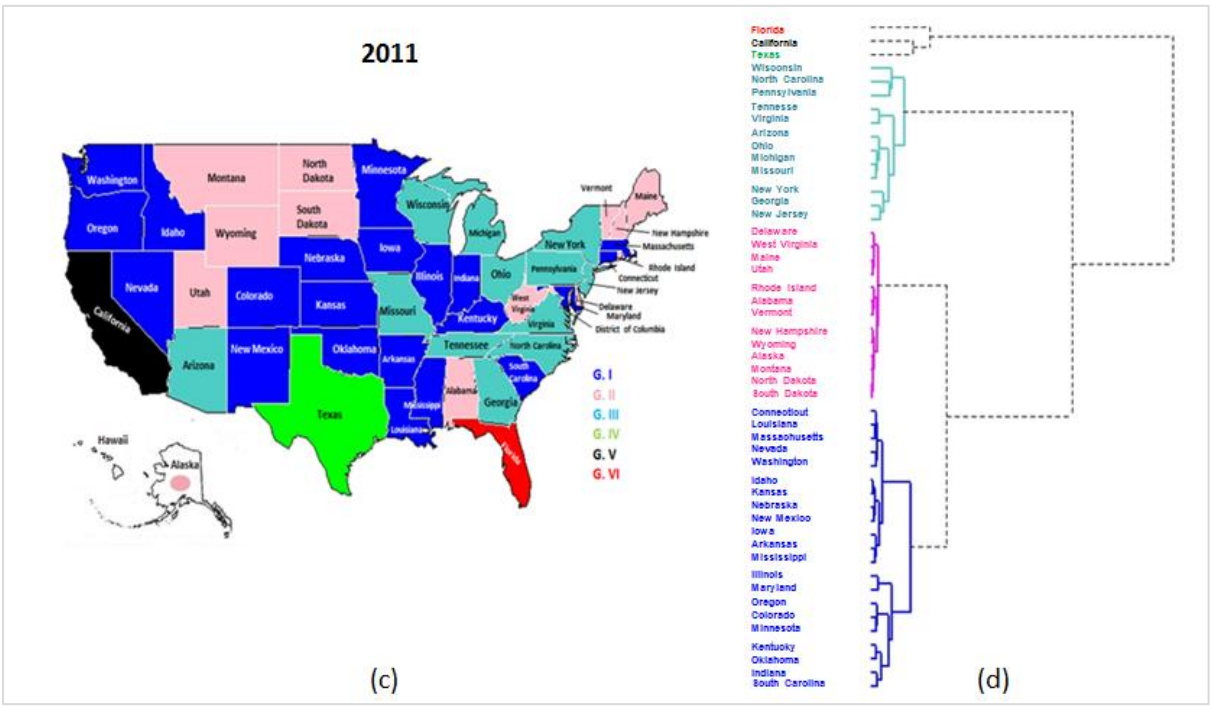


Figura 4.2 – Representação dos padrões geográficos relativos a 2009 e 2011, seguindo o mesmo esquema da Figura 4.1. De acordo com o critério definido na secção 3.3.4 são estabelecidos 7 grupos em 2009, (a) e (b), e 6 grupos em 2011, (c) e (d).

Através da análise dos dendrogramas para os quatro anos em estudo (2005, 2007, 2009 e 2011), é possível observar seis grupos, com exceção do dendrograma relativo ao ano de 2009 no qual são estabelecidos sete grupos. Nas estruturas encontradas, existem três grandes grupos e três estados isolados (*California*, *Florida* e *Texas*). Estes últimos apresentam diferenças significativas entre si e diferenças relativamente aos restantes grupos, surgindo, portanto, em três grupos distintos (grupos IV, V e VI em 2005, 2007 e 2011, respetivamente e grupos V, VI e VII para o ano de 2009). Dos três grupos de maior dimensão, um é constituído em média por 20 estados (grupo III em 2005 e 2007 e grupo I em 2009 e 2011), os restantes dois grupos são formados aproximadamente por 10 a 15 estados (grupos I e II para 2005, grupo I para 2007 e grupos II e III para o ano de 2009 e 2011). Nos dendrogramas referentes a 2007 e 2009 surge um grupo comum, constituído pelos estados da *Pennsylvania* e *Wisconsin*, correspondendo respetivamente aos grupos II e IV.

Definida a estrutura dos grupos, e para permitir uma análise preliminar caracterizadora, inspecionemos os padrões geográficos. Nos mapas das Figuras 4.1 e 4.2 é possível observar que estados vizinhos estão agrupados e por isso partilham características comuns. Por exemplo, de forma geral, é visível nos quatro mapas uma zona central (a rosa, formada por estados pertencentes à região Oeste e Centro-Oeste). Este grupo engloba 5 estados (*Idaho*, *Montana*, *Wyoming*, *North Dakota* e *South Dakota)* em 2005. No entanto, nos anos seguintes, a estrutura do grupo considerado sofre alteração devido à evolução de alguns estados para grupos vizinhos. Dispersos no mapa encontram-se outros estados pertencentes a este grupo, principalmente a Nordeste, correspondendo ao estados de *West Virginia*, *Vermont*, *Maine*, *Rhode Island*, *Delaware* e *New Hampshire*, a Este os estados de *New Mexico, Utah* e *Nevada*, a Sul do mapa e mais notório para o ano de 2007, os estados de *Arkansas*, *Louisiana*, *Mississippi* e *Alabama*. Um outro grupo, a azul, é constituído por estados maioritariamente da região Centro-Oeste e Sul. Ao longo do tempo, os estados constituintes deste grupo tendem a dispersar para outros grupos, como se pode verificar pelo padrão evidenciado nos mapas correspondentes a 2009 e 2011. Note-se que nos mapas, este padrão forma um "arco" em redor da zona central (a rosa).

As Tabelas 4.1 e 4.2 reúnem os grupos formados através da análise hierárquica de agrupamentos referentes aos dois extremos temporais, referentes aos anos de 2005 e 2011. Nas tabelas estão discriminados seis grupos explicitamente numerados e as cores seguem o padrão resultante da análise hierárquica.

Tabela 4.1 – Grupos formados através da análise hierárquica de agrupamentos para o ano de 2005. Cada estado é caracterizado pelo número total de cada tipo de crime, com base na ligação de *Ward.* Para o estabelecimento dos grupos foi utilizado uma linha de corte adequada definida na secção 3.3.4.

Grupo VI	Grupo V	Grupo IV	Grupo III	Grupo II	Grupo I
California	Florida	Texas	Arizona	Alaska	Alabama
			Colorado	Delaware	Arkansas
			Georgia	Idaho	Connecticut
			Illinois	Kansas	Iowa
			Indiana	Maine	Louisiana
			Kentucky	Montana	Massachusetts
			Maryland	New Hampshire	Mississippi
			Michigan	New Mexico	Nebraska
			Minnesota	North Dakota	Nevada
			Missouri	Rhode Island	Utah
			New Jersey	South Dakota	
			New York	Vermont	
			North Carolina	West Virginia	
			Ohio	Wyoming	
			Oklahoma		
			Oregon		
			Pennsylvania		
			South Carolina		
			Tennessee		
			Virginia		
			Washington		
			Wisconsin		

Tabela 4.2 – Grupos formados através da análise hierárquica de agrupamentos para o ano de 2011. Cada estado é caracterizado pelo número total de cada tipo de crime, com base na ligação de *Ward*. Para o estabelecimento dos grupos foi utilizado uma linha de corte adequada definida na secção 3.3.4.

Grupo VI	Grupo V	Grupo IV	Grupo III	Grupo II	Grupo I
Florida	California	Texas	Arizona	Alabama	Arkansas
			Georgia	Alaska	Colorado
			Michigan	Delaware	Connecticut
			Missouri	Maine	Idaho
			New Jersey	Montana	Illinois
			New York	New Hampshire	Indiana
			North Carolina	North Dakota	Iowa
			Ohio	Rhode Island	Kansas
			Pennsylvania	South Dakota	Kentucky
			Tennessee	Utah	Louisiana
			Virginia	Vermont	Maryland
			Wisconsin	West Virginia	Massachusetts
				Wyoming	Minnesota
					Mississippi
					Nebraska
					Nevada
					New Mexico
					Oklahoma
					Oregon
					South Carolina
					Washington

Para finalizar a análise preliminar deste perfil referente às detenções nos quatro anos considerados, e que será refinada em secções subsequentes, é possível adiantar uma hipótese justificativa para os padrões encontrados. De facto é patente para os vários anos em estudo, uma zona central caracterizada por estados com populações na ordem de um milhão de habitantes, registando um número de crimes proporcional a esta população e que se considera como uma baixa criminalidade total. O avanço dessa zona para a periferia corresponde, genericamente, a uma transição de um grupo de criminalidade intermédia (região a azul nos mapas) correspondendo aos grupos I e III para o ano de 2005, para áreas de criminalidade média alta, a Leste, ou alta, a Sul e Sudoeste. Este padrão geral dos mapas para os quatro anos em estudo está representado na Figura 4.3 e é, em grande medida, o resultado do aumento da população estadual no mesmo sentido. A proporcionalidade já referida entre o crime e a população traduz o padrão criminal num padrão claramente geográfico. De facto, os estados centrais de menor população estão rodeados por estados de população crescente.

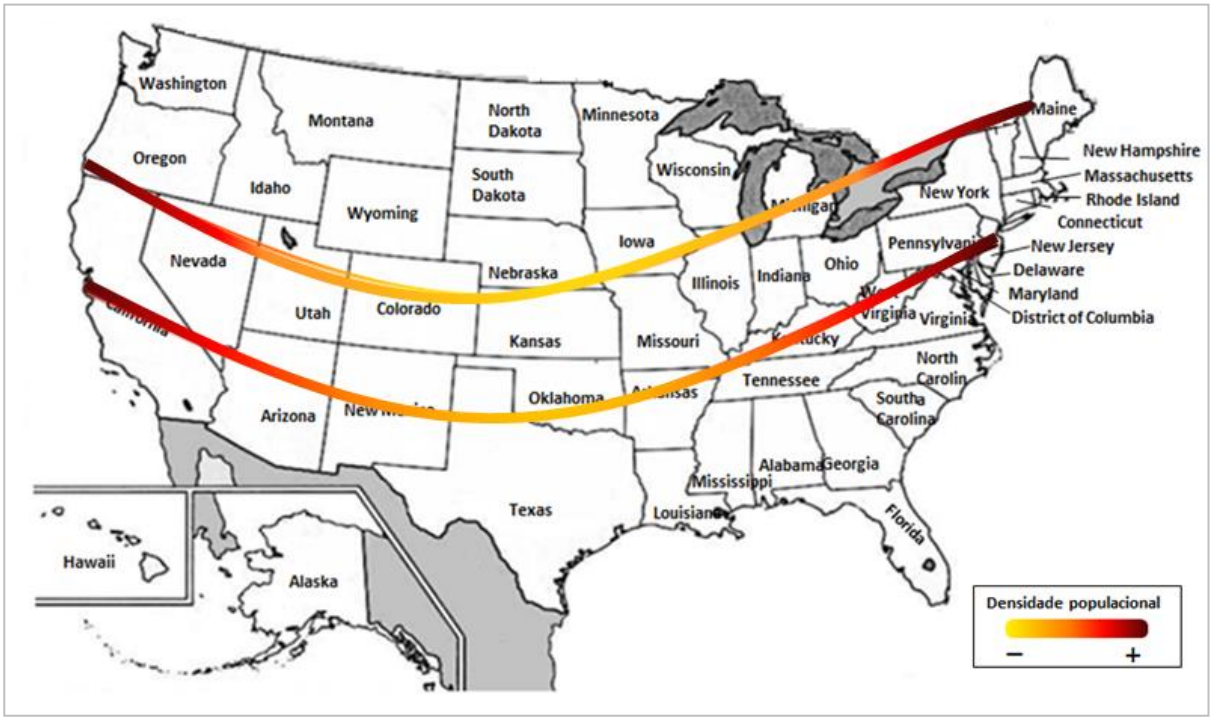
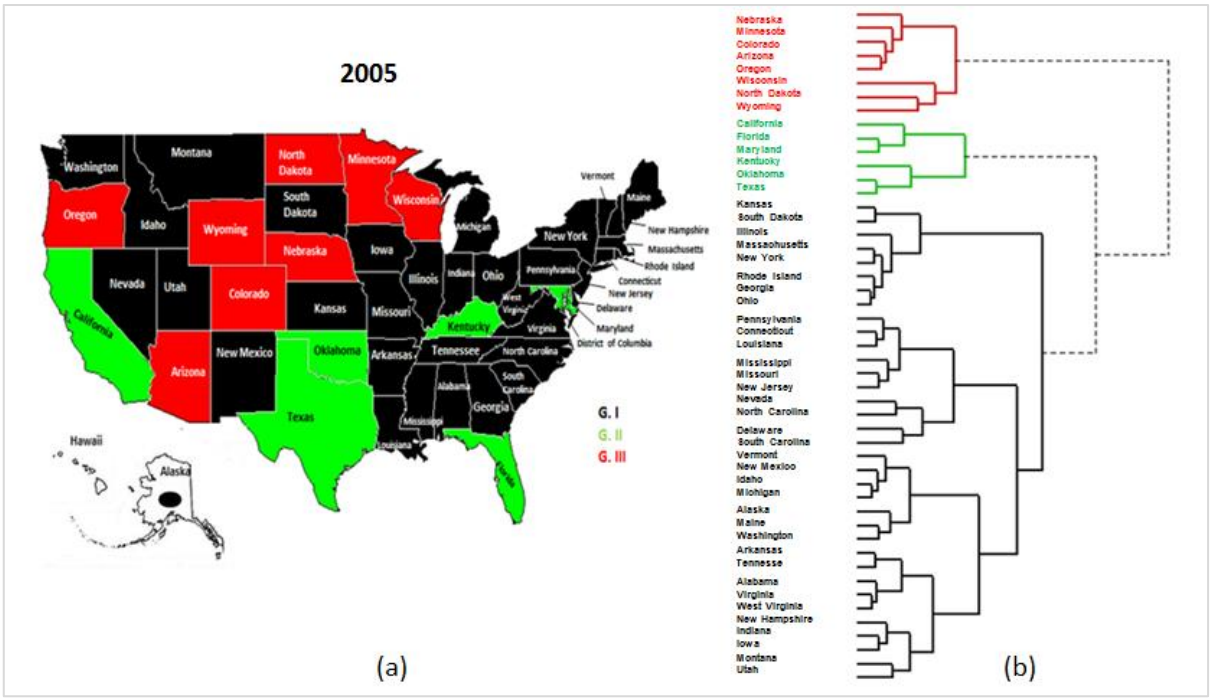


Figura 4.3 – Esquema ilustrativo do perfil geral patente nos mapas, para os dados correspondentes ao número de crimes referentes aos quatro anos em estudo (2005, 2007, 2009 e 2011). O padrão geográfico criminal nos mapas traduz a proporcionalidade entre o crime e a população.

4.1.2 Taxa de criminalidade

Prossiga-se agora para uma análise semelhante à realizada na secção anterior, mas na qual cada estado é agora definido por um vetor no qual a frequência de cada tipo de crime é dividida pela população total de cada estado. Quanto ao procedimento geral, será idêntico ao anterior.

Na Figura 4.4 estão representados os dendrogramas correspondentes a 2005 e 2007, bem como a representação dos padrões geográficos obtidos através da análise hierárquica para os anos considerados. O mesmo esquema é apresentado na Figura 4.5 para 2009 e 2011. Registam-se alterações significativas relativamente às estruturas definidas com base apenas na frequência de crimes. Podemos agora identificar apenas três grupos em 2005 e 2009, descritos nas Tabelas 4.3 e 4.4. O número de elementos de cada grupo varia entre seis e trinta e cinco em 2005 e entre um e vinte e sete em 2011.



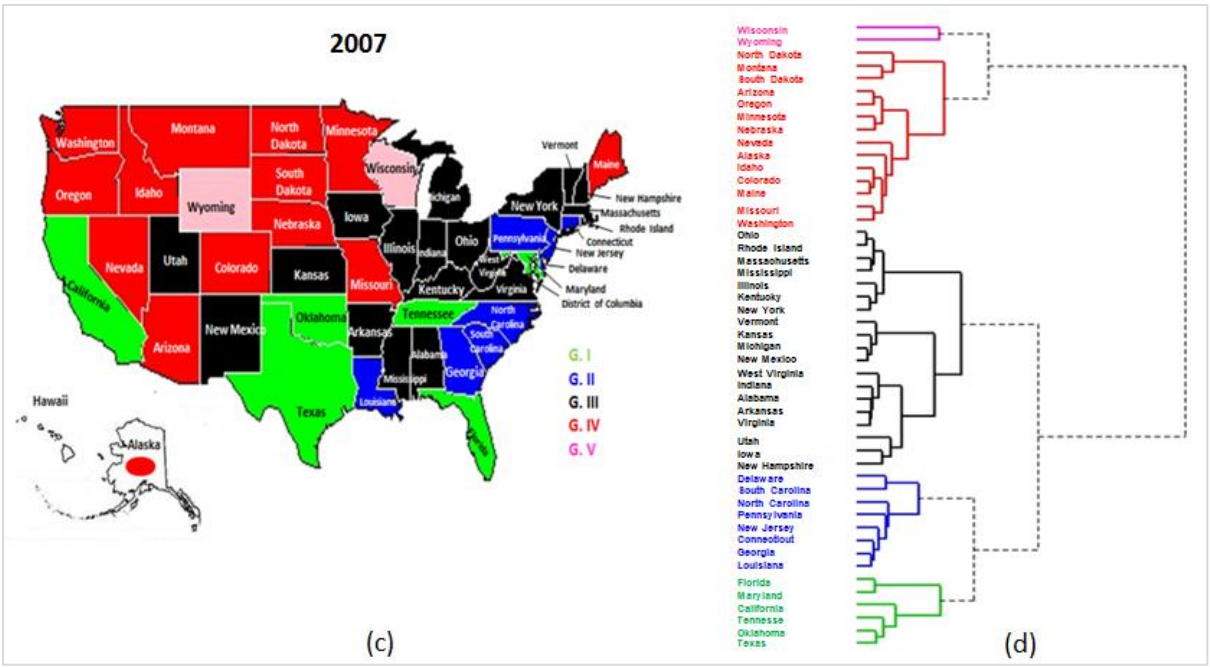


Figura 4.4 – Dendrogramas construídos por HCA (método de ligação *Ward*) sobre o conjunto de dados correspondentes a 49 estados dos EUA, de acordo com a tabela 69 do FBI UCR para o ano de 2005 (b) e 2007 (d). Para o estabelecimento dos grupos foi usada uma linha de corte definida pela função descrita na secção 3.3.4, da qual resultam 3 e 5 grupos, respetivamente. Na representação dos padrões geográficos obtidos por HCA para o ano de 2005 (a) 2007 (c) são usadas cores distintas para representar os grupos de estados, de acordo com as estruturas definidas em (b) e (d), respetivamente.

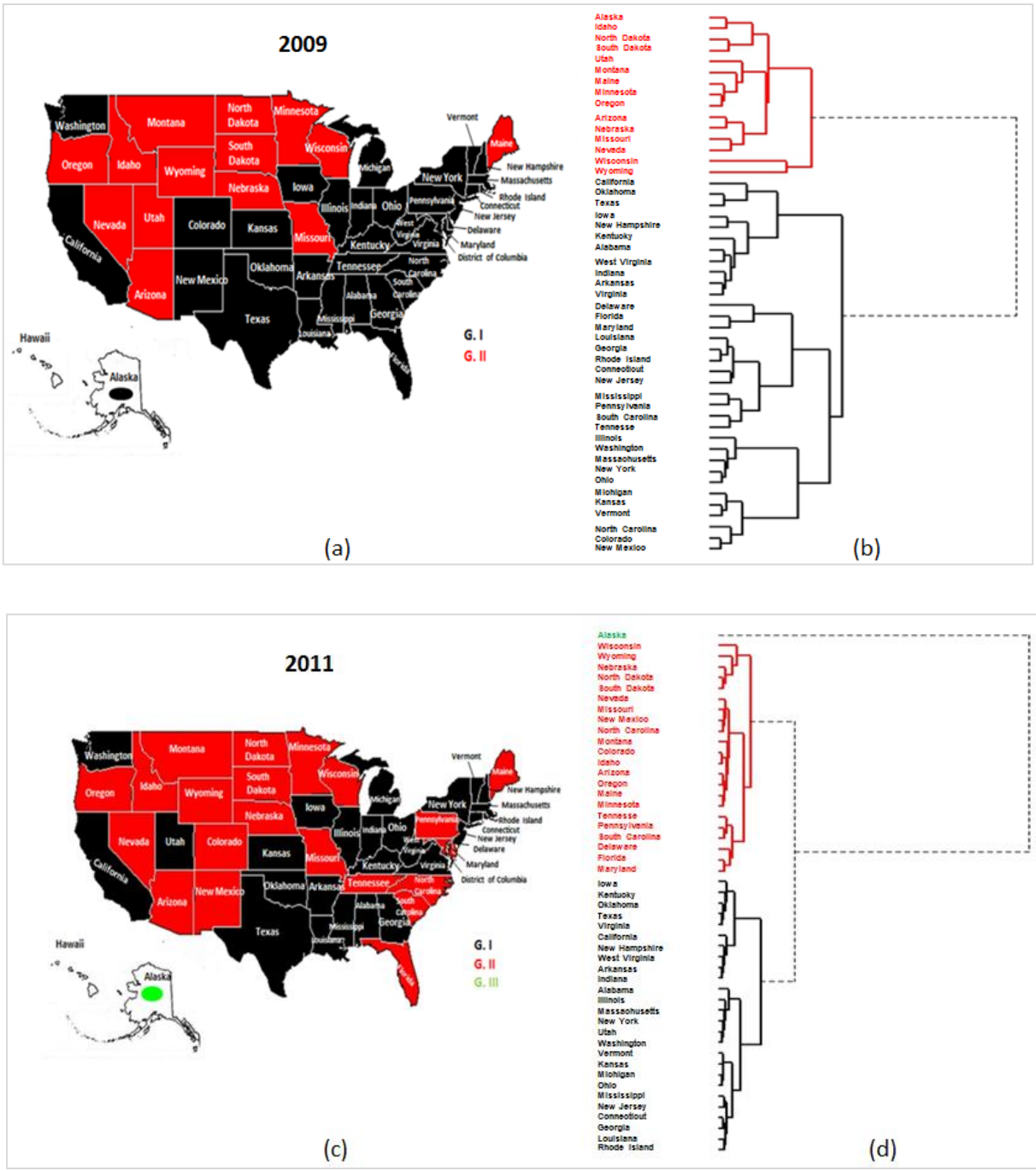


Figura 4.5 – Representação dos padrões geográficos relativos a 2009 e 2011, seguindo o mesmo esquema da Figura 4.4. De acordo com o critério definido na secção 3.3.4 são estabelecidos 2 grupos em 2009, (a) e (b) e 3 grupos em 2011, (c) e (d).

Tabela 4.3 – Representação dos grupos formados através de análise hierárquica de agrupamentos para o ano de 2005. Cada estado é caracterizado pela frequência de cada tipo de crime dividida pela população total do estado, com base na ligação de *Ward.* Para o estabelecimento dos grupos é usada uma linha de corte adequada definida na secção 3.3.4.

Grupo III	Grupo II	Grupo I	
Arizona	California	Connecticut	Alabama
Colorado	Florida	Delaware	Alaska
Minnesota	Kentucky	Georgia	Arkansas
Nebraska	Maryland	Illinois	Idaho
North Dakota	Oklahoma	Kansas	Indiana
Oregon	Texas	Louisiana	Iowa
Wisconsin		Massachusetts	Maine
Wyoming		Mississippi	Michigan
		Missouri	Montana
		Nevada	New Hampshire
		New Jersey	New Mexico
		New York	Tennessee
		North Carolina	Utah
		Ohio	Vermont
		Pennsylvania	Virginia
		Rhode Island	Washington
		South Carolina	West Virginia
		South Dakota	

Tabela 4.4 – Representação dos grupos formados através de análise hierárquica de agrupamentos para o ano de 2011. Cada estado é caracterizado pela frequência de cada tipo de crime dividida pela população total do estado, com base na ligação de *Ward*. Para o estabelecimento dos grupos é usada uma linha de corte adequada definida na secção 3.3.4.

Grupo III	Grupo II		Grupo I	
Alaska	Colorado	Arizona	Alabama	Connecticut
	Missouri	Delaware	Arkansas	Georgia
	Montana	Florida	California	Kansas
	Nebraska	Idaho	Illinois	Louisiana
	Nevada	Maine	Indiana	Michigan
	New Mexico	Maryland	Iowa	Mississippi
	North Carolina	Minnesota	Kentucky	New Jersey
	North Dakota	Oregon	Massachusetts	New York
	South Dakota	Pennsylvania	New Hampshire	Ohio
	Wisconsin	South Carolina	Oklahoma	Rhode Island
	Wyoming	Tennessee	Texas	Utah
			Virginia	Vermont
			West Virginia	Washington

Neste tipo de análise os estados têm tendência para se agruparem em zonas contíguas, como é observável pelo padrão geográfico, sendo estas constituídas por um grande número de estados. No dendrograma referente ao ano de 2005 são visíveis três grupos, sendo o grupo I o mais compacto constituído por 35 estados. No ano de 2007, verifica-se um aumento do número de grupos, passando para cinco grupos, assemelhando-se mais ao padrão obtido na abordagem anterior (dados em bruto).

Nas estruturas referentes a 2009 e 2011, é visível uma redução do número de grupos, os quais naturalmente incluem mais estados na sua composição, verificando-se uma tendência para os grupos se agruparem, criando zonas no mapa, como referido anteriormente. No mapa relativo a 2009, é possível observar que a zona central (a vermelho) se mantêm preservada com alguns estados mais dispersos, seguindo toda uma zona mais alargada correspondendo aos estados do grupo I. Resumindo, geograficamente existe alguma tendência para que os grupos se juntem, criando zonas constituídas por um grande número de estados, como é notório no padrão geográfico, com algumas exceções, em que alguns não seguem esta tendência. Relativamente às observações anteriores mantém-se uma zona central até à fronteira com o Canadá, alguma prevalência de estados a Este e, em 2011, *California* e *Texas* ainda fazem parte de um mesmo grupo, agora alargado. Nesta fase, não é trivial apresentar uma hipótese explicativa para este perfil.

4.1.3 Fração de crime

Uma questão fundamental diz respeito à normalização dos dados. Para este efeito, a nossa opção é autonormalizar cada estado, isto é, cada variável é usada na forma de uma fração,

$$f_{ij} = n_{ij} / \sum_{i=1}^{N} n_{ij} \tag{4.1}$$

onde f_{ij} corresponde à fração de detenções do tipo *i* no estado *j*; n_{ij} corresponde ao número de detenções do tipo *i* nesse estado e $\sum n_{ij}$, corresponde ao somátorio de detenções do tipo *i* para cada estado. Isto significa que, o número de detenções para cada tipo de crime é dividido pelo número total correspondente àquele estado. Deste modo, cada estado é descrito por um conjunto de variáveis que são a fração das detenções previstas para cada tipo de crime. Os estados mais similares são os que apresentam o mesmo perfil de detenções, independentemente da magnitude de incidência.

O conjunto de dados resultante é processado seguindo, novamente, procedimento idêntico ao das abordagens anteriores.

Nas Figuras 4.6 e 4.7 encontram-se representadas as estruturas e os padrões geográficos encontrados no período considerado, usando a fração de crime.

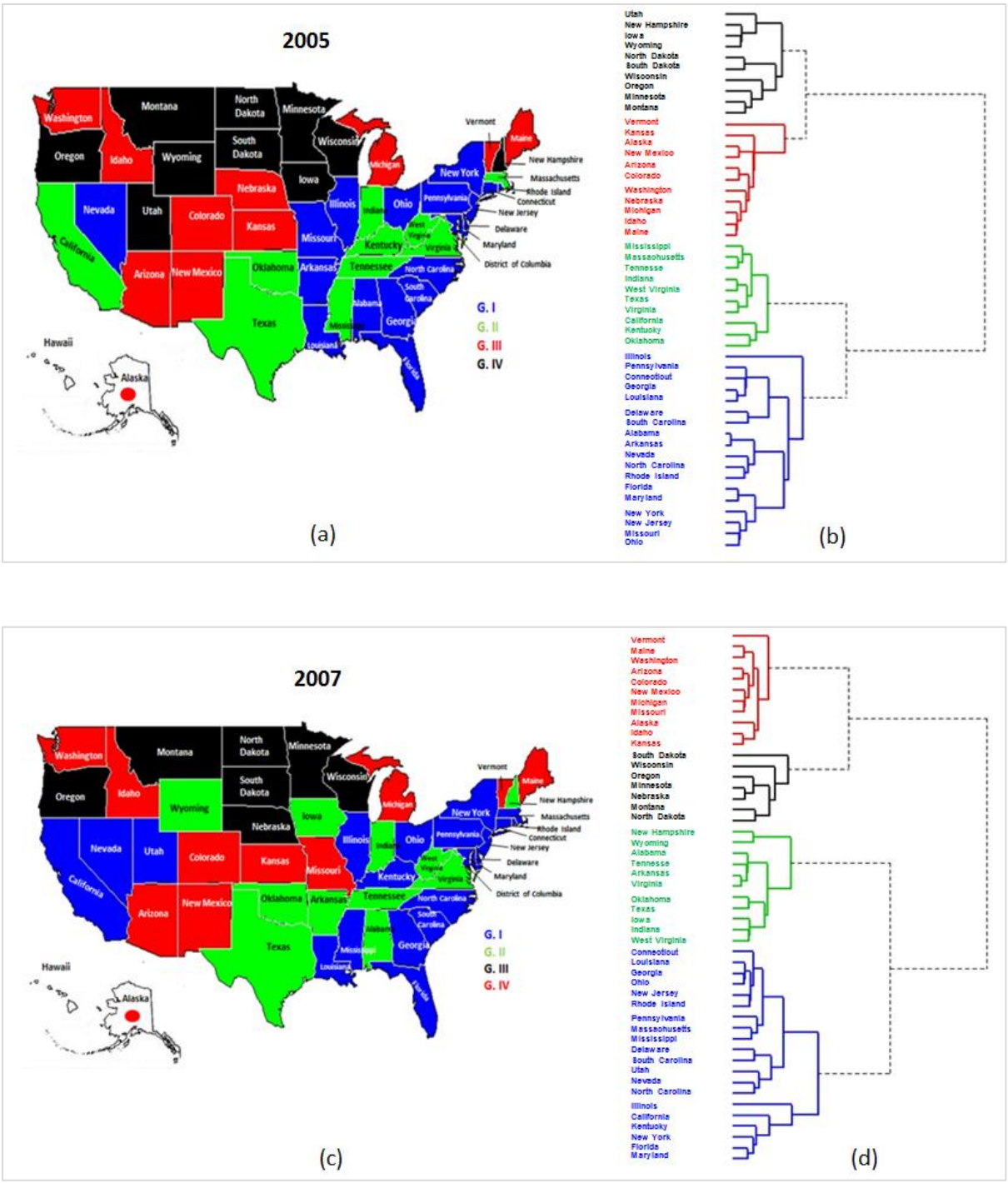


Figura 4.6 - Dendrogramas construídos por HCA (método de ligação *Ward*) sobre o conjunto de dados correspondentes a 49 estados dos EUA, de acordo com a tabela 69 do FBI UCR para o ano de 2005 (b) e 2007 (d). Para o estabelecimento dos grupos foi usada uma linha de corte definida pela função descrita na secção 3.3.4, da qual resultaram 4 grupos para ambos os anos. Na representação dos padrões geográficos obtidos por HCA para o ano de 2005 (a) 2007 (c) são usadas cores distintas para representar os grupos de estados, de acordo com as estruturas definidas em (b) e (d), respetivamente.

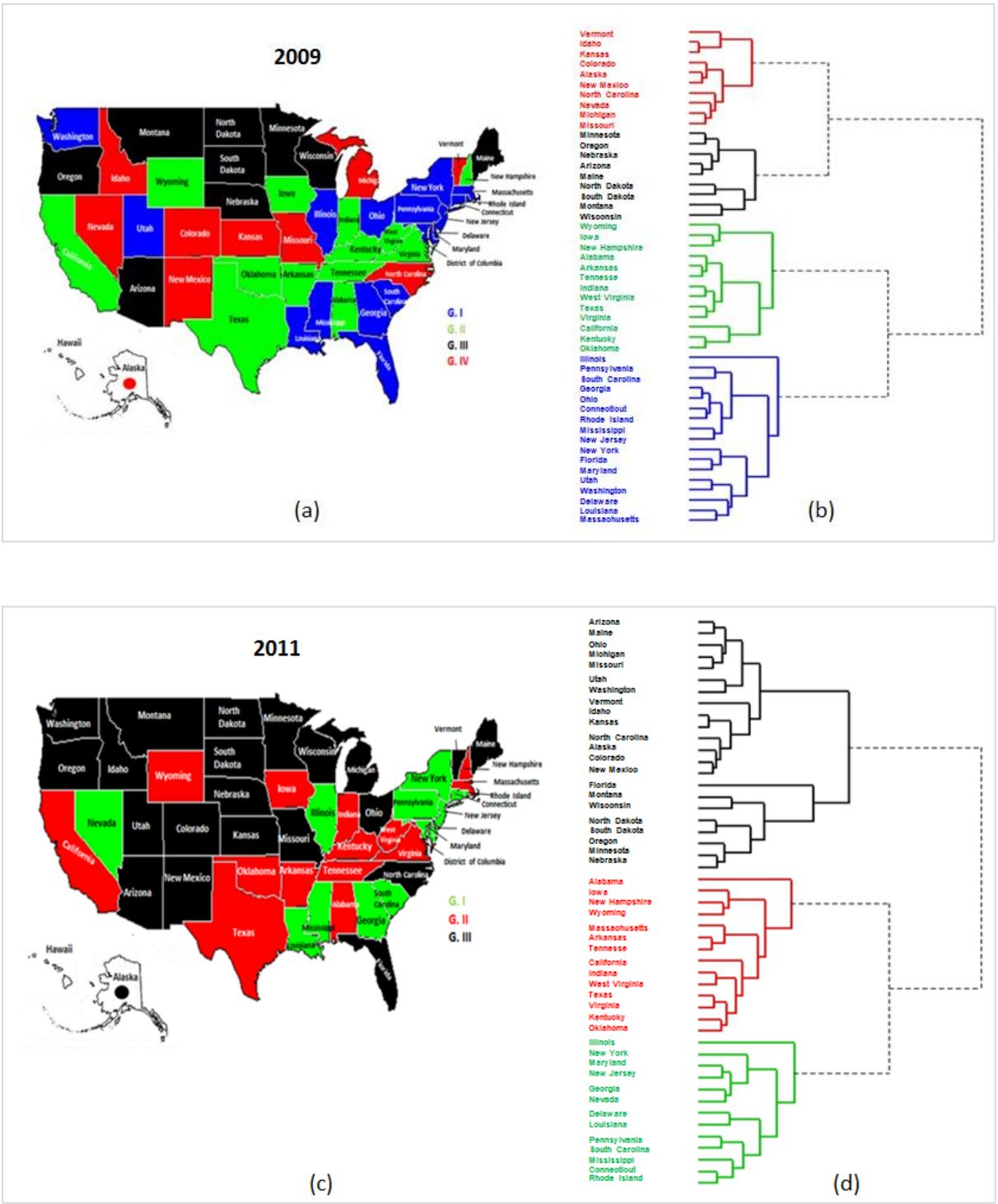


Figura 4.7 – Representação dos padrões geográficos relativos a 2009 e 2011, seguindo o mesmo esquema da Figura 4.6. De acordo com o critério definido na secção 3.3.4 são estabelecidos 4 grupos em 2009, (a) e (b), e 3 grupos em 2011, (c) e (d).

Neste perfil observam-se diferenças significativas relativamente aos dois perfis anteriores (número de detenções e taxa de crime). Relativamente aos dendrogramas para os quatro anos em estudo (2005, 2007, 2009 e 2011), têm o mesmo número de grupos, ou seja quatro à exceção do dendrograma para o ano de 2011 em que houve uma redução e são visíveis três grupos, resultando um deles da fusão de outros dois.

O padrão geográfico resultante da análise hierárquica assemelha-se mais ao perfil com base no número de detenções, estando os estados homogeneamente dispostos nos mapas. A zona central (região a preto) para o ano de 2005, a partir da fronteira do Canadá é observável e mantida para os restantes mapas em estudo. A Este no mapa (região a azul) para o ano de 2005, é observável um conjunto de estados pertencentes ao mesmo grupo, que se mantêm para os restantes mapas. Pertencentes a essa região, alguns estados se encontram dispersos, como o estado de *Nevada* para o mapa referente ao ano de 2005, *California*, *Nevada* e *Utah* para o mapa de 2007 e *Washington* e *Utah* para o mapa de 2009. Os grupos II e III referente ao ano de 2005 que para os anos de 2007 e 2009 correspondem aos grupos II e IV, são os grupos em que os estados se encontram mais dispersos.

Relativamente ao padrão geográfico para o ano de 2011, verifica-se a perda de formação que parecia existir nos anos anteriores. Na análise anterior verificava-se que os estados se encontravam homogeneamente distribuídos pelos grupos, o que em 2011 não se verificou. Para os três anos (2005, 2007 e 2009) o padrão dispunha-se em quatro grupos, em 2011 cingiu-se a três grupos, ou seja, dois grupos fundiram-se para surgir num único grupo. Tal como se verificou para a o sistema anterior (taxa de criminalidade) também neste sistema se observou a mesma tendência. Sendo assim, os estados parecem aproximar-se em termos numéricos de criminalidade para valores muito idênticos.

Em seguida, as Tabelas 4.5 e 4.6 reúnem os grupos formados, através da análise hierárquica de agrupamentos referentes aos anos de 2005 e 2011. Nas tabelas estão discriminados os grupos explicitamente numerados, com as cores dos grupos de acordo com a análise hierárquica.

Tabela 4.5 – Representação dos grupos formados através de análise hierárquica de agrupamentos para o ano de 2005, na qual cada estado é um objeto que é caracterizado pela fração de cada tipo de crime relativamente ao número total cometido nesse estado, com base na ligação de *Ward.* Para o estabelecimento dos grupos foi utilizado uma linha de corte adequada.

Grupo IV	Grupo III	Grupo II	Grupo I
Iowa	Alaska	California	Alabama
Minnesota	Arizona	Indiana	Arkansas
Montana	Colorado	Kentucky	Connecticut
New Hampshire	Idaho	Massachusetts	Delaware
North Dakota	Kansas	Mississippi	Florida
Oregon	Maine	Oklahoma	Georgia
South Dakota	Michigan	Tennessee	Illinois
Utah	Nebraska	Texas	Louisiana
Wisconsin	New Mexico	Virginia	Maryland
Wyoming	Vermont	West Virginia	Missouri
	Washington		Nevada
			New Jersey
			New York
			North Carolina
			Ohio
			Pennsylvania
			Rhode Island
			South Carolina

Tabela 4.6 – Representação dos grupos formados através de análise hierárquica de agrupamentos para o ano de 2011, na qual cada estado é um objeto que é caracterizado pela fração de cada tipo de crime relativamente ao número total cometido nesse estado, com base na ligação de *Ward.* Para o estabelecimento dos grupos foi utilizado uma linha de corte adequada.

Grupo III		Grupo II	Grupo I
Alaska	Colorado	Alabama	Connecticut
Arizona	Florida	Arkansas	Delaware
Idaho	Minnesota	California	Georgia
Kansas	Montana	Indiana	Illinois
Maine	Nebraska	Iowa	Louisiana
Michigan	New Mexico	Kentucky	Maryland
Missouri	North Dakota	Massachusetts	Mississippi
North Carolina	Oregon	New Hampshire	Nevada
Ohio	South Dakota	Oklahoma	New Jersey
Utah	Wisconsin	Tennessee	New York
Vermont		Texas	Pennsylvania
Washington		Virginia	Rhode Island
		West Virginia	South Carolina
		Wyoming	

4.2 Caracterização no espaço e no tempo

Nesta secção procederemos a uma identificação das variáveis caracterizadoras dos três sistemas (número de detenções, taxa de criminalidade e fração de crime) anteriormente sujeitos ao estudo de agrupamentos. Tentaremos também extrair informação que nos permita confirmar ou esclarecer a formação dos grupos formados nas análises anteriores. O critério para a seleção das contribuições mais significativas é baseado na comparação com o valor médio esperado. Sabendo que uma componente principal constituiu uma base vetorial ortonormada, para um caso *m* dimensional esperamos um valor médio de $1/\sqrt{m}$. Nesta fase o conjunto de dados contém a informação relativa aos 4 anos considerados, constituindo assim uma matriz 196x27. Esta análise assentará na matriz de covariância. Lembramos que a matriz de covariância privilegia as grandes variações no conjunto de dados.

Nesta análise, o resultado do PCA assentou na matriz global, ou seja, inclui os dados referentes aos anos de 2005, 2007, 2009 e 2011, traduzindo uma visão temporal sobre os vários sistemas.

Para uma caracterização mais específica foi excluído da análise duas variáveis pouco específicas que juntam infrações de ordem variada como as *other assaults* e *all other offenses*.

4.2.1 Número de detenções

O resultado da utilização de PCA sobre o número de detenções revela um padrão de interpretação bastante simples, como já indicado nos resultados HCA. De facto, apenas uma componente é suficiente para explicar mais de 80% da variabilidade total, esta contribui com 85.08%. Na Figura 4.8 representam-se as *loadings* para as duas primeiras componentes.

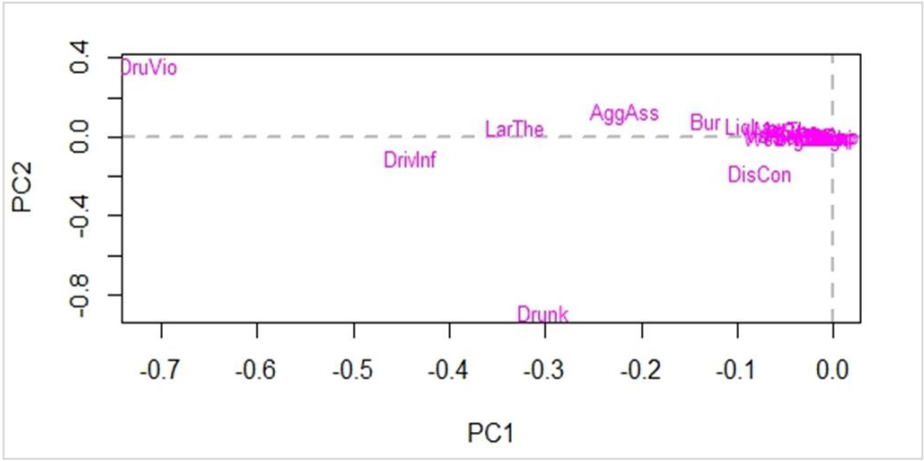


Figura 4.8 – Representação das *loadings* obtidas por análise de componentes principais para o número de detenções. Note-se que, embora por facilidade de representação constem as duas primeiras componentes, com uma recuperação de 90.50 % da variabilidade total.

É visível que todas as *loadings* da primeira componente se encontram na mesma zona do eixo. Tal significa, simplesmente, que avançar ao longo desse eixo corresponde a evoluir de um maior número de crimes total para um menor número. Nota-se que apesar de esta evolução não ser específica para determinadas variáveis, existem algumas de maior relevância para a caracterização dos dados. Citem-se entre elas a *drug abuse violations* e *driving under the influence*. Tratam-se, simplesmente das que ocorrem com maior frequência nos anos de 2005, 2007, 2009 e 2011 em cada estado e que, portanto, dominam nesta análise.

A Tabela 4.7 resume os resultados do PCA para as primeiras quatro componentes principais.

Tabela 4.7 – Resultados do PCA para as quatro primeiras componentes principais para o perfil, número de detenções utilizando a matriz de covariância.

#	Variância explicada (%)	Variância explicada cumulativa (%)
PC1	85.1	85.0
PC2	5.42	90.5
PC3	4.52	95.0
PC4	2.35	97.4

A primeira componente principal descreve 85.08% da informação inicial. Neste caso, em que é usada a matriz de covariância bastava uma componente. No entanto, como a representação é bidimensional, são necessárias as duas primeiras, a segunda componente não fornece informações adicionais significativas.

Inspeccionaremos agora o impacto da evolução temporal, ano a ano, nos limites dos grupos com a utilização do algoritmo *convex hull*. Deste modo, os limites dos conjuntos correspondentes a cada ano são definidos pelos estados mais afastados da origem (veja-se a Figura 4.9) Por conveniência, os nomes dos estados surgem abreviados.

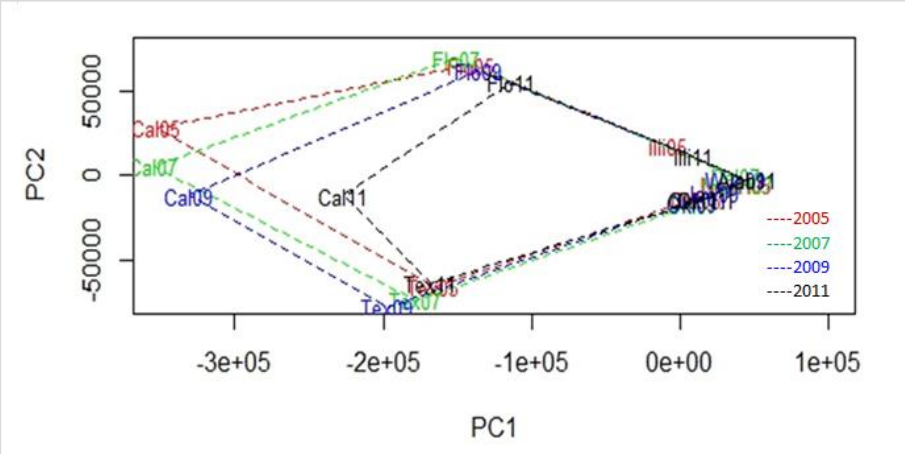


Figura 4.9 – Representação dos limites dos conjuntos de estados correspondente ao número de detenções para os anos de 2005, 2007, 2009 e 2011 sob a forma de *convex hull*, com 90,5% de informação inicial recuperada. As cores referem-se aos anos nos quais foram efetuadas as estimativas.

É notória uma forte compressão da fronteira ao longo do eixo PC1 e uma ligeira compressão ao longo do eixo PC2 para os quatro anos em estudo. Esta é mais evidente para os estados da *California*, *Texas*, *Florida* e Illinois, estando os restantes restritos a uma pequena área no gráfico. Assim sendo, a compressão ao longo do eixo PC1 é marcada pelo decréscimo das *drug abuse violations* e a compressão ao longo do eixo PC2 com uma diminuição das infracções relacionadas com a bebida: *drunknness*. Desta forma, é possível visualizar as grandes alterações durante os anos em estudo para as variáveis caracterizadoras do sistema. Os resultados obtidos pelo PCA são, genericamente, concordantes ou complementares em relação aos resultados HCA. Note-se que uma redução no número de grupos é compatível com uma redução na variabilidade.

4.2.2 Taxa de criminalidade

Recorrendo a uma definição de cada estado com base na respetiva taxa de criminalidade, são necessários quatro componentes para recuperar cerca de 85.3% da variabilidade original. Temos, assim, perto de 39% para a primeira componente e 21% para a segunda. A existência de quatro componentes é um claro indicador de que o padrão obtido se reveste de maior complexidade do que na análise da frequência de crimes. Opta-se por utilizar as duas primeiras componentes, porque recuperam uma percentagem da variabilidade suficiente para que o gráfico reflita o comportamento do sistema. Os padrões geográficos formados a partir da análise de agrupamentos para ambos os perfis são reveladores da complexidade deste sistema.

Na Figura 4.10 apresentam-se as *loadings* correspondentes ao perfil taxa de criminalidade em análise.

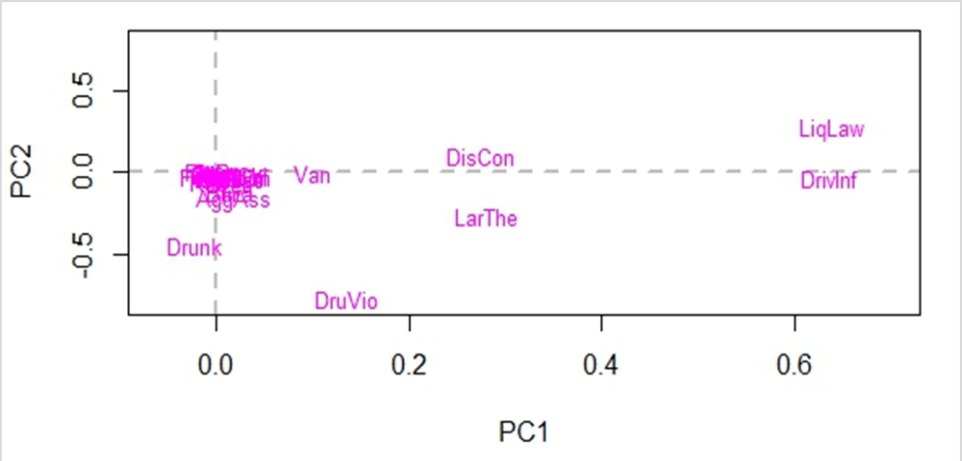


Figura 4.10 - Representação das loadings obtidas por análise de componentes principais, para a matriz global (2005, 2007, 2009 e 2011) do perfil taxa de criminalidade. As duas componentes produzem uma recuperação de 60% da variabilidade total.

É evidente que *liquor laws* e *driving under influence* dominam claramente na primeira componente enquanto que na segunda se destacam crimes relacionados com *drug abuse violations* e *drunknness*. Refira-se que as *loadings* para a segunda componente têm sinais opostos às restantes.

A Tabela 4.8 resume os resultados do PCA para as primeiras cinco componentes principais.

Tabela 4.8 – Resultados do PCA para as cinco primeiras componentes principais para o perfil correspondente à taxa de criminalidade utilizando a matriz de covariância.

#	Variância explicada (%)	Variância explicada cumulativa (%)
PC1	39.3	39.3
PC2	20.7	60.0
PC3	15.6	75.6
PC4	9.72	85.3
PC5	5.71	91.0

Na Figura 4.11 encontra-se a representação dos limites dos conjuntos correspondentes aos quatro anos em estudo, sob a forma de *convex hull*, para a taxa de criminalidade.

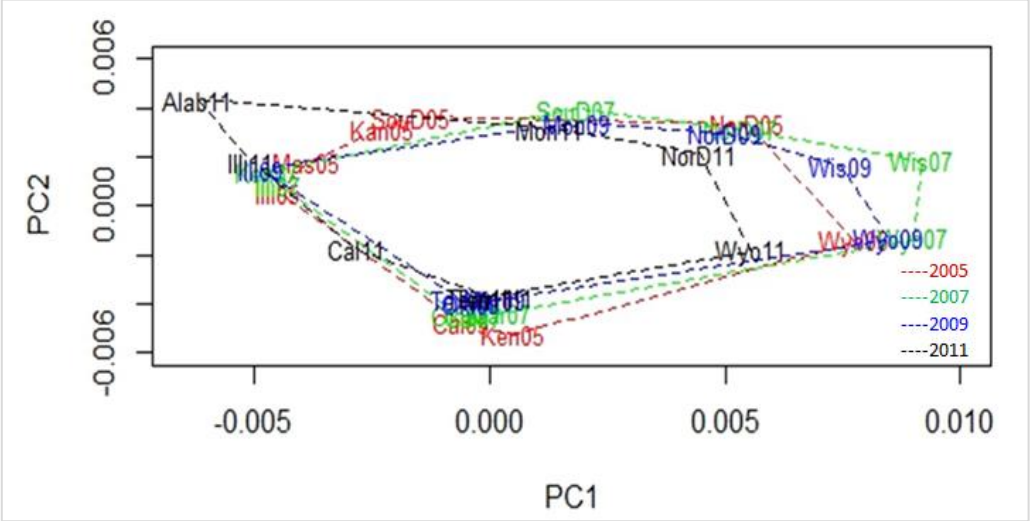


Figura 4.11 - Representação dos limites dos conjuntos de estados correspondente à taxa de criminalidade para os anos de 2005, 2007, 2009 e 2011 sob a forma de *convex hull*, com 60% de informação inicial recuperada. As cores referem-se aos anos nos quais foram efetuadas as estimativas.

Nota-se claramente nesta figura uma expansão dos limites de fronteira ao longo do eixo PC1 de 2005 a 2007. De 2007 a 2011, verifica-se uma compressão dos limites ao longo do eixo PC1, que é marcada, por exemplo para o estados de *North Dakota*, *Wisconsin* e *Wyoming*. A expansão e compressão ao longo do eixo PC1 é resultante de um aumento até 2007 seguido de uma diminuição até 2011, para infracções relacionadas com a bebida como sejam *liquor laws* e *driving under the influence*. Mais uma vez, grandes alterações são vistas para os anos em estudo, estando os resultados concordantes com os verificados pela análise hierárquica de agrupamentos (HCA).

4.2.3 Fração de crime

Utilizando-se para cada estado um perfil resultante do valor relativo da frequência de cada crime sobre o total de crimes cometidos para esse estado, necessitamos de quatro componentes para recuperar perto de 80% da variabilidade original. A primeira componente corresponde a cerca de 39%, obtendo-se 58% incluindo a segunda.

Na Figura 4.12 apresentam-se as *loadings* correspondentes ao perfil sobre análise.

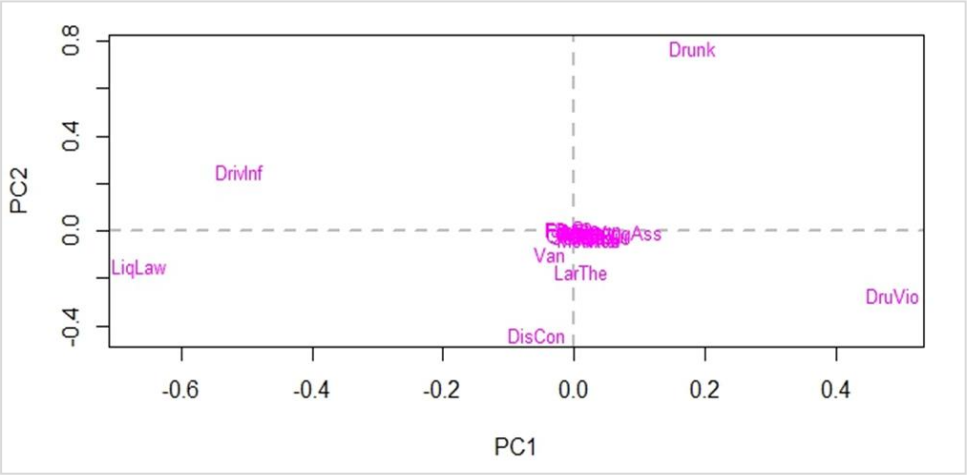


Figura 4.12 - Representação das *loadings* obtidas por análise de componentes principais, para a matriz global (2005, 2007, 2009 e 2011) do perfil fração de crime. Com uma recuperação de 58% da variabilidade total.

É evidente que *liquor laws, driving under the influence* e *drug abuse violations* dominam no eixo PC1, sendo que as duas primeiras estão na parte negativa do eixo. Na segunda componente, dominam as infracções relacionadas com a bebida como *drunknness* e *disorderly conduct*, sendo esta última pertencente à parte negativa do eixo PC2.

A Tabela 4.9 resume os resultados do PCA para as primeiras cinco componentes principais.

Tabela 4.9 – Resultados do PCA para as cinco primeiras componentes principais para o perfil fração do crime utilizando a matriz de covariância.

#	Variância explicada (%)	Variância explicada cumulativa (%)
PC1	38.5	38.5
PC2	19.8	58.3
PC3	12.0	70.3
PC4	9.25	79.5
PC5	7.71	87.3

A primeira componente principal descreve 39% da informação inicial acrescentando a segunda componente 20%, perfazendo as duas um total de variância explicada cumulativa de 59%. Na Figura 4.13 encontra-se a correspondente representação *convex hull*.

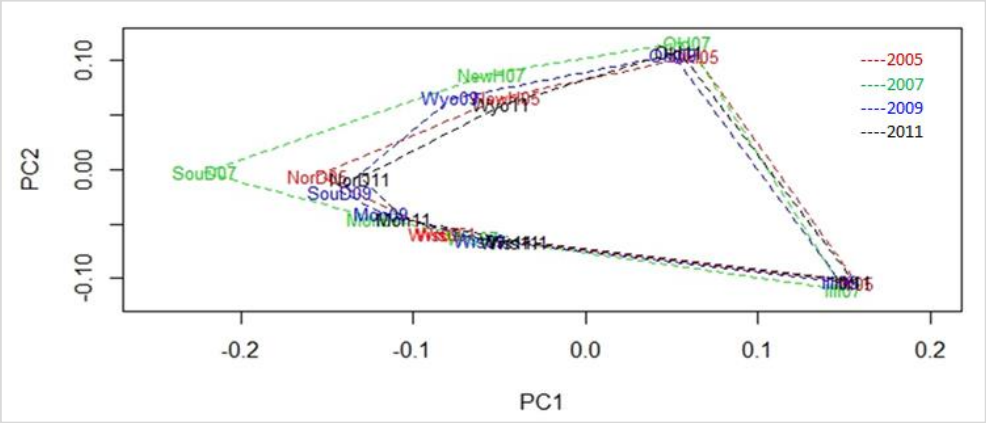


Figura 4.13 - Representação dos limites dos conjuntos de estados correspondente à fração de crime para os anos de 2005, 2007, 2009 e 2011 sob a forma de *convex hull*, com 58.28% de informação inicial recuperada. As cores referem-se aos anos nos quais foram efetuadas as estimativas.

É evidente nesta figura uma expansão dos limites de fronteira ao longo do eixo PC1 de 2005 a 2007, mais notório para os estados que se situam na parte negativa do eixo PC1. Tal é evidente para o estados de *South Dakota*, *North Dakota* e *New Hampshire*, seguido de uma compressão nos limites de fronteira ao longo do eixo PC1 de 2007 a 2011. A expansão notória deve-se a um aumento das infrações *liquor laws* e *driving under the influence* de 2005 a 2007 e a compressão a um decréscimo destas mesmas infrações. Novamente, infrações relacionadas com a bebida estão na base das alterações do perfil.

Os resultados das secções anteriores, devido à utilização de três diferentes perfis (número de detenções, taxa de criminalidade e fração de crimes, todas expressas por tipo de crime) tornam complicada uma análise global. Será, no entanto, de referir que o tipo de perfil escolhido para a caracterização é determinante no tipo de resultados obtidos. Quando utilizamos como perfil, por exemplo, o número de detenções encontramos uma distribuição de estados para os quatro anos de estudo, aparentemente determinada por fatores geográficos, observando-se para o ano de 2009 uma grande diferença nas características gerais, face aos restantes anos.

Capítulo 5

Taxa de criminalidade e população

5.1 Perspetiva geral

No capítulo 4, fez-se uma caracterização no espaço e no tempo dos três sistemas em estudo (número de detenções, taxa de criminalidade e fração de crime). Note-se, no entanto, que o sistema, número de detenções é marcado por fatores geográficos enquanto que a taxa de criminalidade para além desses acresce os demográficos. Para além disso, estes sistemas podem ser globais, considerando todos os delitos, ou podem ser estabelecidos de acordo com características específicas, por exemplo, distinguindo o crime violento do crime contra a propriedade.

Neste âmbito, a escolha da variável dependente neste tipo de estudos torna-se extremamente controversa. Por exemplo, a validade teórica da utilização de quocientes como os que surgem na taxa de criminalidade, tem sido questionada. No entanto, a sua utilidade é reconhecida, uma vez que, a taxa de criminalidade representa o número de delitos numa base *per capita*, permitindo fazer comparações entre jurisdições correspondentes a populações diferentes [112].

Na Figura 5.1 encontra-se demostrada a dependência positiva existente entre o número total de detenções e o tamanho da população, para os quatro anos (2005, 2007, 2009, 2011) considerados.

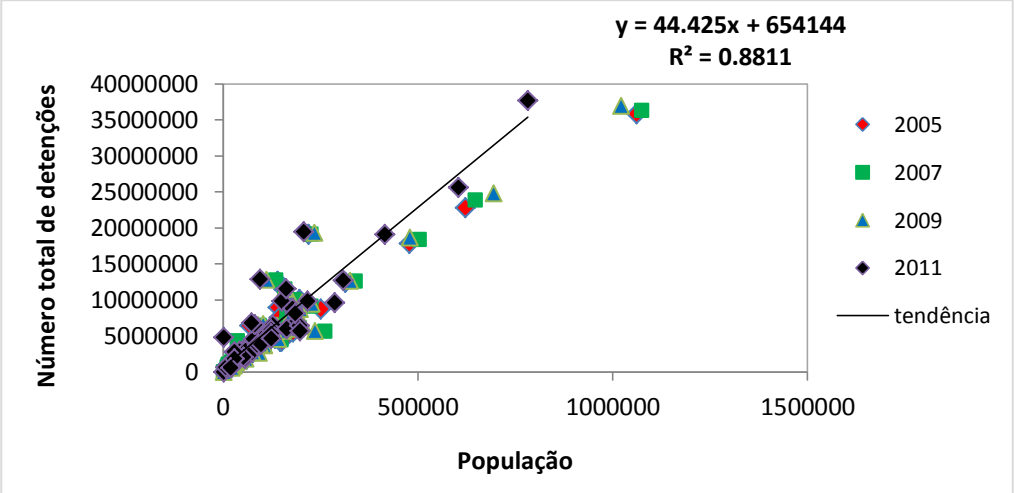


Figura 5.1 – Representação do Número total de detenções vs. População, dos estados americanos para os anos em estudo (2005, 2007, 2009 e 2011).

Pela análise do gráfico da Figura 5.1 é notório que o número total de detenções correlaciona-se muito bem com o tamanho da população sobre o qual é determinado, sendo um facto bem estabelecido

No entanto, a relação entre taxa de criminalidade e tamanho da população é menos clara. Genericamente, considera-se que existe uma evidência considerável, baseada na informação policial, que substancia uma relação positiva entre a população e o crime *per capita*, apesar de existirem algumas exceções [113-114]. Estas surgem para certos delitos, [115-117], em certos pontos temporais, [118-119], ou em certos locais [120-121]. Alguns estudos sugerem que esta dependência pode estar a enfraquecer [113,122]. Além disso, a relação positiva entre tamanho de uma população e o crime é frequentemente citada como sendo um dos factos da criminologia, [123-126]. As incoerências são, no entanto, muitas e como demonstrado na referência [127], a análise multivariada revela que enquanto o tamanho da população não tem um efeito notório na taxa de criminalidade violenta ou na taxa de criminalidade dos crimes contra a propriedade, continua a ser, de longe, o melhor fator de previsão para o número de crimes violentos ou contra a propriedade.

Mesmo quando se considera que a taxa de criminalidade aumenta com o tamanho da população, têm sido registadas discrepâncias entre as chamadas análises transversais (estudo observacional através da coleta de dados de uma população) e as longitudinais [126]. Nas primeiras, faz-se o congelamento temporal e analisam-se as várias jurisdições. Caso a premissa seja verdadeira, espera-se que a representação da taxa de criminalidade vs. população tenha um declive positivo. Por outro lado, na evolução temporal, espera-se, que um aumento na população produza um aumento, eventualmente com decalagem, na taxa de criminalidade. Estudos que verificaram este comportamento para um determinado espaço temporal, não conseguiram identificar o mesmo na evolução temporal referida [126], talvez devido a um grande investimento em medidas de prevenção e redução do crime, principalmente nos grandes centros urbanos [25,127]. Um outro aspeto tem a ver com o facto de, a dependência da taxa de criminalidade na população ser variável, alterando-se com a gama que se esteja a considerar. Numa abordagem em que foram considerados grupos de cidades, desde as menos populosas até às de elevada população, observou-se que apenas para populações mais elevada, a taxa de criminalidade pode diminuir com a população [112]. Tendo em conta os estudos encontrados na literatura, podemos afirmar que os resultados e abordagens apresentadas são díspares, por vezes mesmo contraditórios e pouco conclusivos.

5.2 Distribuição do crime usando a curva de Lorenz

Nesta secção pretende-se avaliar a dispersão do crime nos estados americanos no período de 2005 a 2011, usando medidas de desigualdade estabelecidas na econometria. A desigualdade na distribuição do crime pode ser avaliada utilizando várias medidas como o coeficiente de Gini, o coeficiente de variação e o índice de Theil [128]. Estas medidas permitem aos dirigentes políticos determinar se a prevenção do crime e a actividade policial para determinados tipos de crime deve ser direcionada para um número limitado de estados e regiões ou de forma mais abrangente, para todo o território americano.

A fim de enumerar as diferenças entre os crimes responsáveis pela discriminação dos estados, são consideradas as categorias de crime de maior contribuição para as duas primeiras componentes (PC1 e PC2), resultantes da análise de componentes principais sobre os dados relativos às detenções efetuadas em cada estado no período de 2005 a 2011.

Nesta fase, a desigualdade da distribuição do crime tendo em conta o número de detenções nos estados americanos é analisada de um modo mais ilustrativo e simples através da construção de curvas de Lorenz.

A curva de Lorenz é construída para cada uma das categorias de crime com base no número de detenções efetuadas, em cada estado, durante o período de 4 anos. O índice de Gini é usado para representar a extensão da desigualdade.

Construção da curva de Lorenz e coeficiente de Gini

O coeficiente de Gini é uma medida de concentração que pode ser utilizada em análises de concentração ou distribuição de indicadores sociais e económicos. Neste contexto, o coeficiente de Gini é usado para avaliar a distribuição dos crimes nos estados americanos tendo em conta o número de detenções. Este índice pode ser calculado com base na expressão,

$$G = 1 - \sum_{i=1}^{n} (X_i - X_{i-1})(Y_i + Y_{i-1}) \tag{5.1}$$

sendo X_i a proporção acumulada da população dos estados; Y_i, a proporção acumulada das detenções relativas aos crimes considerados e n, o número de estados.

Seja p o valor da proporção da população num dado estado i e d o valor correspondente da proporção acumulada das detenções, os pares de valores (p,d) para os diversos estados definem um conjunto de pontos, que constituem a curva de Lorenz.

Na Figura 5.2 é apresentada uma curva de Lorenz teórica para a distribuição dos crimes. Esta representação mostra como a proporção acumulada das detenções varia em função da proporção acumulada da população, com os estados organizados por ordem crescente de número detenções.

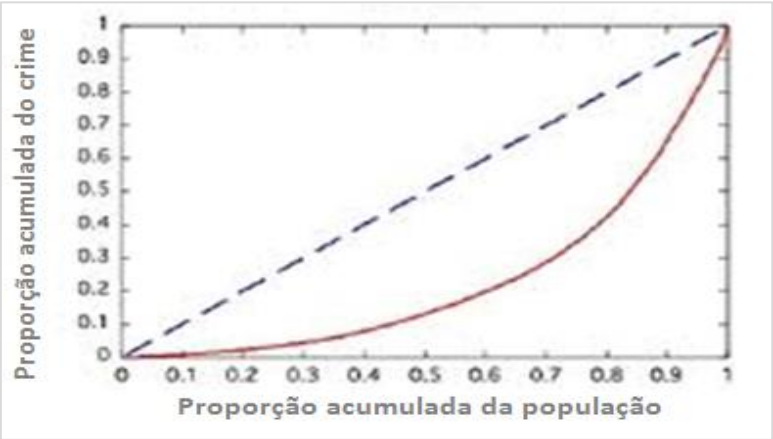


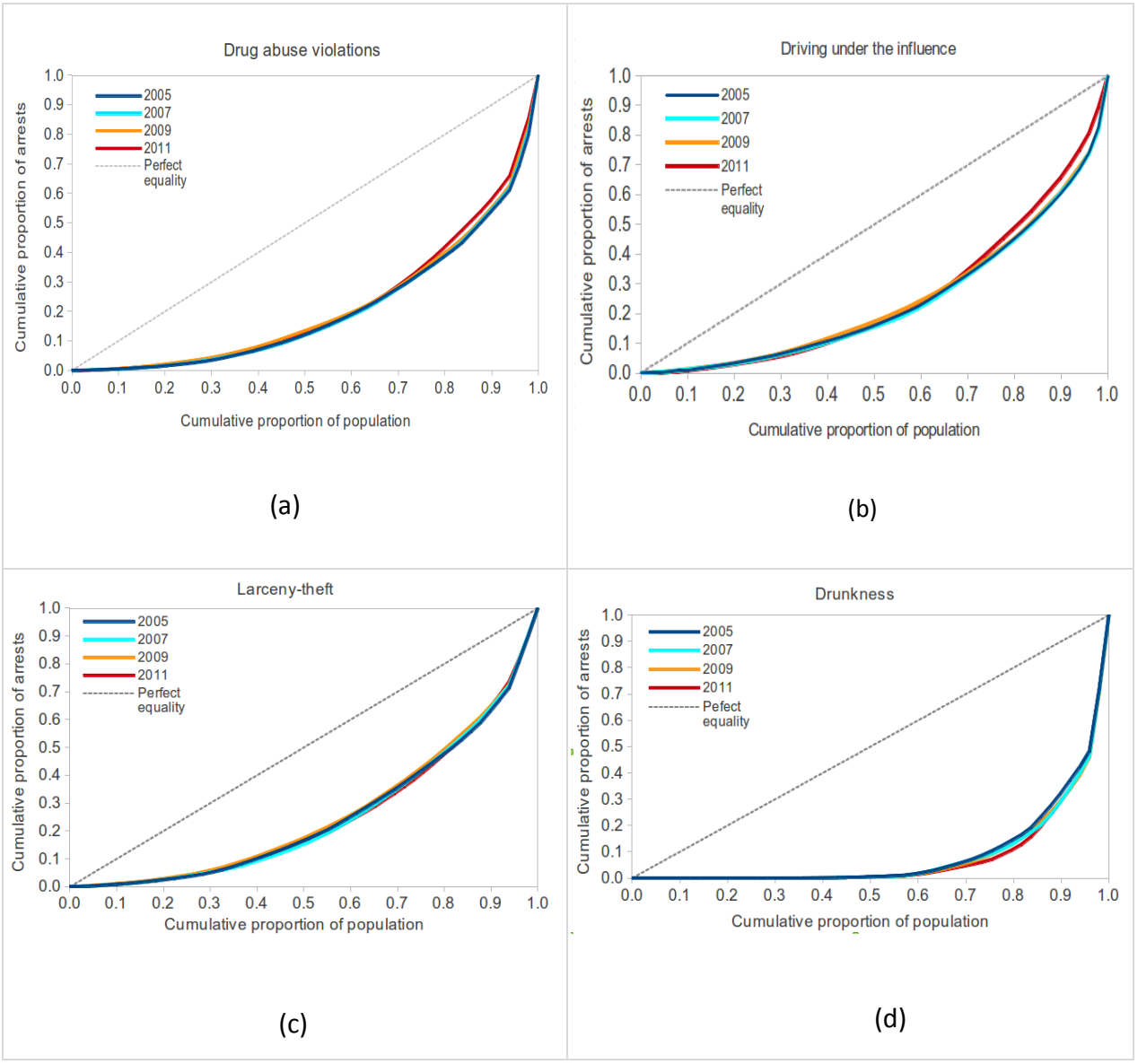
Figura 5.2 – Representação da curva de Lorenz teórica para a distribuição dos crimes.

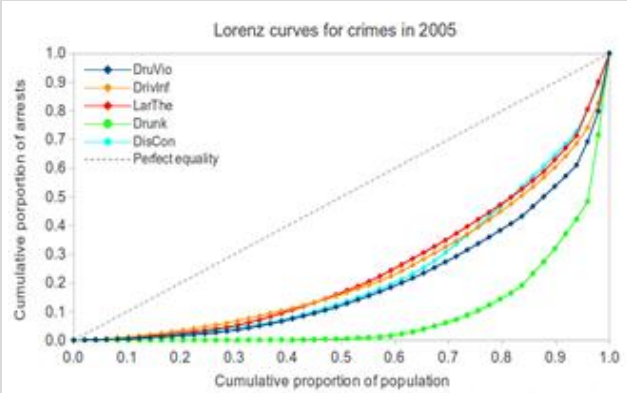
Por exemplo, ao representar a distribuição da variável associada a *larceny-theft* utilizando a curva de Lorenz, é possível observar quanto da fração acumulada desse crime varia em função da fração acumulada da população dos estados. No eixo das abscissas é representada a proporção acumulada dos estados em ordem crescente do número de detenções e no eixo das ordenadas é representada a proporção acumulada das detenções. Numa distribuição perfeitamente uniforme, a 10% dos estados com menor número de detenções deve corresponder 10% das detenções, metade dos estados, metade das detenções, e assim sucessivamente. Neste caso, a curva de Lorenz seria uma linha recta com inclinação de 45 graus como consta no gráfico da Figura 5.2. Quanto mais convexa for a curva, mais os estados se distinguem em termos de número de detenções.

Os valores observáveis da curva de Lorenz relativos às detenções em cada categoria de crime, num ano particular, são calculados da seguinte forma: (i) para cada tipo de crime, os estados são ordenados de forma crescente tendo em conta o número total de detenções; (ii) o total de ocorrências nos estados é calculado, representando o número total de detenções de uma categoria específica de crime ao longo de todos os estados; (iii) em cada passo é calculada a soma cumulativa das detenções nos estados, isto é, no primeiro passo o número total de detenções para um tipo de crime correspondente ao estado de menor incidência, é calculado como uma proporção do total de detenções relativas esse tipo de crime para todos os estados. Este valor está associado ao eixo das ordenadas. O valor correspondente no eixo das abcissas é a proporção de população total no mesmo estado. Por outras palavras, o número total de detenções para um estado específico é calculado

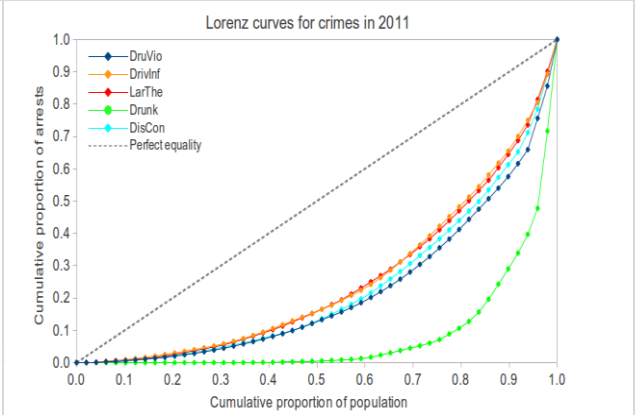
como a proporção do total de detenções e representado em função da proporção de população total nesse estado.

A Figura 5.3 apresenta as curvas de Lorenz obtidas para 4 categorias de crimes selecionadas com base nos resultados apresentados no capítulo 4.





(e)



(f)

Figura 5.3 – Curvas de Lorenz correspondentes aos tipos de crime selecionados, (a) *drug abuse violations*, (b) *driving under the influence*, (c) *larceny-theft* e (d) *drunkness*. As curvas representadas em (e) e (f) correspondem às perspetivas globais para os anos considerados como extremos do período considerado.

Considere-se a variável associada ao crime *drug abuse violations*. Em 2011, o estado de *Alabama* com o menor número de detenções, apresenta apenas 300 ocorrências nesta categoria. No total existem ca. 1.3 milhões de ocorrências reportadas nos Estados Unidos em 2011. Assim, este estado retém 0.02% do total de detenções para *drug abuse violations*. Este estado tem uma população de 4.76 milhões, o que corresponde, aproximadamente, a 1.6% da população global.

Se considerarmos que o número de detenções é distribuído ao longo dos estados nas mesmas proporções em que é distribuída a população, então as detenções são exatamente as mesmas em cada estado. Esta situação é representada pela linha diagonal. Quanto mais próxima a curva estiver da diagonal mais uniforme é a distribuição do crime associado a uma determinada categoria. Observando a forma das distribuições dos gráficos relativos às curvas de Lorenz para as detenções em 2005 e 2011, verifica-se que a categoria *drunkness* apresenta a maior diferença relativamente à propagação da população. Isto significa que, as detenções associadas a esta categoria estão mais concentradas em determinadas regiões.

Na Tabela 5.1 encontram-se os coeficientes de Gini calculados para cada categoria. Os valores confirmam as diferenças observadas entre as categorias de crime. Em geral, as categorias *drug abuse violations*, *drive under influence, larceny-theft* e *drunkness* não seguem o padrão de dispersão da população, dado que os valores variam entre 0.48 e 0.84. Deste conjunto destaca-se a categoria *drunkness* com coeficientes de Gini superiores a 0.8 no período considerado. Este resultado sugere que as detenções nesta categoria estão concentradas num pequeno grupo de estados americanos.

Tabela 5.1 – Valores obtidos para o coeficiente de Gini, no período de 2005 a 2011, considerando os crimes *drug abuse violations*, *driving under the influence*, *larceny-theft* e *drunkness.*

Tipo de crime	Coeficiente de *Gini*			
	2005	2007	2009	2011
DrugVio[a]	0.590	0.589	0.573	0.563
DrivInf[b]	0.516	0.522	0.507	0.494
LarcThe[c]	0,497	0,501	0.481	0.500
Drunk[d]	0,818	0.829	0,830	0.836

[a] DrugVio, *drug abuse violations*; [b] DrivInf, *driving under the influence*; [c]LarcThe, *larceny-theft*; [d] Drunk, *drunkness.*

5.3 Discussão

A caracterização efetuada no presente capítulo permite afirmar que as infrações relacionadas com a droga e álcool como sejam *drug abuse violations*, *driving under the influence*, *drunknness*, *liquor laws* e *larceny-theft* constituem as categorias de crime de maior variação no período considerado. Confrontando estes resultados com os perfis das curvas de Lorenz, podemos concluir que estas categorias são mais representativas em certos estados. Geograficamente, os estados caracterizados por uma maior variação de infrações nestas categorias constituem o grupo central. Esta observação é concordante com as observações anteriores, para os dados relativos ao número de detenções. Deste grupo central fazem parte o estados de *South Dakota*, *North Dakota*, *Montana*, *Wyoming* e *Wisconsin*. Este padrão é bem visível nas Figuras 4.1 e 4.2 que evidenciam os padrões geográficos decorrentes dos dendrogramas para o número de detenções.

Estados com grandes dimensões como sejam *California*, *Texas*, *Florida* destacam-se dos restantes, pois traduzem também uma densidade populacional superior a outros grupos. Muitas das medidas de prevenção implementadas são direcionadas para os grandes centros urbanos [25,128], sendo colocadas à margem as restantes zonas. Uma conclusão relevante é que o padrão estadual não reflete, frequentemente, o comportamento médio do país.

Capítulo 6

Comentários finais

Com o presente trabalho realizou-se uma caracterização global de informação recolhida com base nos dados UCR do FBI para os anos de 2005, 2007, 2009 e 2011. Foi considerado um conjunto de variáveis associado a tipos de crime e três perfis diferentes (número de detenções, taxa de criminalidade e fração de crimes). Tornou-se patente que o tipo de perfil escolhido para a caracterização é determinante no tipo de resultados obtido, pelo que existe um valioso grau de complementaridade entre estas representações.

Quando se utiliza para a descrição o número total de detenções, deparamo-nos com uma distribuição de estados aparentemente governada por fatores geográficos. Verifica-se que para a taxa de criminalidade a situação se altera, já que o tipo de crimes dominante na caracterização se torna mais específico. Com efeito, ao contrário do anterior, este perfil não assenta numa variação global em que o tipo de crime é irrelevante. Por último, quando é considerada a fração do crime, temos de ter em atenção que tratando-se de uma fração, o aumento de relevância de um ou mais crimes é sempre feito à custa da perda de importância de outros.

Na generalidade, verificou-se que muita da criminalidade, caracterizadora dos estados, está associada a infrações relacionadas com a droga e álcool, como *drug abuse violations*, *driving under the influence*, *drunknness* e *liquor laws*. Também *larceny-theft* surge com alguma frequência. Note-se que, para o período considerado, estas observações se mantêm válidas.

Outro aspeto relevante tem a ver com o facto de certas variáveis caracterizadoras serem prevalentes apenas em certos estados, conclusão a que se chegou através das medidas econométricas utilizadas nesta dissertação.

A representação do crime total em função da população fornece-nos a relação marcante entre estas duas componentes. Daí a maioria dos grupos formados para o perfil (número de crimes) como sejam os estados da *California*, *Texas* e *Florida* a destacarem-se dos restantes grupos.

Finalmente, é notória uma diminuição da variabilidade ao longo dos anos, associada à diminuição da frequência de crimes de tipos mais prevalentes.

Referências bibliográficas

[1] Cusson, M., *Criminologia*, 2ª Edição, Cruz Quebrada, Casa das Letras, 2007. (ISBN: 978-972-46-1620-9).

[2] McCollister, K. E., M.T. French, and H. Fang, The cost of crime to society: New crime-specific estimates for policy and program evaluation. *Drug and Alcohol Dependence*, **108**(1-2), 2010, 98-109.

[3] Rosenthal, S.S and A. Ross, Violent crime, entrepreneurship, and cities. *Journal of Urban Economics*, **67**(1), 2010, 135-149.

[4] Christens, B, and P. W. Speer, Predicting violent crime using urban and suburban densities. *Behavior and Social Issues*, **14**(1), 2005, 113-127.

[5] McDowall, D. and C. Loftin, Do US city crime rates follow a national trend? The influence of nationwide conditions on local crime patterns. *Journal of Quantitative Criminology*, **25**(3), 2009, 307-324.

[6] Martin, A. A., Testing for similarity in area-based spatial patterns: A nonparametric Monte Carlo approach. *Applied Geography*, **29**(3), 2009, 333-345.

[7] Ye, X. and L. Wu, Analyzing the dynamics of homicide patterns in Chicago: ESDA and spatial panel approaches. *Applied Geographic*, **31**(2), 2011, 800-807.

[8] Herraiz, D. S., Project Safe Neighborhoods. *Washington, DC: U.S. Department of Justice, Federal Bureau of Investigation,* **1**(1), 2004, 1-9.

[9] McGarrel, E., N. Corsaro, N. Hipple and T. Bynum., Project Safe Neighborhoods and violent crime trends in US cities: Assessing violent crime impact. *Journal of Quantitative Criminology*, **26**(2), 2010, 165-190.

[10] Braga, A. and D. Weisburd, Editors´ Introduction: empirical evidence on the relevance of place in criminology. *Journal of Quantitative Criminology*, **26**(1), 2010, 1-6.

[11] Groff, E., D. Weisburd, and S. M. Yang, Is it important to examine crime trends at a local "micro" level? : A longitudinal analysis of street to street variablity in crime trajectories. *Journal of Quantitative Criminology*, **26**(1), 2010, 7-32.

[12] Bernasco, W., Modeling micro-level crime location choice: Application of the discrete choice framework to crime at places. *Journal of Quantitative Criminology*, **26**(1), 2010, 113-138.

[13] Cherry, T. L. and J. A. List, Aggregation bias in the economic model of crime. *Economics Letters*, **75**(1), 2002, 81-86.

[14] Lu, Y. and X. Chen, On the false alarm of planar K-function when analyzing urban crime distributed along streets. *Social Science Research*, **36**(2), 2007, 611-632.

[15] Lin, W. H. and R. Dembo, An integrated model of juvenile drug use: A cross-demographic groups study. *Western Criminology Review*, **9**(2), 2008, 33-51.

[16] Keith, H., Extreme spatial variations in crime density in Baltimore County, MD. *Geoforum*, **37**(3), 2006, 404-416.

[17] John, R. H., Micro-structure in micro-neighborhoods: A new social distance measure, and its effect on individual and aggregated perceptions of crime and disorder. *Social Networks*, **32**(2), 2010, 148-159.

[18] Julie, A. P., Explaining discrepant findings in cross-sectional and longitudinal analyses: An application to U.S. homicide rates. *Social Science Research*, **35**(4), 2006, 948-974.

[19] Liska, A., and P. Bellair, Violent-crime rates and racial composition: Convergence over time. *The American Journal of Sociology*, **101**(3), 1995, 578-610.

[20] John, L. W., Racial composition, unemployment, and crime: Dealing with inconsistencies in panel designs. *Social Science Research*, **37**(3), 2008, 787-800.

[21] Sun, I., R. Triplett and R. Gainey, Neighborhoods characteristics and crime: A test of Sampson and Groves` Model of social disorganization. *Western Criminology Review*, **5**(1), 2004, 1-16.

[22] Oh, J. H., Social disorganization and crime rates in U.S. central cities: Toward an explanation of urban economic change. *The Social Science Journal*, **42**(4), 2005, 569-582.

[23] Kawachi, I., B. P. Kennedy, and R. G. Wilkinson, Crime: social disorganization and relative deprivation. *Social Science* & *Medicine*, **48**(6), 1999, 719-731.

[24] Ractcliffe, J. H., Aoristic signatures and the spatio-temporal analysis of high volume crime patterns. *Journal of Quantitative Criminology*, **18**(1), 2002, 23-45.

[25] Ractcliffe, J. H., The hotspot matrix: A framework for the spatio-temporal targeting of crime reduction. *Police Pratice and Research*, **5**(1), 2004, 05-23.

[26] Block, C. R., Is crime seasonal? Chicago: Illinois criminal justice information authority. *U.S. Department of Justice National. Institute of Justice*, 1984.

[27] Cohen, J., W. Gorr, and C. Durso, Estimation of crime seasonality: A cross-sectional extension to time series classical decomposition. *U.S. Department of Justice National. Institute of Justice,* 2003, 1-30.

[28] Cohen, L. E. and M. Felson, Social change and crime rate trends: A routine activity approach. *American Sociological Review*, **44**(1), 1979, 588-608.

[29] LeBeau, J. L., Four case studies illustrating the spatial-temporal analysis of serial rapists. *Police Studies*, **15**(1), 1992, 124-145.

[30] Brundson, C., J. Corcoran, and G. Higgs, Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems*, **31**(1), 2007, 52-75.

[31] Oatley, G. C. and B. W. Ewart, Crimes analysis software: "pins in maps", clustering and Bayes net prediction. *Expert Systems with Applications*, **25**(4), 2003, 569-588.

[32] Xue, Y. and D. E. Brown, Spatial analysis with preference specification of latent decision makers for criminal event prediction. *Decision Support Systems*, **41**(3), 2006, 560-573.

[33] Liu, H. and D. E. Brown, Criminal incident prediction using a point-pattern-based density model. *International Journal of Forecasting*, **19**(4), 2003, 603-622.

[34] Corcoran, J. J., I. D. Wilson, and J. A. Ware, Predicting the geo-temporal variations of crime and disorder. International Journal of Forecasting, **19**(4), 2003, 623-634.

[35] Sparks, C. S., Violent crime in San Antonio, Texas: An application of spatial epidemiological methods. *Spatial and Spatio-temporal Epidemiology*, **2**(4), 2011, 301-309.

[36] Butts, J. A., Violent crime rates continue to fall among juveniles and young adults. *Research and Evaluation Center*, **6**(1), 2012, 1-1.

[37] Blonigen, D. M., Explaining the relationship between age and crime: Contributions from the developmental literature on personality. *Clinical Psychology Review*, **30**(1), 2010, 89-100.

[38] Mears, D. P. and S. H. Field, Closer look at the age, peers, and delinquency relationship. *Western Criminology Review*, **4**(1), 2002, 20-29.

[39] Shicor, D., D. L. Decker, and R. M. O`Brien, The relationship of criminal victimization, police per capita and population density in twenty-six cities. *Journal of Criminal Justice*, **8**(5), 1980, 309-316.

[40] Chalfin, A. and J. McCrary, The effect of police on crime: New evidence from U. S. cities, 1960-2010.*UC Berkeley, NBER,* **11**(1), 2012, 1-84.

[41] Lo, C. C. and H. Zhong, Linking crime rates to relationship factors: The use of gender-specific data. *Journal of Criminal Justice*, **34**(3), 2006, 317-329.

[42] Lo, C. C., Sociodemographic factors, drug abuse, and other crimes: How they very among male and female arrestees. *Journal of Criminal Justice*, **32**(5), 2004, 399-409.

[43] Agha, S., Structural correlates of female homicide: A cross-national analysis. *Journal of Criminal Justice*, **37**(6), 2009, 576-585.

[44] Blackwell, B. S. and A. R. Piquero, On the relationships between gender, power control, self-control, and crime. *Journal of Criminal Justice*, **33**(1), 2005, 1-17.

[45] Vowell, P. R., A partial test of an integrative control model: Neighborhood context, social control, self-control, and youth violent behavior. *Western Criminology Review*, **8**(2), 2007, 1-15.

[46] Beaver, K. M. and J. P. Wright, The association between county-level IQ and county-level crime rates. *Intelligence*, **39**(1), 2011, 22-26.

[47] Bartels, J. M., J. J. Ryan, L. S. Urban, and L. A. Glass. Correlations between estimates of state IQ and FBI crime statistics. *Personality and Individual Differences*, **48**(5), 2010, 579-583.

[48] Canter, D., T. Coffey, M. Huntley, and C. Missen, Predicting serial killers' Home base using a decision support system. *Journal of Quantitative Criminology*, **16**(4), 2000, 457-478.

[49] Jesse, B., Does income inequality lead to more crime? A comparison of cross-sectional and time-series analyses of United States counties. *Economics Letters*, **96**(2), 2007, 264-268.

[50] Jongmook, C., Income inequality and crime in the United States. *Economics Letters*, **101**(1), 2008, 31-33.

[51] Masih, R., Modelling the dynamic interactions among crime, deterrence and socio-economic variables: evidence from a vector error-correction model, in selected papers of the MSSA/IMACS 10th biennial conference on Modelling and simulation. *Elsevier Science Publisher B. V.: Perth Australia*, **39**(3-4), 1995, 411-416.

[52] Grubesic, T. H., E. A. Mack, and M. T. Kaylen, Comparative modeling approach for understanding urban violence. *Social Science Research*, **41**(1), 2012, 92-109.

[53] McCall, P. L., K. F. Parker, and J. M. MacDonald, The dynamics relationship between homicide rates and social, economic, and political factors from 1970 to 2000. *Social Science Research*, **37**(3), 2008, 721-735.

[54] Land, K. C., K. F. Parker, and P. L. McCall, Heterogeneity in the rise and decline of city-level homicide rates 1976-2005: A latent trajectory analysis. *Social Science Research*, **40**(1), 2011, 363-378.

[55] Tang, C. F. and H. H. Lean, New evidence from the misery index in the crime function. *Economics Letters*, **102**(2), 2009, 112-115.

[56] Boggers, L. N. and J. R. Hipp, Violent crime, residential instability and mobility: Does the relationship differ in minority neighborhoods? *Journal of Quantitative Criminology*, **26**(3), 2010, 351-370.

[57] Sara, M., Alcohol, Drugs and violent crime. *International Review of Law and Economics*, **25**(1), 2005, 20-44.

[58] Zimmerman, P. R. and B. L. Benson, Alcohol and rape: An "economics-of-crime" perspective. *International Review of Law and Economics*, **27**(4), 2007, 442-473.

[59] Barber, N., Evolutionary social science: A new approach to violent crime. *Aggression and Violent Behavior*, **13**(3), 2008, 237-250.

[60] Nevin, R., Understanding international crime trends: The legacy of preschool lead exposure. *Environmental Research*, **104**(3), 2007, 315-336.

[61] Rotton, J. and E. G. Cohn, Global warming and U.S. Crime Rates. *Environment and Behavior,* **35**(6), 2003, 802-825.

[62] Yearwood, D. L. and G. Koinis, Revisiting property crime and economic conditions: An exploratory study to identify predictive indicators beyond unemployment rates. *The Social Science Journal*, **48**(1), 2011, 145-158.

[63] Cohen, L. E. and M. Felson, Social change and crime rate trends: a routine activity approach. *American Sociological Review*, **44**(1), 1979, 588-608.

[64] United States Department of Justice, F.B.I. *Crime in the United States,* 2011. September 2012, http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2011/crime-in-the-u.s.-2011.

[65] Steven, A. S., The Federal Bureau of Investigation: Its history, organization, functions and publications. *Government Publications Review* (1973), **6**(3), 1979, 213-239.

[66] Chilton, R. and J. Jarvis, Victims and offenders in two crime statistics programs: A comparison of the national incident-based reporting system (NIBRS) and the national crime victimization survey (NCVS). *Journal of Quantitative Criminology,* **15**(2), 1999, 193-205.

[67] Skogan, W. G., Measurements problems in official and survey crime rates. *Journal of Criminal Justice,* **3**(1), 1975, 17-31.

[68] Maltz, M. D., Crime statistics: A mathematical perspective. *Journal of Criminal Justice,* **3**(3), 1975, 177-194.

[69] Bursik Jr, R. J. and H. G. Grasmick, The use of multiple indicators to estimate crime trends in American cities. *Journal of Criminal Justice,* **21**(5), 1993, 509-516.

[70] O`Brien, R. M., Comparing detrended UCR and NCS crime rates over time: 1973-1986. *Journal of Criminal Justice*, **18**(3), 1990, 229-238.

[71] O`Brien, R. M., Detrended UCR and NCS crime rates: Their utility and meaning. *Journal of Criminal Justice,* **19**(6), 1991, 569-574.

[72] Maltz, M. D. and J. Targonski, A note on the use county- level UCR data. *Journal of Quantitative Criminology,* **18**(3), 2002, 297-318.

[73] RStudio version 2.15.2, Copyright 2012, The R Foundation for Statistical Computing (ISBN: 3-900051-07-0), fonte: http://www.rstudio.com/(consultado em: 2012-10-26).

[74] RStudio, fonte: http://www.rstudio.com/ide/screenshots/(consultado em: 2012-10-19).

[75] Esbensen, K. H., Multivariate data analysis-in pratice: an introduction to multivariate data analysis and experimental design. *Aalborg University, Esbjerg* (5th Edition), 2002.

[76] Tan, P. N., M. Steinbach, and V. Kumar, Introduction to data mining. *Pearson Addison-Wesley*, 2006.

[77] Brereton, R. G., Chemometrics: data analysis for the laboratory and chemical plant. *Wiley*, 2003.

[78] Wold, S., Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems*, **30**(1), 1995, 109-115.

[79] Mutihac, L. and R. Mutihac, Mining in chemometrics. *Analytic Chimica Acta*, **612**(1), 2008, 1-18.

[80] Wold, S., K. Esbensen and P. Gelati, *Chemometrics Intelligent Laboratory,* **2**(1), 1987,

[81] Esbensen, K. and P. Gelati, The start and early history of chemometrics: selected interviews. Part 2. *Journal of Chemometrics*, **4**(6), 1990, 389-412.

[82] Steven D, B., Has the chemometrics revolution ended? Some views on the past, present and future of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, **30**(1), 1995, 49-58.

[83] Gemperline, P., Pratical guide to chemometrics. 2006: CRC/Taylor & Francis.

[84] Philip K, H., The evolution of chemometrics. *Analytic Chimica Acta*, **500**(1-2), 2003, 365-377.

[85] Martens, H. and M. Martens, Multivariate analysis of quality: an introduction.*Wiley*, 2003.

[86] Andrew, J. and T. Hancewicz, Rapid analysis of Raman image using two-way multivariate curve resolution. *Applied Spectroscopy*, **52**(6), 1998, 797-807.

[87] Liu, S., S. Kokot and G. Will, Photochemistry and chemometrics-An overview. *Journal of Photochemistry and Photobiology C: Photochemistry Reviews*, **10**(4), 2009, 159-172.

[88] Roggo, Y., P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond and N. Jent, A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of Pharmaceutical and Biomedical Analysis*, **44**(3), 2007, 683-700.

[89] Ni, Y. and S. Kokot, Does chemometrics enhance the performance of electroanalysis? *Analytica Chimica Acta*, **626**(2), 2008, 130-146.

[90] Piacenti da Silva, M., O. Zucchi, A. Silva, M. Poletti, Discriminant analysis of trace elements in normal, benign and malignant breast tissues measured by total reflection X-ray fluorescence. *Spectrochimica Acta Part B: Atomic Spectroscopy*, **64**(6), 2009, 587-592.

[91] Madsen, R., T. Lundstedt, and J. Trygg, Chemometrics in metabolomics – A review in human disease diagnosis. *Analytica Chimica Acta*, **659**(1-2), 2010, 23-33.

[92] Romualdi, C., S. Camparo, D. Campagno, B. Celegato, N. Cannata, S. Toppo, G. Valle and G. Lanfranchi, Pattern recognition in gene expression profiling using DNA array: A comparative study of different statistical methods applied to cancer classification. *Human Molecular Genetics*, **12**(8), 2003, 823-836.

[93] Engreitz, J., B. Daigle, J. Marshall and R. Altman, Independent component analysis: Mining microarray data for fundamental human gene expression modules. *Journal of Biomedical Informatics*, **43**(6), 2010, 932-944.

[94] Giacomino, A., O. Abollino, M. Malandrino and E. Mentasti, The role of chemometrics in single and sequential extraction assays: A Review. Part II. Cluster analysis, multiple linear regressions, mixture resolution, experimental design and other techniques. *Analytic Chimica Acta*, **688**(2), 2011, 122-139.

[95] Jain, A. K., M. N. Murty, and P, J. Flynn, Data clustering: a review. *ACM Computing Surveys*, **31** (3), 1999, 264-323.

[96] Processo de agrupamento, fonte: http://www.steema.com/tags/cluster (consultado em: 2013-04-22).

[97] Kaufman, L. and P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis. *John Wiley and Sons, Ltd.* 2008, 1-67.

[98] Downs, G. and J. M. Barnard, Clustering methods and their uses in computational chemistry, in: K.B. Lipkowitz, D.B. Boyd (Eds.). *Reviews in Computational Chemistry*, *Wiley*, *United Kingdom*, **18**(1), 2002, 1–40.

[99] Steinbach, M., L. Ertoz and V. Kumar, Challenges of clustering in high dimensional data. *University of Minnesota Supercomputing Institute Research Report*, **213**(1), 2003, 1–33.

[100] Daszykowski, M., B. Walczak and D. L. Massart, Density-based clustering for exploration of analytical data, *Analytical Bioanalytical Chemistry*, **380**(1), 2004, 370–372.

[101] Almeida, J.A.S., L.M.S. Barbosa, A.A.C.C. Pais and S.J. Formosinho, Improving hierarchical cluster analysis: A new method with outlier`s detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems*, **87**(1), 2007, 208-217.

[102] Campanella, L., G. De Angelis and G. Visco, Chemometrics investigation of the efficiency of different TiO_2 based catalysts as principal components of TOC photochemical under development. *Analytical and Bioanalytical Chemistry*, **376**(4), 2003, 467-475.

[103] Kokot, S., M. Grigg, H. Panayiotou and T. Phuong, Data interpretation by some common chemometrics methods. *Electroanalysis*, **10**(16), 1998, 1081-1088.

[104] Davies, A. M. C. and T. Fearn, Back to basics: the principles of principal component analysis. *Spectroscopy Europe*. Tony Davies Column. 2005.

[105] Adams M.J., Chemometrics and statistics: Multivariate classification techniques. *Elsevier: Australia*, 2005, 21-26.

[106] Jolliffe, I. T., Principal component analysis 2nd edition. *Springer series in statistics*, 2002.

[107] Barber C. B., D. P. Dobkin and H. Huhdanpaa, The quickhull algorithm for convex hulls. *ACM Transactions Mathematical Software*, **22**(4), 1996, 469-483.

[108] Lerman, R. and S. Yitzhaki, Improving the accuracy of estimates of Gini coefficients, Journal of Econometrics, **42**(1), 1989, 43-47.

[109] Barreto, F., A. Menezes, R. Dantas, Entendendo o índice de Gini. *Instituto de Pesquisa e Estratégia Económica do Ceará*, **1**(1), 2009, 2-8.

[110] Ogwang, T. and U. L. Rao, Hybrid models of the Lorenz curve. *Economics Letters*, **69**(1), 2000, 39-44.

[111] Curve Lorenz, fonte: http://economicsconcepts.com/lorenz_curve_and_gini_coefficient.htm (consultado em: 2013-05-15).

[112] Nolan III, J. J., Establishing the statistical relationship between population size and UCR crime rates: Its impact and implication. *Journal of Criminal Justice*, **32**(6), 2004, 547-555.

[113] Ackerman, W. V., Socioeconomic correlates of increasing crime rates in smaller communities. *The Professional Geographer*, **50**(3), 1998, 372-387.

[114] Archer, D. and R. Gartner, Violence and crime in cross-national perspective. *Yale University Press*, 1987.

[115] Conklin, J. E. Criminology. *New York*: *Macmillan Publishing Co*. 1981.

[116] Ousey, G. C., Explaining regional and urban variation in crime: a review. *The nature of crime: continuity and change*. Department of Justice, Office of Justice Programs, National Institute of Justice, Washington DC, **1**(1), 2000, 261-308.

[117] Tittle, C. R., Sanctions and social deviance: the questions of deterrence. *Praeger*. 1980.

[118] Land, K. C., P. L. McCall and L. E. Cohen, Structural covariates of homicide rates: are there any invariances across time and social space? *American Journal of Sociology*, **95**(4), 1990, 922-963.

[119] McCall, P. L., C. L. Kenneth and L. E. Cohen, Violent criminal behavior: is there a general and continuing influence of the south? *Social Science Research*, **21**(3), 1992, 286-310.

[120] Berman, Y., Size of population and juvenile delinquency in cities in Israel. *Criminology*, **11**(1), 1973, 105-114.

[121] Krohn, M. D., L. Lanza-Kaduce and R. L. Akers, Community context and theories of deviant behavior: an examination of social learning and social bonding theories. *Sociological Quarterly*, **25**(3), 1984, 353-371.

[122] Sandim, H. R. Z., Profile of social disadvantage in the 100 largest cities of the United States, 1980 to 1990/1993. *Cities*, **15**(5), 1998, 317-326.

[123] Braithwaite, J., Crime, shame and reintegration. *Cambridge University Press*, 1999.

[124] Gottfredson, M. R. and T. Hirschi, A general theory of crime. *Stanford University Press*, 1990.

[125] Siegel, L. J., Criminology. *Wadsworth/Thomson Learning*, 2003.

[126] Rotolo, T. and C. Tittle, Population size, change, and crime in U.S. cities. *Journal of Quantitative Criminology*, **22**(4), 2006, 341-367.

[127] Chamlin, M., J. Cochran, An excursus on the population size-crime relationship. *Western Criminology Review*, **5**(2), 2004, 119-130.

[128] Ramos, L., Interpretando variação dos índices de desigualdade de Theil. *Planeamento Económico*, **20** (3), 1990, 479-488.

Anexo

Anexo 1

- **Definição das ofensas segundo a base de dados *Uniform Crime Reporting* do FBI**

  As infrações designadas como do tipo I são as que a seguir se descriminam.

Criminal homicide

a) ***Murder and nonnegligent manslaughter****:* Consiste no assassinato intencional de um ser humano por outro. Estão incluídas as mortes causadas por negligência, tentativas de homicídio e assaltos com o objetivo de matar. Suicídios e mortes acidentais são excluídos. O programa classifica os homicídios justificáveis de forma separada e limita a sua definição a (1) a morte de um criminoso por um polícia no cumprimento do dever ou (2) a morte de um criminoso, durante a prática de um crime, por um cidadão.

b) ***Manslaughter by negligence***: é definido como a morte de outra pessoa por negligência grave. A morte de pessoas devido à sua própria negligência, mortes acidentais não resultantes de negligência grosseira e fatalidades originadas pela condução não se incluem nesta categoria.

Forcible rape — é definido como o conhecimento carnal de uma mulher à força e contra a sua vontade. Estão incluídos estupros pela força e tentativas ou assaltos para violações, independentemente da idade da vítima. Situações de infrações legais (não recorrendo a força – vitima menor de idade) são excluídas.

Robbery — consiste em tomar ou tentar tirar algo de valor, que esteja ao cuidado ou no controle de uma pessoa ou pessoas, pela força, ameaça de força ou violência e /ou introduzindo medo na vítima.

Aggravated assault — é definido como um ataque ilegal, de uma pessoa sobre outra, com o objetivo de infligir lesões corporais graves ou agravadas. Este tipo de ataque é normalmente acompanhado pelo uso de uma arma ou por meios suscetíveis de produzir morte ou uma grande lesão corporal. As agressões simples encontram-se excluídas.

Burglary (breaking or entering) — consiste na entrada ilegal numa estrutura para cometer um crime ou um roubo. A tentativa de entrada forçada encontra-se incluída.

Larceny-theft (except motor vehicle theft) — é definido como a toma ilegal e o transporte de um bem que se encontra na posse de outrem. Exemplos deste tipo de ofensa são: roubos de bicicleta, peças e acessórios de veículos automóveis, furtos ou o roubo de qualquer artigo ou bem que não é

conduzido pela ação da força, violência ou fraude. As tentativas de roubo estão incluídas. Os desfalques, jogos de confiança, falsificações, fraudes, cheques, etc. não estão incluídos nesta categoria.

Motor vehicle theft — consiste no roubo ou tentativa de roubo de um veículo motorizado.
Um veículo motorizado é por definição um que é auto movível e usa estradas e não carris para se locomover. As lanchas, equipamentos de construção, aviões e equipamentos agrícolas estão excluídos desta categoria.

Arson — é definido como qualquer queima ou tentativa de queima, intencional ou mal-intencionada, com ou sem intenção fraudulenta, de bens pessoais ou de outra pessoa tais como uma casa de habitação, edifícios públicos, veículos automóveis, aviões, etc.

As infrações designadas do tipo II, para as quais apenas são recolhidos os dados das detenções, descriminam-se de seguida.

Other assaults (simple) — assaltos ou tentativas de assalto onde não foi usada qualquer arma e a vítima não sofreu uma lesão grave ou experimentou qualquer lesão. Estão incluídos comportamentos de perseguição, intimidação, coerção e rituais iniciáticos.

Forgery and counterfeiting — consiste na alteração, cópia ou imitação de algo, sem autoridade ou direito, com a intenção de enganar ou defraudar, fazendo passar a cópia ou o item alterado pelo que é original ou único. Também engloba a venda, a compra ou a posse de algo que tenha sido copiado ou imitado com a intenção de enganar ou defraudar. Incluí a forma tentada.

Fraud — definida como a perversão intencional da verdade com o propósito de induzir outra pessoa ou entidade, tendo como base a confiança gerada, a entregar algo de valor ou a conceder um direito legal. Consiste na conversão fraudulenta e obtenção de dinheiro sob falsos pretextos. Os jogos de confiança e os cheques sem fundos, excepto falsificações, estão incluídos nesta categoria.

Embezzlement — consiste na apropriação ilegal ou uso inadequado, por parte de um agressor, para o seu uso próprio de dinheiro, propriedade, ou qualquer outra coisa de valor confiado ao seu cuidado, custódia ou controlo.

Stolen property: buying, receiving, possessing — compra, receção, posse, venda, ocultação, ou transporte de qualquer propriedade com o conhecimento de que esta foi tomada ilegalmente quer por roubo, fraude, peculato, furto, etc. Incluí a forma tentada.

Vandalism — destruir ou danificar, de forma intencional ou maliciosa, qualquer propriedade pública ou privada, real ou pessoal, sem o consentimento do proprietário ou pessoa que tenha o controlo ou a custódia, através de ações que impliquem cortar, rasgar, quebrar, marcar, pintar, desenhar, sujar, ou qualquer outra ação especificada na lei local. Também se incluem tentativas.

Weapons: carrying, possessing, etc. — violação de leis ou decretos que proíbam o fabrico, compra, venda, transporte, posse, ocultação ou uso de armas de fogo, instrumentos cortantes, explosivos, dispositivos incendiários ou outras armas mortais. Incluí a forma tentada.

Prostitution and commercialized vice — consiste na promoção ilegal ou participação em atividades sexuais com fins lucrativos, incluindo as tentativas. Procurar clientes ou transporte de pessoas para fins de prostituição; possuir, administrar ou gerir uma habitação ou outro estabelecimento com a finalidade de proporcionar um lugar onde a prostituição é realizada; ou de qualquer outra forma assistir ou promover a prostituição.

Sex offenses (except forcible rape, prostitution, and commercialized vice) — crimes contra a castidade e decência moral. O incesto e o atentado ao pudor fazem parte desta categoria. Também se incluem tentativas.

Drug abuse violations — Violação das leis que proíbem a produção, distribuição e/ou uso de determinadas substâncias controladas. Incluem o cultivo ilegal, o fabrico, a distribuição, a venda, o uso, a posse, o transporte ou importação de qualquer droga ou narcótico.
Contabilizam as detenções por violações de leis locais e estaduais, especificamente as relacionadas com a posse ilegal, venda, uso, cultivo, produção e confeção de estupefacientes. As categorias de drogas a controlar são as seguintes: ópio ou cocaína e seus derivados (morfina, heroína, codeína); maconha; narcóticos sintéticos (narcóticos fabricados que podem causar dependência real – demerol, metadona) e medicamentos não narcóticos perigosos (barbitúricos, benzedrina).

Gambling — definido como a aposta ilegal recorrendo ao uso de dinheiro ou qualquer outra coisa de valor. Inclui: ajudar, promover ou explorar um jogo de azar ou algum outro jogo; possuir ou

transmitir informações sobre apostas; fabricar, vender, comprar ou possuir qualquer equipamento, dispositivos ou mercadorias de jogo; adulterar o resultado de um evento desportivo ou competição para ficar com vantagem no jogo.

Offenses against the family and children — consistem em atos ilícitos, não violentos, realizados por um membro da família ou por um responsável legal, que ameaçam o bem-estar físico, mental, económico ou moral de um outro membro da família e que não são classificados como outros crimes tais como assalto e ofensas sexuais. Também se incluem tentativas.

Driving under the influence — conduzir, operar um veículo motorizado ou um veículo de transporte comum enquanto se encontrar incapaz física e mentalmente como resultado do consumo de bebidas alcoólicas ou pelo uso de drogas e estupefacientes.

Liquor laws — violação das leis ou decretos estaduais que proíbam a produção, a venda, a compra, o transporte, a posse ou o uso de bebidas alcoólicas. Não inclui a condução sob o efeito do álcool. As violações federais também estão excluídas.

Drunkenness — consiste na ingestão de bebidas alcoólicas até ao ponto das faculdades mentais e da coordenação motora ficarem substancialmente prejudicadas. Não inclui a condução sob o efeito do álcool.

Disorderly conduct — Qualquer comportamento que tenda a perturbar a ordem pública ou o decoro, escandalize a comunidade ou choque o senso de moralidade pública.

Vagrancy — consiste na violação de uma ordem judicial, regulamento, portaria ou lei que exijam a remoção de pessoas das ruas ou de outras áreas específicas; proíbam as pessoas de permanecerem numa determinada área ou lugar de forma ociosa ou sem rumo; proíbam as pessoas de vaguear de lugar para lugar sem terem meios visíveis de subsistência.

All other offenses — Todas as violações de leis estaduais ou locais que não estejam identificadas de forma específica como ofensas de tipo I ou tipo II, exceto as violações ao tráfego.

Suspicion — prisão sem que tenha sido cometido um delito específico e libertação sem que tenha sido apresentada uma acusação formal.

Curfew and loitering laws (persons under age 18) — violações ao toque de recolher e às leis de vadiagem por pessoas menores de 18 anos de idade.

Runaways (persons under age 18) — limitado a jovens que foram colocados sob custódia protetora ao abrigo das disposições dos estatutos locais.