

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA APLICADA À MEDICINA E BIOLOGIA

**A Caminhada do Turista como ferramenta na
identificação de padrões**

Mônica Guimarães Campiteli

Ribeirão Preto
2007

UNIVERSIDADE DE SÃO PAULO

FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO

PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA APLICADA À MEDICINA E BIOLOGIA

**A Caminhada do Turista como ferramenta na
identificação de padrões**

Mônica Guimarães Campiteli

Tese submetida ao Programa de Pós-Graduação em Física Aplicada à Medicina e Biologia da Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo como parte dos requisitos para a obtenção do título de Doutor em Física Aplicada à Medicina e Biologia.

Orientador: Prof. Dr. Osame Kinouchi Filho

Ribeirão Preto, Maio de 2007.

AUTORIZO A REPRODUÇÃO TOTAL OU PARCIAL DESTE DOCUMENTO, POR MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Campiteli, M.G.

A Caminhada do Turista como ferramenta na identificação de padrões/ Mônica Guimarães Campiteli; orientador Prof. Dr. Osame Kinouchi Filho.

– Ribeirão Preto/SP, 2007.

128 f.

Tese (Doutorado – Programa de Pós-Graduação em Física Aplicada à Medicina e Biologia) – Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo.

Palavras Chaves: caminhada determinista, caminhada do turista, clusterização, reconhecimento de padrões, análise de imagens.

Banca Examinadora:

Agradecimentos

Manifesto meus agradecimentos mais sinceros ao meu orientador Prof. Dr. Osame Kinouchi Filho pela paciência e atenção com que me ensinou e guiou na área da Física, me ajudando a formar minhas primeiras conexões de um “cérebro físico”.

Ao meu co-orientador Prof. Dr. Alexandre Souto Martinez que com seu bom humor e praticidade motivou e impulsionou todas as fases deste trabalho.

Agradeço aos colegas do grupo de Física Estatística pelas trocas de experiências e de conhecimento.

Agradeço ao Pablo Diniz Batista pelo suporte dado na parte de programação e pelo apoio.

Agradeço ao Prof. Dr. Nelson Augusto Alves pelas discussões valiosas e pela atenção. Ao Prof. Dr. Alessandro Martins da Costa e ao Dr. Adriano Holanda pela prestatividade e pelas dicas sobre o LaTeX. Gostaria também de agradecer ao Dr. Marcelo Mazza pelo apoio e preocupação nos momentos delicados desta fase.

Aos meus queridos colegas de república pelo apoio e suporte sempre que precisei. Agradeço especialmente ao Rodrigo Rotta pelo carinho, apoio e prestatividade que foram muito importantes nesta fase final.

Gostaria de agradecer aos meus pais Miriam e Doire Campiteli, ao meu irmão Ricardo e aos meus tios Nilma e Luís Eduardo Gerardi pelo apoio constante, fundamental para a concretização deste trabalho.

Finalmente, agradeço a todos os amigos do departamento de Física e Matemática que participaram de alguma forma desta fase pela motivação sem dúvida valiosa para a conclusão desta tese.

À Capes pelo fomento.

Resumo

A caminhada do turista pode ser enunciada num meio desordenado formado por N pontos espalhados aleatoriamente num hipercubo de d dimensões. Um caminhante, partindo de um ponto qualquer desse meio, se desloca seguindo a regra determinista de dirigir-se sempre ao ponto mais próximo que não tenha sido visitado nos últimos μ passos. Esta dinâmica de movimentação leva a trajetórias formadas por uma parte inicial transiente de t pontos, e uma parte final cíclica de p pontos. As trajetórias obtidas são altamente dependentes da configuração do meio. Este cenário sugere que este modelo possa ser usado como uma ferramenta de reconhecimento de padrões em conjuntos de dados. O objetivo desta tese é mostrar que as propriedades da caminhada do turista permitem a sua utilização na caracterização e exploração de diversos tipos de sistemas. Aplicamos o modelo descrito em dois tipos distintos de sistemas, sistemas contínuos e redes regulares, estudando suas propriedades em função de parâmetros como tamanho do sistema, valor de memória (μ), condições de contorno e regras de movimentação. Finalmente, propomos e exploramos duas novas metodologias de reconhecimento de padrões baseadas nesta caminhada. A primeira consiste de um algoritmo de análise de imagens para caracterizar texturas que utiliza os resultados da matriz conjunta $S(t, p)$ que carrega as informações sobre todas as trajetórias obtidas, reduzindo sua dimensionalidade e permitindo a classificação eficiente de diferentes classes de imagens por um algoritmo de análise discriminante. O diferencial desta metodologia está em sua capacidade de extrair da imagem as informações presentes em diversas escalas simultaneamente. A segunda metodologia é um algoritmo de agrupamento de dados não supervisionado que considera cada atrator formado num dado valor de μ como um agrupamento natural e tem como resultado final uma árvore hierárquica geral, onde os grupos se conectam conforme se aumenta o valor de μ . Os resultados desta metodologia comparam-se em eficiência aos resultados obtidos pela metodologia tradicional para os dados testados e, entre as vantagens obtidas, podemos citar (i) independência de uma métrica relacionando os elementos do conjunto, já que trabalha apenas com uma matriz de vizinhanças, (ii) respeito à estrutura natural embutida no conjunto de dados, gerando uma árvore geral ao invés de uma árvore binária e (iii) a representação de maneira idêntica de conjuntos que sofreram transformação de escala devido à independência de uma métrica.

Palavras Chaves: caminhada determinista, caminhada do turista, clusterização, reconhecimento de padrões, análise de imagens.

Abstract

The tourist walk is defined in a disordered environment characterized by N points randomly distributed in a d -dimensional hypercube. Leaving from a given point, a walker moves according to the deterministic rule of going to next point not visited in the last μ time steps. This dynamics leads to trajectories consisting in a transient part of t points e a final cyclic part of p points. The obtained trajectories are strongly dependent on the configuration of points. This described scenario suggests that the model can be treated as a tool for pattern recognition. The aim of this thesis is to demonstrate that the tourist walk's properties allow for its use in the characterization and exploration of various kinds of systems. We have applied the model in two distinct kinds of systems - continuous systems and regular networks and studied its properties as a function of the following parameters: system size, memory (μ), boundary conditions and movimentation rule. Eventually we have proposed and explored two new pattern recognition methodologies based on this deterministic walk. The first one consists of an image analysis algorithm to characterize textures that makes use of the joint matrix $S(t, p)$ which carries the data about all trajectories obtained, reducing its dimensionality and allowing an efficient classification of different classes of images by a discriminant analysis algorithm. Its distinctive feature is its ability to extract informations in all scales from an image simultaneously. The second methodology proposed is a non-supervised clustering algorithm that considers each attractor in a given μ as a natural cluster. Its final result is a general hierarchical tree where groups coalesce as μ is increased. The results obtained with this methodology are comparable in efficiency with the results obtained with the tradicional method for the datasets tested. Among the advantages presented we can cite (i) independence from a metrics relating the elements since it works only with a neighborhood ranking table, (ii) respect for the natural structure hidden in the dataset, generating a general tree instead of a binary one and (iii) the representation of two sets transformed by scale in an identic manner due to the independence from a metrics.

Keywords: deterministic walk, tourist walk, clustering, pattern recognition, image analysis.

Sumário

Lista de Siglas	p. 17
Lista de Símbolos	p. 19
Lista de Figuras	p. 21
1 Introdução	p. 29
2 Caminhadas Deterministas	p. 33
2.1 A Caminhada da Rainha Vermelha	p. 34
2.2 A Formiga de Langton	p. 36
2.3 Redes de variáveis lógicas	p. 36
2.4 Mapeamento aleatório de Derrida-Flyvbjerg	p. 37
2.4.1 Mapeamento aleatório de ligações simétricas	p. 38
2.5 A caminhada do turista	p. 38
2.5.1 O Turista Preguiçoso ($\mu = 0$)	p. 41
2.5.2 O Turista sem memória ($\mu = 1$)	p. 41
2.5.3 O Turista com memória ($1 < \mu < N$)	p. 42
2.6 Conclusões	p. 43
3 Aspectos numéricos da caminhada do turista	p. 45
3.1 Tabela de vizinhança	p. 45
3.2 Condição de contorno	p. 45
3.3 A Caminhada	p. 47

3.4	Tratamento estatístico	p. 48
4	A caminhada do turista em meio contínuo	p. 51
4.1	Sistemas unidimensionais	p. 51
4.1.1	Distribuição de tempos de transiente	p. 52
4.1.1.1	Condição de contorno periódica	p. 56
4.1.1.2	Condição de contorno aberta	p. 60
4.1.2	Distribuição de ciclos	p. 62
4.1.2.1	Condição periódica de contorno	p. 64
4.1.2.2	Condição aberta de contorno	p. 65
4.2	Caminhadas no espaço bidimensional	p. 68
4.2.1	Distribuição de tempos de transiente	p. 69
4.2.2	Distribuição de ciclos	p. 72
4.2.3	Condição aberta de contorno	p. 76
4.3	Conclusões	p. 76
5	A caminhada do turista em redes regulares	p. 81
5.1	Redes aleatórias	p. 83
5.1.1	Regra de caminhada (i) - caminhante se move na direção de menor gradiente de energia	p. 83
5.1.1.1	Distribuição de tempos de transiente	p. 84
5.1.1.2	Distribuição de ciclos	p. 86
5.1.1.3	Transiente versus período	p. 90
5.1.2	Regra de caminhada (ii) - caminhante se move na direção de maior gradiente de energia	p. 90
5.1.3	Regra de caminhada (iii) - caminhante se move na direção do sítio de menor energia	p. 93
5.2	Redes estruturadas - texturas	p. 95

5.3	A Caminhada do turista como metodologia de análise de imagens . . .	p. 100
5.3.1	Motivação	p. 100
5.3.2	Experimentos	p. 101
5.3.3	Análise Discriminante Flexível (ADF)	p. 101
5.3.4	Resultados	p. 102
5.4	Conclusões	p. 106
6	A caminhada do turista como algoritmo de reconhecimento de padrões	p. 107
6.1	Algoritmos de agrupamentos de dados	p. 108
6.1.1	Métodos hierárquicos	p. 108
6.1.2	Método baseado na caminhada do turista	p. 110
6.2	Os conjuntos de dados	p. 114
6.3	Resultados	p. 115
6.4	Conclusões	p. 119
7	Conclusão	p. 121
	Referências	p. 125

Lista de Siglas

DFM Departamento de Física e Matemática

FAMB Física Aplicada à Medicina e Biologia

FFCLRP Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto

USP Universidade de São Paulo

cpc condição periódica de contorno

cac condição aberta de contorno

DAc Distribuição Acumulada complementar

ADF Análise Discriminante Flexível

Lista de Símbolos

N Número de pontos do meio

d Dimensionalidade do meio

μ Memória do turista

t Tempo de transiente

p Período de atrator

μ_c Memória crítica, na qual ocorre uma transição de fase na caminhada do turista

$\mu_c^{(cpc)}$ Memória crítica para sistemas contínuos com condição de contorno periódica

$\mu_c^{(cac)}$ Memória crítica para sistemas contínuos com condição de contorno aberta

$\mu_c^{(*)}$ Memória crítica obtida através de cálculo analítico

D_{ij} Matriz de distâncias

d_{ij} Distância entre os sítios i e j

$V_{N \times \mu}$ Tabela de vizinhança de um sistema de N pontos com memória μ

$P_{\mu,d}^N(t)$ Distribuição de probabilidade de tempos de transiente em um sistema de N pontos em d dimensões com memória μ

$P_{\mu,d}^N(p)$ Distribuição de probabilidade de ciclos de período p em sistemas de N pontos em d dimensões com memória μ

$P_A(p)$ Distribuição de probabilidade de um ponto pertencer à bacia de atração dos ciclos com período p

$S_{\mu,d}^N(t, p)$ Matriz conjunta de tempos de transiente e períodos de atrator

Lista de Figuras

- 1 Exemplo de um caminhante em um espaço bidimensional com $\mu = 4$ ilustrando a janela de memória como uma cauda associada ao caminhante (traço vermelho). A trajetória percorrida pelo caminhante está representada pelo traço preto contínuo e os passos futuros pelo traço pontilhado. p. 40
- 2 Figura 2(a): Exemplo de superposição de atratores em meio bidimensional para $\mu = 2$. Partindo-se do sítio A , obtém-se o ciclo $ABCD$, porém, partindo-se do sítio C , obtém-se o ciclo CBE . Figura 2(b): Exemplo de caminhada em duas dimensões ilustrando o problema de $t = 0$ ($\mu = 4$). p. 43
- 3 Condições abertas e periódicas de contorno para um meio unidimensional (Figura 3(a)) e bidimensional (Figura 3(b)). p. 46
- 4 Exemplo de um vetor de sítios visitados para uma caminhada com $\mu = 4$. Os últimos 3 pontos do vetor são comparados ao restante da trajetória até que uma seqüência idêntica seja encontrada. O período do atrator é dado pela diferença entre a posição j e a posição i do vetor. Neste exemplo, a trajetória tem um tempo de transiente igual a 3 e período igual a 6. p. 48
- 5 Exemplo de caminhadas do turista em uma dimensão. Ilustramos as trajetórias obtidas a partir de quatro diferentes pontos iniciais com memória 2 e $N = 30$ p. 52
- 6 Distribuições de tempos de transiente em função da memória μ para $N = 64, 256, 512$ e 1024 p. 53
- 7 Distribuição de tempos de transiente em função do tamanho do sistema para $\mu = 1$ (Figura 7(a)) e $\mu = 3$ (Figura 7(b)). p. 53
- 8 Ampliação dos primeiros tempos de transiente da Figura 6. Note que $t = 1$ é menos provável que os tempos adjacentes para qualquer valor de μ p. 54

-
- 9 Distribuições de tempos de transiente em função do tamanho do sistema para $\mu = 5$ e $\mu = 7$ p. 55
- 10 Distribuições de tempos de transiente em função de μ para $N = 64, 256, 512$ e 1024 em condições de contorno periódica. Podemos notar uma mudança de comportamento das curvas de distribuição num valor característico de μ que varia com o tamanho do sistema. p. 56
- 11 Ampliação das Figuras 10(b) e 10(d) mostrando a região de transição. A transição parece ocorrer em $\mu = 6$ para $N = 2^8$ e $\mu = 7$ para $N = 2^9$. Após a transição, nota-se o aparecimento de um pico em $t = N - (\mu + 1)$. Os valores dos picos estão indicados nas Figuras. p. 57
- 12 O parâmetro τ (equação 4.2) diverge nos sistemas acima da memória crítica dada pela equação 4.3. Mostramos esta divergência em função do complementar da variável α dada pela equação 4.4. Vemos aqui que, para $\alpha > 1$, ou seja, para valores de memória menores que a memória crítica, o parâmetro τ é finito. Para $\alpha < 1$, valores de memória acima da memória crítica, o parâmetro τ diverge conforme aumentamos N p. 58
- 13 A Figura mostra a distribuição de ciclos das caminhadas com $t = 49$ (Figura 13(a)), $t = 249$ (Figura 13(b)) e $t = 349$ (Figura 13(c)) para $N = 256$ e $\mu = 6$. As curvas foram obtidas a partir da distribuição conjunta, fixando-se t p. 59
- 14 Distribuição de tempos de transiente para $N = 512$ e $\mu = 1, 2, 3$ e 5 nas condições de contorno aberta e fechada. p. 60
- 15 Distribuição de tempos de transiente em função de μ e $N = 64, 256, 512$ e 1024 em condição de contorno aberta. p. 61
- 16 Ampliação das Figuras 15(b) e 15(c) mostrando a região de transição. As curvas representadas em verde indicam a transição. Os valores dos picos em $t = N - (\mu + 1)$ após a transição estão indicados nas Figuras. p. 61
- 17 Distribuição de ciclos em função de μ para $N = 256$ e 512 p. 63
- 18 Distribuição acumulada complementar de ciclos para os sistemas mostrados na Figura 17. p. 64
- 19 Distribuições acumuladas complementares para $N = 256$ e $N = 512$ e μ variando de 2 a 4. p. 64

- 20 Figuras 20(a) e 20(c): distribuição de ciclos para $N = 256$ e $N = 512$ respectivamente em condições de contorno periódicas. Figuras 20(b) e 20(d) mostram as DAC's para os mesmos sistemas anteriores. Note que as DAC's permitem a distinção entre as fases antes e depois da transição. p. 65
- 21 As Figuras 21(a) e 21(c) trazem respectivamente as distribuições de ciclos para $N = 256$ e $N = 512$ em condições de contorno abertas. As Figuras 20(b) e 20(d) mostram as DAC's para os mesmos sistemas. . . p. 66
- 22 Figuras 22(a) e 22(b): caminhadas em meio unidimensional ilustrando a fase localizada (Figura 22(a)) e delocalizada (Figura 22(b)). Figura 22(c): o pêndulo simples mostrando a analogia entre as fases localizada e delocalizada da caminhada em meio unidimensional. p. 67
- 23 Exemplos de caminhadas do turista em um meio bidimensional de $N = 500$ e memória $\mu = 5$. As trajetórias de caminhadas partindo de quatro pontos diferentes estão ilustradas com as partes transientes em vermelho e os atratores em preto. p. 68
- 24 Caminhada numa paisagem fixa, variando-se o valor de μ . A Figura mostra apenas os atratores para cada valor de μ . Figura 24(a): $\mu = 2$; Figura 24(b): $\mu = 3$; Figura 24(c): $\mu = 4$; Figura 24(d): $\mu = 5$ p. 69
- 25 Distribuição de tempos de transiente em função de μ para um sistema de $N = 512$ p. 70
- 26 Distribuição de tempos de transiente em função de N para valores de $\mu = 1, 2$ e 10 p. 71
- 27 Tempo de transiente médio em função do tamanho N do sistema para $\mu = 2$ e $\mu = 10$ p. 72
- 28 Tempo de transiente médio em função de μ para $N = 512$ p. 73
- 29 Distribuição de ciclos para um sistema bidimensional de $N = 512$ em função de μ . A Figura 29(b) mostra as respectivas distribuições acumuladas complementares. p. 73
- 30 Figura 30(a): Período médio de atratores em função de N para $\mu = 2$ e $\mu = 10$. Figura 30(b): Período médio de atratores em função de μ para $N = 512$ p. 74

- 31 A Figura mostra os tempos de transiente contra os períodos de atrator obtidos para cada trajetória percorrida num sistema de $N = 512$ e $\mu = 2$ (Figura pxt2) e $\mu = 25$ (Figura 31(b)). Note que as figuras estão mostradas em escalas diferentes. p. 75
- 32 Média do total de pontos percorridos por trajetória num sistema de $N = 512$ pontos em função de μ p. 75
- 33 Comparação entre as condições de contorno aberta (cac) e periódica (cpc) para $N = 512$ e $\mu = 1$, $\mu = 2$ e $\mu = 10$ p. 77
- 34 Comparação das distribuições acumuladas complementares de ciclos em condições de contorno aberta e periódica para $N = 512$ e $\mu = 2, 3$ e 10 . p. 78
- 35 Exemplo de rede regular quadrada. O sítio destacado em vermelho apresenta quatro primeiros vizinhos, destacados em cinza escuro e quatro segundos vizinhos destacados em tom mais claro. p. 81
- 36 Diagrama mostrando a regra de desempate utilizada nas redes regulares. Neste exemplo, o caminhante inicia o movimento no sítio 1. Pela regra de caminhada utilizada neste exemplo (regra de caminhada (i)), há 5 sítios possíveis para o próximo passo de iteração. O sítio escolhido entre estes será o primeiro sítio que encontrar-se disponível saindo em posição norte e movendo-se em direção horária, neste caso, o sítio de número 2. A mesma regra vale para os próximos 3 passos ilustrados. p. 82
- 37 Proporção de trajetórias interrompidas em função de L e μ p. 83
- 38 Distribuições de tempos de transiente em função de L para $\mu = 1, 2, 3, 5, 6$ e 7 p. 85
- 39 Distribuições de tempo de transiente em função de μ para $L = 8$ e $L = 200$. p. 86
- 40 A Figura 40(a) mostra o tempo médio de transiente em função do tamanho da rede para os diversos valores de μ . Vemos um aumento inicial e uma tendência à saturação para redes grandes. A Figura 40(b) mostra a variação de $\langle t \rangle$ em função de μ . O tempo médio de transiente nestes sistemas parece variar linearmente com μ até a memória estudada. . . p. 86
- 41 $P_A(p)$ em função de L para $\mu = 2$ e 3 . As Figuras 41(b), 41(d) mostram as distribuições acumuladas complementares desta mesma grandeza, definidas na equação 4.9. p. 87

- 42 $P_A(p)$ em função de L para $\mu = 5$ e 7 . As Figuras 42(b), 42(d) mostram as distribuições acumuladas complementares desta mesma grandeza, definidas na equação 4.9. p. 88
- 43 $P_A(p)$ em função de μ para uma rede de $L = 64$ p. 89
- 44 Figura 44(a): Distribuição de ciclos com período p em função de μ para uma rede de $L = 64$. Figura 44(b): Comparação entre as distribuições $P_A(p)$ e $P(p)$ p. 89
- 45 A Figura 45(a) mostra como $\langle p \rangle$ varia em função do tamanho da rede para os diversos valores de μ . As curvas indicam uma independência entre as duas grandezas. A queda observada em redes pequenas para $\mu > 3$ é devida ao grande número de trajetórias interrompidas observadas nestes sistemas. A Figura 40(b) mostra a variação de $\langle p \rangle$ em função de μ . O período médio destes sistemas aumenta linearmente com μ até a memória estudada. p. 90
- 46 Tempo médio de transiente em função do período do atrator, para uma rede de $L = 64$ e $\mu = 2$ (Figura 46(a)) e $\mu = 7$ (Figura 46(b)). Não há correlação entre tempo médio de transiente e período. p. 91
- 47 A Figura mostra o período do atrator contra o tempo de transiente de cada trajetória para as caminhadas realizadas em uma rede de $L = 64$ e $\mu = 2$ (Figura 47(a)) e $\mu = 7$ (Figura 47(b)). As curvas são limitadas superiormente pela reta $p + t = c$, e c depende do valor de μ p. 91
- 48 Exemplo de caminhada seguindo a regra (ii). Partindo de um sítio qualquer, o caminhante busca as regiões de maior contraste e passa a alternar entre os extremos. p. 92
- 49 Distribuições $P(t)$ (Figura 49(a)) e $P_A(p)$ (Figura 49(b)) para a regra de caminhada (ii). Vemos o colapso das distribuições para μ e $\mu + 1$ p. 92
- 50 Exemplo de caminhada seguindo a regra (ii). Partindo do sítio com valor 51, as flechas indicam a seqüência de passos até o atrator. A trajetória é a mesma para $\mu = 2$ ou $\mu = 3$ p. 93
- 51 Distribuições de $P(t)$ (Figura 51(a)) e $P_A(p)$ (Figura 51(b)) para a regra de caminhada (iii). p. 94

52	Comparação entre as distribuições de tempos de transiente e $P_A(p)$ para as três diferentes regras de caminhada e um sistema contínuo bidimensional. As redes regulares possuem $L = 32$ e o sistema contínuo, $N = 1024$. $\mu = 2$	p. 94
53	A Figura 53(a) mostra uma figura geométrica utilizada para avaliação de técnicas de reconstrução de imagens chamada fantoma. A caminhada do turista foi aplicada a esta figura e as trajetórias obtidas (trajetória transiente + atrator) estão mostradas na Figura 53(b). As caminhadas foram rodadas com μ variando de um a sete e as trajetórias foram sobrepostas.	p. 96
54	Texturas utilizadas nas estatísticas apresentadas.	p. 97
55	Distribuição de tempos de transiente das texturas da Figura 54.	p. 98
56	Figura 56(a): $P_A(p)$ das texturas para $\mu = 2$. Figura 56(b): distribuição acumulada complementar de $P_A(p)$ para os mesmos sistemas.	p. 98
57	Distribuições de tempos de transiente (Figura 57(a)) e de bacias de atração (Figura 57(b)) para as texturas (d) e (e) (Figura 54) variando-se o valor de μ	p. 99
58	Imagem gerada de uma rede aleatória com valores de pixels sorteados dentro da faixa $[0, 255]$	p. 99
59	Distribuições de tempos de transiente (Figura 59(a)) e $P_A(p)$ (Figura 59(b)) para as texturas (a), (d) e (e) (Figura 54). No mesmo gráfico estão as distribuições de tempo de transiente e $P_A(p)$ para uma rede aleatória de lado $L = 200$ e com valores dos pixels sorteados dentro da faixa $[0, 255]$	p. 100
60	Texturas utilizadas no experimento 1.	p. 102
61	Distribuição de tempos de transiente das texturas 2 e 3 para diferentes valores de μ	p. 103
62	Gráficos de superfície das distribuições conjuntas das texturas ilustradas na Figura 60. Os painéis de (a) a (f) correspondem respectivamente às Figuras 60(a) a 60(f).	p. 104
63	Resultado das análises discriminantes flexíveis do experimento 1 para os valores $\mu = 2$ (Figura 63(a)), $\mu = 3$ (Figura 63(b)), $\mu = 7$ (Figura 63(c)) e $\mu = 10$ (Figura 63(d)).	p. 105

- 64 Exemplo de um conjunto de dados e a matriz de distâncias associada a ele. No método hierárquico aglomerativo tradicional, a cada passo de iteração o par mais próximo é identificado. Este par é então colapsado em um único ponto e as linhas e colunas da matriz correspondentes a eles são atualizadas (Figura 64(b)). Este processo se repete até que haja apenas um elemento na matriz. Neste exemplo, as distâncias foram atualizadas baseando-se na média entre as distâncias de cada ponto original. p. 109
- 65 Exemplo de uma árvore hierárquica binária com dois agrupamentos identificados pela linha pontilhada. p. 109
- 66 Exemplo de um conjunto de dados bidimensional com três agrupamentos bem definidos (Figura 66(a)). As Figuras 66(b) a 66(g) mostram os atratores formados no conjunto para os seguintes valores de memória $\mu = 1, 2, 3, 5, 7$ e 15 , respectivamente. p. 111
- 67 Exemplo do processo de construção da árvore geral baseada na caminhada do turista. Na Figura 67(a) mostramos um conjunto de 20 pontos aleatoriamente distribuídos num espaço bidimensional. Realizamos as caminhadas partindo de todos os pontos do mapa e os atratores encontrados para cada valor de μ (valores mostrados apenas de 1 a 3 por motivos didáticos) estão ilustrados em cores diferentes na figura. $\mu = 1$, linha cheia verde; $\mu = 2$, linha cheia preta; $\mu = 3$, linha pontilhada vermelha. Note que conforme aumentamos o valor de μ , alguns agrupamentos se sobrepõem. Esta sobreposição corresponde a um nó no valor hierárquico (valor de μ) correspondente na árvore (Figura 67(b)). . . . p. 113
- 68 Conjunto de dados 1. 33 elementos organizados em quatro grupos de 8 elementos cada equidistantes de um ponto central. p. 114
- 69 Conjunto de dados 2. O conjunto consiste de 100 elementos dispostos em dois grupos de 50 elementos cada. Os dois grupos são idênticos na organização dos pontos, diferenciados entre si apenas pela escala. . . . p. 115
- 70 Conjunto de dados 3. Os dados originais estão num espaço de $d = 4$. A Figura mostra a projeção dos dados nas duas componentes principais. As cores representam as três diferentes categorias para este conjunto de dados. p. 116

-
- 71 Resultados obtidos para o conjunto de dados 1. Figura 71(a): árvore hierárquica resultante da técnica de *single linkage*. Figura 71(b): árvore hierárquica resultante da técnica de caminhada do turista. p. 116
- 72 Resultados obtidos para o conjunto de dados 2. Figura 72(a): árvore hierárquica resultante da técnica de *single linkage*. Figura 72(b): árvore hierárquica resultante da técnica de caminhada do turista. p. 118
- 73 Resultados obtidos para o conjunto de dados 3. Figura 73(a): árvore hierárquica resultante da técnica de *single linkage*. Figura 73(b): árvore hierárquica resultante da técnica de caminhada do turista. p. 119

1 *Introdução*

Muitos avanços nas áreas de engenharia, medicina e tecnologia neste e no último século estão intimamente ligadas às inovações nas técnicas de reconhecimento de padrões. O termo reconhecimento de padrões engloba uma ampla gama de problemas de processamento de informações de grande importância prática no reconhecimento de fala, no diagnóstico médico, no sensoriamento remoto e análise geológica, na análise e previsão de comportamento de mercado etc. Estes dados são geralmente caracterizado por sua alta dimensionalidade o que dificulta o tratamento computacional quando se deseja trabalhar diretamente sobre eles. De maneira geral, algoritmos de reconhecimento de padrão atuam extraindo do conjunto de dados, algum tipo de informação que seja capaz de descrever objetivamente as singularidades do conjunto, ao mesmo tempo em que reduz sua dimensionalidade.

Recentemente, caminhadas em conjuntos de dados como forma de exploração e caracterização do meio vêm crescendo em número de publicações, mais comumente na caracterização de grafos e redes complexas [1, 2, 3, 4]. As caminhadas mais estudadas são as caminhadas aleatórias. Neste tipo de modelo, é necessário trabalhar-se com as médias do número total de simulações, já que o resultado de cada simulação é probabilístico. As caminhadas deterministas, por outro lado, vêm experimentando um tímido, porém crescente interesse da comunidade científica e aplicações em diversas áreas têm mostrado o potencial destes modelos na caracterização de ambientes.

A caminhada determinista do turista [5, 6] é definida num espaço contínuo de d dimensões com N pontos distribuídos aleatoriamente. Partindo de um dado sítio i do espaço, a dinâmica é dada pela regra determinista de mover-se sempre para o sítio mais próximo que não tenha sido visitado nos últimos μ passos de tempo. A regra de proximidade entre vizinhos pode ser definida de acordo com o problema tratado. Originalmente, o problema foi definido segundo a distância Euclideana. As trajetórias deste modelo sempre terminam em atratores de período p após um tempo de transiente t . Esta regra de proximidade entre os sítios visitados sugere uma possível aplicação deste modelo na caracterização de

conjuntos de dados. O parâmetro μ ou, a memória do sistema atua como um fator de escala em que o caminhante explora o conjunto.

O objetivo desta tese é apresentar a caminhada determinista do turista como uma ferramenta de reconhecimento de padrões em conjuntos de dados. Exploramos as características da caminhada do turista em dois diferentes meios - meio contínuo e meio discreto - em ambiente aleatório e estruturado visando criar um arcabouço teórico que sustente as metodologias propostas.

No Capítulo 2, fazemos uma breve revisão das caminhadas deterministas mais conhecidas e detalhamos a caminhada que é tema central desta tese, a caminhada do turista. Detalhamos seus aspectos mais relevantes e os resultados conhecidos até agora. Os detalhes da implementação numérica da caminhada do turista são tratados no Capítulo 3. Abordamos algumas etapas da implementação numérica que julgamos relevantes para o entendimento dos resultados, como condições de contorno, construção da tabela de vizinhança, obtenção das trajetórias e determinação dos atratores.

No Capítulo 4, descrevemos os resultados obtidos com a caminhada do turista em meio contínuo aleatório em dimensões $d = 1$ e $d = 2$, enfatizando o papel da memória, das condições de contorno e do tamanho do sistema no comportamento das trajetórias. Os resultados são interpretados a partir das distribuições de tempos de transiente e períodos de atratores.

No Capítulo 5, apresentamos um novo tipo de caminhada do turista adaptado para redes regulares. Apresentamos os resultados das distribuições de tempos de transiente e períodos de atrator para ambientes aleatórios em função de três diferentes regras de caminhada (ou de ordenamento da vizinhança), valor de memória e tamanho da rede. Em seguida, apresentamos os resultados para uma das regras de caminhada em redes estruturadas, ou seja, que apresentam algum padrão de organização dos valores dos sítios representadas por imagens de texturas. A partir dos resultados apresentados, mostramos que a caminhada do turista pode ser eficientemente utilizada como ferramenta de reconhecimento de padrões em imagens e texturas.

Finalmente, no Capítulo 6, apresentamos uma nova metodologia de reconhecimento de padrões em conjuntos de dados multidimensionais baseada na caminhada do turista. A metodologia apresentada é baseada nos algoritmos de agrupamentos de dados de forma aglomerativa cujo produto final é uma árvore hierárquica, neste caso, uma árvore geral. Com o objetivo de avaliar de maneira contextualizada os resultados da metodologia proposta, fazemos uma breve revisão de um dos algoritmos aglomerativos mais populares na

área de reconhecimento de padrões e comparamos seus resultados com os do algoritmo proposto baseado na caminhada do turista.

As considerações finais e perspectivas de trabalhos futuros são tratadas no Capítulo 7.

2 *Caminhadas Deterministas*

Os sistemas dinâmicos discretos no espaço e/ou no tempo têm atraído a atenção da comunidade científica pois constituem modelos simples e eficientes de inúmeros fenômenos biológicos e físicos [7, 8, 9, 10]. Um sistema dinâmico possui variáveis que mudam como função de seus valores atuais. O estudo destes sistemas envolve a análise de seu espaço de fase, que é o espaço definido por todas as suas variáveis e que cobre todos os possíveis estados do sistema: cada ponto no espaço de fase representa uma das possíveis combinações das variáveis do sistema. O comportamento do sistema no tempo é representado como uma trajetória através de seu espaço de fase e o estudo da dinâmica envolve a caracterização geométrica destas trajetórias.

Em sistemas dinâmicos dissipativos existem três possibilidades para o comportamento de longo termo da trajetória do espaço de fase: ela irá parar completamente, ela irá cair em um ciclo ou ela nunca irá se fechar sobre si mesma. No primeiro caso, diz-se que o sistema evoluiu para um *ponto fixo* ou *ponto limite*. No segundo caso, para um *ciclo limite* e no terceiro, para os chamados *atratores estranhos*. Na verdade, todos os três são tipos de atratores. O nome deriva do fato de que se um sistema está num estado próximo a um atrator no espaço de fases, ele geralmente evolui para o estado ou ciclo representado pelo atrator. O conjunto de todos os pontos a partir dos quais um sistema é levado a um atrator é chamado *bacia de atração* daquele atrator. A trajetória entre o ponto inicial e o atrator é chamada *transiente*. Pode haver muitos atratores no espaço de fase de um sistema e, portanto, muitas bacias de atração. Um sistema que apresente bacias de atração é chamado sistema dissipativo pois diversas condições iniciais levam a um mesmo estado e, portanto há perda de informação sobre a condição inicial.

Tradicionalmente, existem duas classes de modelos que descrevem o movimento de objetos em um meio, ou sistemas que evoluem no tempo. A primeira é formada por processos puramente estocásticos que incluem caminhadas aleatórias clássicas. Nesta classe, o resultado de um experimento aleatório determina o próximo movimento da partícula. A segunda classe é formada por modelos puramente determinísticos em que uma regra de

movimentação pré-determinada é aplicada a cada passo de tempo para decidir para onde a partícula deve se mover.

A classe de caminhadas deterministas em meios aleatórios inclui caminhadas em meios rígidos que são sistemas gerados pelo movimento determinístico de uma partícula sobre um meio de espalhadores distribuídos aleatoriamente sobre os vértices de um grafo G [11]. Nestes sistemas, a passagem de uma partícula pode modificar o tipo de espalhador localizado no vértice visitado de acordo com alguma regra determinística. Este tipo de modelo pode ser dividido em duas subclasses. A primeira é formada pelos modelos com meio fixo em que uma partícula que chegue em um vértice g em qualquer momento sempre encontrará o mesmo tipo de espalhador. A segunda subclasse consiste de modelos que permitem uma resposta do espalhador ao objeto em movimento modificando o estado do vértice após a visita da partícula (apenas o número de encontros importa e não a direção de onde a partícula veio). Caminhadas em meios rígidos têm sido extensivamente estudadas em diversos campos e têm representado um paradigma para a propagação de sinal em meios aleatórios [11].

Ao contrário da teoria de caminhadas aleatórias que sustenta-se em solo firme, o campo das caminhadas deterministas é ainda pouco desenvolvido. Contudo, as caminhadas deterministas vêm experimentando um crescente interesse da comunidade científica e muitas e interessantes aplicações vêm sendo apresentadas nos últimos anos não só na área da física como também na sociologia, economia e no comportamento animal [9, 12]. Nesta seção, faremos uma breve revisão dos principais modelos de caminhadas deterministas em meios rígidos encontrados na literatura: a caminhada da rainha vermelha, a formiga de Langton, redes de variáveis lógicas que podem ser interpretadas como caminhadas no espaço de fase, o mapeamento aleatório estudado por Derrida e Flyvbjerg e, por fim, uma revisão mais detalhada da caminhada que é tema desta tese, a caminhada do turista.

2.1 A Caminhada da Rainha Vermelha

Considere um caminhante numa rede quadrada finita de lado L com condições de contorno periódicas. A cada sítio i da rede é vinculado um número real não-negativo h_i que pode ser encarado como a profundidade do sítio. No passo de tempo $t = 0$, assume-se que a paisagem é plana, $h_i = 0$ para todo i e o caminhante inicia sua trajetória a partir de um sítio e com uma direção escolhidos aleatoriamente. Quando o caminhante passa por um sítio, ele o modifica aumentando o valor de h_i .

A caminhada da rainha vermelha [13] surgiu como um modelo de evolução e pode ser aplicado como uma alternativa ao modelo de Lotka-Volterra de presa-predador [14]. Assim, a profundidade h_i representa o quanto o caminhante danificou ou esgotou os recursos do sítio i no passado. Quanto mais o caminhante mantém-se num dado sítio i , maior sua profundidade e este valor aumenta de uma quantidade fixa por unidade de tempo. Pode-se definir esta dinâmica como:

$$h_i(t+1) = h_i(t) + k, \quad (2.1)$$

onde k é uma constante e o caminhante se encontra no sítio i . Os sítios vazios j se recuperam a cada unidade de tempo e a função de reestabelecimento dos sítios deve ser do tipo:

$$h_j(t+1) = f[h_j(t)] < h_j(t). \quad (2.2)$$

A dinâmica do modelo é dada pela regra que garante que o caminhante sempre caminha para o sítio na sua vizinhança com menor profundidade h_j , minimizando desta forma a agressão ao meio. Esta é uma estratégia de exploração local, ou seja, o caminhante possui informação apenas de seus vizinhos adjacentes a cada passo de iteração.

Se houver dois ou mais sítios em sua vizinhança com a mesma profundidade, o caminhante adota uma regra determinística de desempate. Se nos restringirmos a regras isotrópicas (simétricas rotacionalmente), temos três possibilidades de preferência: a frente, esquerda, direita (FED); esquerda, a frente, direita (EFD); e esquerda, direita, a frente (EDF). A escolha da regra de desempate influencia o comportamento do caminhante em redes pequenas e por tempos curtos. Para redes grandes e para trajetórias longas, a regra escolhida é irrelevante [13].

Devido à regra determinista imposta, pode-se demonstrar que a ordem de profundidade dos primeiros vizinhos de um sítio i corresponde exatamente à ordem em que eles serão visitados na próxima vez em que o caminhante passar por este sítio. Podemos, portanto, substituir a variável contínua h_i por uma variável discreta que armazena informação sobre o ranque de seus n primeiros vizinhos. Disto resulta que a dinâmica do sistema é independente da forma da função $f(h)$.

Para este modelo, o espaço de fases é finito e, para tempos suficientemente grandes a caminhada é necessariamente periódica. Os resultados dependem muito dos detalhes da implementação e a periodicidade depende fortemente de L [13].

2.2 A Formiga de Langton

Uma caminhada determinista interessante é a formiga virtual proposta por C. Langton [15, 10, 16]. Este autômato celular simples é definido numa rede quadrada da seguinte forma: cada sítio (ou célula) da rede pode estar em um de dois estados e a formiga é representada por uma seta que aponta para uma das quatro direções cardeais. A cada iteração, o caminhante (formiga) se move para o sítio na direção apontada pela seta. Ele, então sofre uma deflexão de 90 graus para a esquerda ou para a direita, de acordo com o estado do novo sítio. Além disto, o estado do sítio na posição atual é alterado.

Este algoritmo produz um comportamento interessante. O caminhante inicia uma trajetória mais ou menos simétrica durante os primeiros 500 passos; ele desenvolve um padrão aparentemente aleatório durante os próximos 10000 passos e então, repentinamente passa a construir uma trajetória diagonal infinita, chamada “highway”. O aspecto interessante deste algoritmo é que este padrão de comportamento se repetiu para todas as configurações iniciais testadas até hoje.

2.3 Redes de variáveis lógicas

Considere N variáveis lógicas ou booleanas σ_i (sendo i o rótulo do sítio $i = 1, \dots, N$). A configuração do sistema no tempo t é caracterizada por um estado Σ_t :

$$\Sigma(t) = (\sigma_1(t), \sigma_2(t), \dots, \sigma_N(t)) \quad (2.3)$$

de modo que existem $\omega = 2^N$ estados possíveis para cada instante t . Se a evolução temporal do sistema for tal que cada variável σ_i no instante $t + 1$ dependa de k variáveis no instante precedente, tem-se a atualização em paralelo:

$$\Sigma(t + 1) = f[\Sigma(t)] \quad (2.4)$$

que corresponde a:

$$\sigma_i(t + 1) = f_i[\sigma_1(t), \sigma_2(t), \dots, \sigma_k(t)]. \quad (2.5)$$

A função f_i também resulta em valores lógicos. Para k variáveis, existem 2^{2^k} possíveis regras de atualização.

Em $t = 0$ determina-se o valor de k . Para cada variável σ_i , seleciona-se as k outras variáveis com quem ela irá interagir. Se esta escolha for aleatória, temos as chamadas redes de Kauffman. Para N variáveis, existem $\binom{N}{k}$ possibilidades distintas de escolha e 2^{2^k} possíveis regras, resultando em $2^{2^k} \binom{N}{k}$ possíveis realizações.

Nas redes de Kauffman todas as configurações iniciais levam a uma trajetória composta de uma parte transiente e um atrator e é, portanto uma rede dissipativa. Bhattacharjya e Liang [17] estudaram a distribuição de probabilidades dos períodos dos atratores e de tempos de transiente de redes críticas (com $k = 2$) e encontraram que estas distribuições são leis de potência com expoente fortemente dependente do tamanho da rede. Uma vez que a distribuição de períodos segue uma lei de potência, o período característico não é dado pelo valor médio, mas sim pelo valor mediano. Assim, para $k > 2$, o valor mediano de p cresce exponencialmente com N e, para $k < 2$, p mediano cresce polinomialmente com N . No primeiro caso, diz-se que o sistema é caótico, enquanto que no segundo, diz-se que o sistema é ordenado. O valor de $k = 2$ é o valor da conectividade crítica.

Kauffman argumenta que, além da seleção natural de Darwin e da mutação aleatória, a auto-organização em sistemas complexos é também responsável pela complexidade observada em sistemas biológicos [18]. A questão crucial então é: quais as características de sistemas dinâmicos que melhor modelam a evolução? Especula-se que os sistemas ótimos estão posicionados na borda da transição entre o regime caótico e o ordenado, no caso das redes de Kauffman, com $k = 2$.

2.4 Mapeamento aleatório de Derrida-Flyvbjerg

O modelo de mapeamento aleatório foi estudado por Derrida e Flyvbjerg [19] e é definido em um sistema cujo espaço de fase consiste de M pontos. A dinâmica evolui segundo as regras determinísticas:

1. Para cada ponto i do espaço de fases, associa-se um ponto $T(i)$ escolhido aleatoriamente dentre os M pontos do conjunto. Assim, para cada configuração i no tempo t , o sistema evolui para a configuração $T(i)$ no tempo $t + 1$. Observe que $T(i) = i$ é permitido.
2. $T(i)$ é independente do tempo. Assim, se após t iterações o sistema voltar para a configuração i , ele visitará novamente a configuração $T(i)$ no tempo $t + 1$. Isto

significa que, para qualquer estado inicial, a evolução temporal termina em ciclos, já que o espaço de fases é finito.

As propriedades estatísticas deste tipo de sistema foram amplamente estudadas e os resultados estão compilados em [20].

O modelo de mapeamento aleatório é o caso limite das redes de Kauffman em que n e $k \gg 1$. O ponto importante é que nas redes de Kauffman, os n spins estão conectados a k outros spins escolhidos de modo aleatório, independente e não são necessariamente diferentes. Isto implica que limite de $k \rightarrow \infty$, as imagens das dinâmicas de duas configurações estão descorrelacionadas, mesmo que as duas configurações estejam muito próximas.

2.4.1 Mapeamento aleatório de ligações simétricas

No modelo de Derrida-Flyvbjerg, as ligações entre duas configurações i e j não são necessariamente iguais. Uma variação simétrica deste modelo pode ser construída impondo-se que as ligações entre as configurações i e j e j e i sejam iguais. Este modelo de ligações aleatórias simétricas foi explorado em [21] e pode representar distâncias entre dois sítios em espaços com alta dimensionalidade.

2.5 A caminhada do turista

Nosso interesse neste trabalho é um tipo particular de caminhada determinista conhecida como a caminhada do turista. O nome foi proposto por Peter Stadler em visita ao Brasil e pode ser chamada também de caminhada determinista com repulsão parcial (no inglês: partial self-avoiding deterministic walk - PSADW). Define-se a caminhada do turista num meio formado por N pontos distribuídos num espaço d -dimensional. O caminhante parte de um ponto ou sítio inicial e desloca-se sempre para o seu vizinho mais próximo que não tenha sido visitado nos últimos μ passos de tempo, com $\mu \in [0, N - 1]$. Ilustrativamente, representamos cada ponto do universo do caminhante como uma cidade e o caminhante em si como um turista que deseja visitar o maior número possível de cidades utilizando uma regra de otimização local. O parâmetro μ representa a memória do turista e é o fator que define a dinâmica do sistema. Esta regra dinâmica produz uma trajetória inicial transiente de t passos e termina numa parte final periódica de período p . Cada condição inicial produz uma trajetória diferente.

Apesar de sua fácil formulação e fácil implementação numérica, de acordo com a

janela de memória escolhida, a caminhada do turista pode apresentar um comportamento complexo com transientes longos e ciclos com diversos períodos. Considerando-se todos os sítios como condições iniciais, obtém-se como resultado final as distribuições de tempos de transiente $P_{\mu,d}^N(t)$ e de períodos de atratores $P_{\mu,d}^N(p)$.

Uma característica notável deste algoritmo é que o caminhante desloca-se sobre uma tabela de vizinhanças que contém apenas o ordenamento dos sítios. Desta forma, a distância entre os pontos é irrelevante. A tabela de vizinhança deve conter pelo menos os μ vizinhos mais próximos de cada ponto. É definida, portanto como $V_{N \times \mu}$, na qual cada elemento V_{ij} indica o j -ésimo ($1 \leq j \leq \mu$) vizinho do i -ésimo sítio:

$$(2.6) \quad V_{N \times \mu} = \begin{pmatrix} V_{11} & V_{12} & \cdots & V_{1\mu} \\ V_{21} & V_{22} & \cdots & V_{2\mu} \\ \vdots & \vdots & \ddots & \vdots \\ V_{N1} & V_{N2} & \cdots & V_{N\mu} \end{pmatrix}$$

É interessante notar que a tabela de vizinhanças não é simétrica, ou seja, não temos necessariamente $V_{ij} = V_{ji}$. Como exemplo, considere duas cidades A e B do mapa onde o turista caminha. Se A é o j -ésimo vizinho de B , B não é necessariamente o j -ésimo vizinho de A . Assim, com base em uma matriz de distâncias que contenha as distâncias d_{ij} entre todos os pontos de um conjunto, podemos construir a tabela de vizinhanças. Entretanto, o inverso não é verdadeiro: não é possível reconstruir a matriz de distâncias original e nem definir o ranque como uma distância formal pois, como vimos, tal grandeza não possui a simetria necessária para a definição de uma métrica.

O parâmetro μ é a janela de tempo dentro da qual, o turista não pode visitar um dado sítio (ou cidade). Podemos ilustrar o papel da memória como uma cauda associada ao caminhante (Figura 1). A auto-repulsão é limitada a esta janela e a trajetória pode se interceptar fora deste intervalo. A capacidade de exploração do meio pelo caminhante é alta e mesmo para valores baixos de memória, o turista visita grande parte do meio.

Podemos descrever a caminhada do turista como uma caminhada em meios rígidos, ou seja, uma vez visitado pelo caminhante, o estado do sítio muda de disponível para

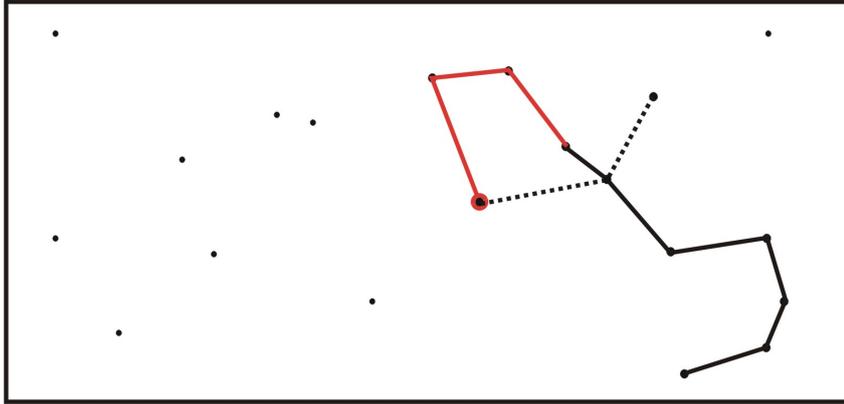


Figura 1: Exemplo de um caminhante em um espaço bidimensional com $\mu = 4$ ilustrando a janela de memória como uma cauda associada ao caminhante (traço vermelho). A trajetória percorrida pelo caminhante está representada pelo traço preto contínuo e os passos futuros pelo traço pontilhado.

proibido e o tempo de reestabelecimento do sítio é dado pelo valor de μ . Este modelo difere da caminhada da rainha vermelha por ser um modelo essencialmente construído num espaço contínuo e não discretizado.

O modelo da caminhada do turista também difere das redes lógicas (de Kauffman) por ter um espaço de fase com μd dimensões (cada ponto é dado por $\vec{X} = \{\vec{x}(t - (\mu + 1)), \dots, \vec{x}(t - 1), \vec{x}(t)\}$) em vez de N -dimensional como nessas redes. No entanto, os resultados da caminhada do turista convergem para os resultados do modelo de ligações aleatórias simétricas no limite de $d \rightarrow \infty$.

Uma importante aplicação da caminhada do turista é como um modelo de estratégias de forrageamento. De fato, Boyer e Larralde [22] propuseram um modelo de forrageamento no qual o território é composto por N fontes de alimento com tamanhos variáveis e estrutura espacial também variável. No modelo, o animal forrageador tem informação sobre a posição e tamanho de todos os sítios. O tamanho de um dado sítio i é dado por um valor aleatório k_i associado ao sítio que representa a atratividade daquele sítio. Partindo de um sítio i do mapa, o animal escolhe o próximo sítio a ser visitado através de uma regra determinista que visa minimizar a razão entre a distância percorrida até o próximo sítio e o tamanho da fonte de alimento. Sítios visitados anteriormente não são mais permitidos nesta dinâmica. Os autores sugerem, no entanto que a utilização de uma janela de memória representando o tempo de restauração da fonte de alimento represente de maneira mais natural o comportamento de animais forrageadores. Os autores obtiveram resultados comparáveis aos dados de campo com macacos aranha. Apesar de a caminhada do turista não prever um valor de atratividade do sítio em seu modelo original, esta

adaptação pode ser facilmente inserida na definição de distância estabelecida entre os sítios, sem qualquer comprometimento ao enunciado do modelo.

A seguir, detalhamos as características da caminhada do turista em algumas condições especiais.

2.5.1 O Turista Preguiçoso ($\mu = 0$)

Para $\mu = 0$, o turista não se movimenta pois o sítio mais próximo é o próprio sítio em que ele se encontra. Chamamos esta condição de turista preguiçoso. Aqui, cabe mencionar que nos primeiros trabalhos sobre a caminhada do turista [5], o parâmetro de memória era designado por τ e $\tau = \mu - 1$. Neste trabalho, qualquer referência à memória do sistema, refere-se ao parâmetro μ . Na condição $\mu = 0$, a solução é trivial, os tempos de transiente são sempre zero e cada ponto do mapa representa um atrator de período unitário. A distribuição de tempos de transiente é, portanto $P_{\mu,d}(t) = \delta_{t,0}$, onde $\delta_{i,j}$ é o delta de Kronecker definido por:

$$\delta_{i,j} = \begin{cases} 0 & \text{se } i \neq j; \\ 1 & \text{se } i = j \end{cases} \quad (2.7)$$

A distribuição de atratores, por sua vez é dada por $P_{0,d}(p) = \delta_{p,1}$. Apesar da trivialidade deste caso, ele se torna interessante como limite de baixa temperatura na caminhada estocástica do turista, tratada em [23], que não será objeto deste estudo.

2.5.2 O Turista sem memória ($\mu = 1$)

O caso onde $\mu = 1$ é chamado de caso do turista sem memória, em que o caminhante tem informação apenas de seu primeiro vizinho, ou seja, do ponto onde ele se encontra. Neste caso, ele é obrigado a se movimentar e caminhará de sítio em sítio até que encontre um par de sítios mutuamente mais próximos e a caminhada termina sempre num ciclo de período 2. Este par de pontos mutuamente mais próximos é chamado aqui de casal e se dá quando $V_{i1} = j$ e $V_{j1} = i$. A distribuição de períodos vale:

$S_{1,d}(p) = \delta_{p,2}$, ou seja, todos os ciclos têm período 2. Já a distribuição de tempos de transiente não é trivial e é característica de cada sistema.

Podemos associar uma energia dada pela distância d_{ij} a cada ligação entre os nodos i e j e, considerando uma rede dual, pensar que a caminhada se dá nas ligações entre os

sítios. Assim, a energia do caminhante sempre diminui até que ele encontre uma ligação de mínima energia (ligações entre casais), onde ficará preso.

2.5.3 O Turista com memória ($1 < \mu < N$)

Para $\mu = N - 1$, onde N é o número de pontos do mapa, a trajetória é totalmente auto-repulsiva e temos um problema do tipo caixeiro viajante com dinâmica *greedy*. A distribuição de períodos é $P_{N-1,d}(p) = \delta_{p,N}$.

Os casos em que $1 < \mu < N - 1$ diferem drasticamente dos anteriores. Neste caso, tanto a distribuição de tempos de transiente como a de períodos é característica do sistema e não trivial. A relação $p \geq \mu + 1$ é a única restrição válida genericamente.

Vale notar que o problema do turista não é um problema relacionado apenas às propriedades geométricas do meio já que os atratores surgem devido à dinâmica introduzida. Ingenuamente, poderia-se pensar que os atratores são objetos geométricos, como um agrupamento onde as distâncias entre os pontos são menores que qualquer distância aos pontos fora do agrupamento. Esta é, de fato, uma condição suficiente mas não necessária para se obter um ciclo. No exemplo da Figura 2(a), para $\mu = 2$, um caminhante parte do sítio A e encontra o ciclo de período 4 $ABCD$. Apesar do sítio E estar próximo ao atrator ($BE < BA$), ele não é visitado pois $BC < BE$ e $CD < CE$. Entretanto, se a caminhada começar do sítio C , obtém-se o ciclo de período 3 CBE . Esta degenerescência e superposição de atratores pode ser entendida notando-se que as trajetórias da Figura 2(a) estão representadas no espaço de configurações e não no espaço de fase. No espaço de fase, cada ponto corresponde a um vetor (\vec{X}_t) de μ variáveis que determinam univocamente o próximo passo:

$$\vec{X}_t = \{\vec{x}_{t-\mu}, \dots, \vec{x}_{t-1}, \vec{x}_t\} \longrightarrow \vec{x}_{t+1} \quad (2.8)$$

onde \vec{x}_t é o vetor de posição do caminhante no tempo t . No espaço de fases, as trajetórias nunca se interceptam. Apenas para $\mu = 1$, o espaço de configurações é equivalente ao espaço de fases. É interessante mencionarmos que, aparentemente o primeiro passo da caminhada do turista representa uma particularidade do cenário descrito acima já que apenas uma variável do vetor \vec{X}_t foi determinada.

$$\vec{X}_0 = \{\vec{0}, \vec{0}, \dots, \vec{x}_0\}. \quad (2.9)$$

Poderíamos contornar esta anomalia escolhendo as demais μ variáveis aleatoriamente dentro do conjunto de sítios do sistema. Para um sistema de N pontos, teríamos $\binom{N}{\mu}$ possíveis configurações iniciais. Entretanto, para $N \gg \mu$, a probabilidade de que alguma das μ variáveis sorteadas aleatoriamente esteja dentro do raio de vizinhança de \vec{x}_t tende a zero (Figura 2(b)). Assim, temos que a escolha do sítio \vec{x}_{t+1} para $t = 0$ depende apenas de \vec{x}_t independentemente da configuração inicial escolhida e, portanto este caso não configura uma particularidade da caminhada.

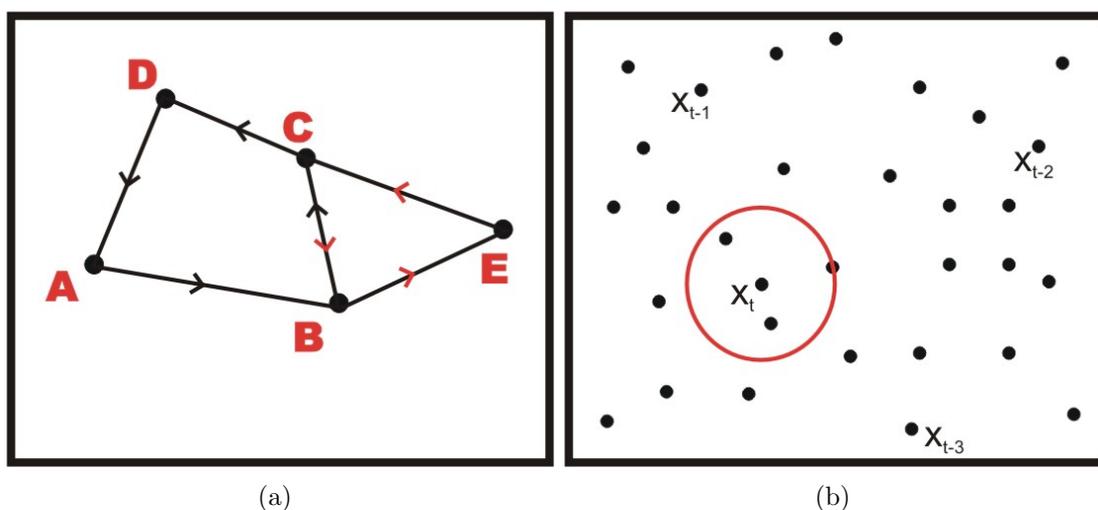


Figura 2: Figura 2(a): Exemplo de superposição de atratores em meio bidimensional para $\mu = 2$. Partindo-se do sítio A , obtém-se o ciclo $ABCD$, porém, partindo-se do sítio C , obtém-se o ciclo CBE . Figura 2(b): Exemplo de caminhada em duas dimensões ilustrando o problema de $t = 0$ ($\mu = 4$).

2.6 Conclusões

Apresentamos uma breve revisão sobre algumas das caminhadas deterministas mais conhecidas e descrevemos os aspectos da caminhada do turista em meios contínuos desordenados. As aplicações das caminhadas determinísticas como modelos de sistemas naturais vêm crescendo e o estudo deste tipo de caminhadas vem atraindo a atenção da comunidade científica. Mostramos que a caminhada do turista apresenta diversos pontos em comum com as caminhadas bem estudadas e o estudo de sua dinâmica revela novas e interessantes propriedades.

Nesta tese, apresentamos as propriedades da caminhada do turista em meios discretos, sistemas que equivalem à caminhada da rainha vermelha. Apresentamos também algumas propriedades da caminhada do turista em meios contínuos que não haviam sido exploradas

até agora.

Mostramos ainda que as propriedades da caminhada do turista abordadas aqui representam uma poderosa ferramenta para a aplicação em exploração de dados e reconhecimento de padrões.

3 *Aspectos numéricos da caminhada do turista*

Neste capítulo, detalharemos alguns aspectos da implementação numérica da caminhada do turista.

3.1 Tabela de vizinhança

A caminhada do turista é realizada totalmente sobre uma tabela de vizinhanças como a descrita na equação 2.6. Esta tabela contém os μ primeiros vizinhos de cada ponto do mapa. A ordenação da tabela é dada segundo uma regra conveniente ao problema estudado. Neste trabalho, utilizamos diferentes medidas de distância como critério para a ordenação dos pontos. Assim, definindo uma medida de distância arbitrária, constrói-se uma matriz de distâncias D_{ij} e, a partir dela obtém-se a tabela $V_{N \times \mu}$ segundo:

Para $i = 1 : N$

Para $j = 1 : \mu$

$$V_{ij} = \min(D_i);$$

onde $\mu = N - 1$ para todas as aplicações estudadas nesta tese. Nestes casos, a caminhada do turista é desenvolvida sobre um meio estático. Assim, não há necessidade de se reordenar os pontos a cada passo de tempo. A complexidade computacional deste processo é da ordem de $N \log(N)$.

3.2 Condição de contorno

Num meio limitado, os sítios situados próximo às fronteiras se relacionam com seus vizinhos de maneira diferente que os sítios localizados na região central. Podemos dizer que tais sítios sofrem um certo isolamento com relação aos demais pois têm uma região

de vizinhança achatada ou deformada. Este isolamento, denominado *efeito de borda* se acentua à medida em que se aumenta a dimensionalidade do sistema. Dizemos que este meio apresenta *condições abertas de contorno*.

Uma maneira de se tornar todos os sítios igualmente acessíveis aos demais é conectar as extremidades do meio. Diz-se então que o meio apresenta *condições periódicas (ou fechadas) de contorno*. As Figuras 3(a) e 3(b) ilustram este processo para 1 e 2 dimensões, respectivamente.

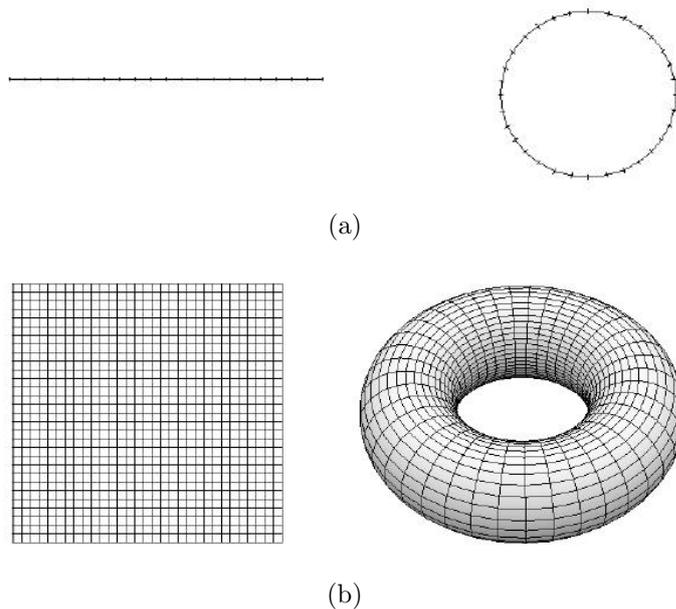


Figura 3: Condições abertas e periódicas de contorno para um meio unidimensional (Figura 3(a)) e bidimensional (Figura 3(b)).

Devemos observar que a circunferência da Figura 3(a) é geometricamente bidimensional, mas topologicamente unidimensional. Do mesmo modo, a toróide da Figura 3(b) tem três dimensões geométricas, porém é topologicamente bidimensional já que os pontos se distribuem apenas sobre sua superfície.

A distância entre dois pontos i e j quaisquer de um meio com condições de contorno periódicas e cujas coordenadas se distribuem dentro do intervalo $[0, 1)$ é dada por:

$$D_{ij} = \min(d_{ij}, 1 - d_{ij}), \quad (3.1)$$

em que d_{ij} representa uma medida arbitrária de distância entre os pontos i e j em condições abertas de contorno.

3.3 A Caminhada

Dada uma tabela de vizinhanças, esta etapa consiste em gerar as trajetórias de um caminhante que parte de cada um dos N sítios do mapa. O valor da memória é o parâmetro fornecido. Ao final desta etapa, tem-se N trajetórias, cada uma com um tempo de transiente e um período de atrator armazenados na matriz conjunta $S_{\mu,d}^N(t,p)$. Vale notar que esta matriz tem tamanho (t_{max}, p_{max}) representando o tempo máximo de transiente e período máximo de atrator que podem atingir valores muito maiores que N . O algoritmo básico da caminhada do turista está representado na Tabela 3.3.

Caminhada(matriz_de_vizinhanças, μ)

For site = 0 : N

stop = 0;

A = site;

Do

B = primeiro vizinho de A;

IF (B não visitado nos últimos μ passos anteriores)

A = B;

ELSE

B = próximo vizinho de A;

IF (Atrator formado)

Armazena Atrator;

Armazena transiente;

stop = 1;

While (stop = 0)

Detalhamos a seguir as etapas do algoritmo. Partindo-se de um sítio i do mapa:

1. Determinar o próximo sítio.

Lê-se a matriz de vizinhanças para se determinar o vizinho de ordem k ($k = 0 : \mu$) do sítio atual. Chamamos este k -ésimo vizinho de **site**. Cada vez que um sítio é visitado, o passo em que isso ocorreu é armazenado no vetor **address**. Assim, para determinar se **site** está disponível, podemos calcular a diferença entre o passo atual e o passo em que ele foi visitado pela última vez. Se esta diferença for maior que μ , o sítio está disponível, caso contrário, incrementa-se k e repete-se o teste para o próximo vizinho.

2. Mover-se para o próximo vizinho.

Neste passo, apenas armazena-se o sítio escolhido no passo anterior na próxima posição do vetor chamado **sítios_visitados**. Este vetor armazenará toda a trajetória percorrida pelo turista até que ele caia num atrator. Atualiza-se também o vetor **address**, armazenando na posição **site** o valor do passo em que ele foi visitado.

3. Verificar se o turista já está num atrator.

Esta etapa consiste em deslizar uma janela com os $\mu - 1$ últimos pontos visitados ao longo do vetor **sítios_visitados**. Se esta janela deslizante encontrar um intervalo com a mesma seqüência de pontos ao longo da trajetória, conclui-se que o caminhante já está num atrator e encerra-se a caminhada.

4. Identificar tempo de transiente e período de atrator.

Ao identificar a repetição de $\mu - 1$ passos dentro da trajetória, identifica-se o atrator. A Figura 4 ilustra como este processo ocorre. O período do ciclo é a diferença entre a posição do vetor em que a janela deslizante encontrou a repetição (posição j) e o início da janela (posição i). Uma vez determinado o período do atrator, o tempo de transiente é dado pela posição j . As grandezas obtidas são armazenadas em uma matriz conjunta $S(t, p)$.

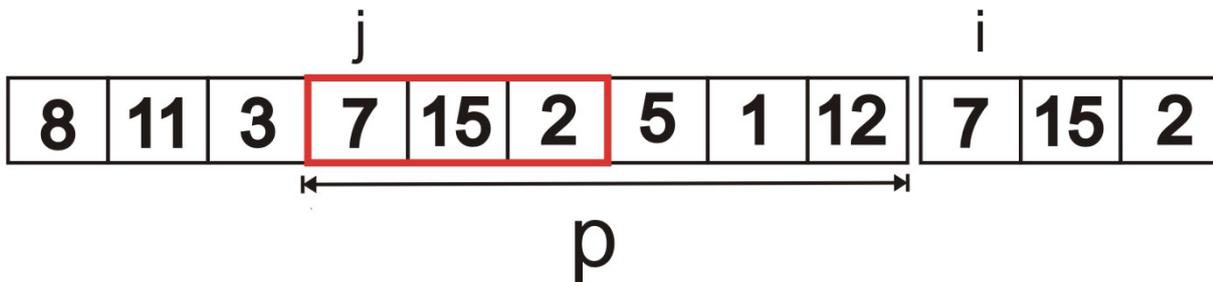


Figura 4: Exemplo de um vetor de sítios visitados para uma caminhada com $\mu = 4$. Os últimos 3 pontos do vetor são comparados ao restante da trajetória até que uma seqüência idêntica seja encontrada. O período do atrator é dado pela diferença entre a posição j e a posição i do vetor. Neste exemplo, a trajetória tem um tempo de transiente igual a 3 e período igual a 6.

3.4 Tratamento estatístico

Para as caminhadas em meios aleatórios, são gerados N_R mapas com N pontos cada cujas coordenadas são geradas aleatoriamente com a função *rand* do pacote do Borland C++ [24]. Sobre cada um dos N_R mapas gerados, são realizadas N caminhadas e a

este processo chamamos realização. A semente utilizada pelo gerador de números pseudo-aleatórios não varia, assim, as seqüências de paisagens geradas são idênticas para todas as simulações.

Para cada paisagem, contabilizam-se as grandezas de interesse somando-se os resultados das NN_R caminhadas. Neste trabalho, estudamos as distribuições de tempos de transiente, período de atrator e as distribuições pontos que levam aos atratores de período p . As distribuições de tempos de transiente ($P(t)$) e de pontos que levam aos atratores de período p ($P_A(p)$) são obtidas facilmente a partir da matriz conjunta $S(t, p)$ conforme:

$$P(t) = \sum_p (S(t, p)); \quad (3.2)$$

$$P_A(p) = \sum_t (S(t, p)). \quad (3.3)$$

Por sua vez, a matriz conjunta é obtida somando-se o número de caminhadas com tempo de transiente t e com período de atrator p obtidos em todas as realizações e dividindo pelo número total de caminhadas NN_R .

A distribuição $P_A(p)$ representa o tamanho da bacia de atração do atratores com período p . Ou seja, o número de pontos que levam a atratores com período p numa dada paisagem. Esta grandeza é conceitualmente diferente da grandeza $P(p)$, que fornece o número de atratores diferentes com período p encontrados numa dada paisagem e que é dada por:

$$P(p) = \frac{n(p)}{N^*} \quad (3.4)$$

onde $n(p)$ é o número de atratores diferentes com período p e N^* é o número total de atratores de uma dada paisagem. Note que N^* é sempre menor que N para $\mu \geq 1$. A distribuição de atratores $P(p)$ é computacionalmente mais cara que a distribuição $P_A(p)$ já que envolve a comparação entre os ciclos encontrados a cada caminhada realizada.

4 *A caminhada do turista em meio contínuo*

Neste capítulo exploramos a caminhada do turista no espaço euclidiano contínuo. Consideramos N pontos pelas coordenadas $\{x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(d)}\}$. Neste meio, definimos a distância uniformemente distribuídos dentro do hipercubo unitário $[0, 1)^d$. Cada ponto é definido pela distância entre dois pontos i e j quaisquer através da métrica:

$$d_{ij} = \sqrt{\sum_{k=1}^d (x_i^{(k)} - x_j^{(k)})^2}. \quad (4.1)$$

A dimensionalidade do sistema influencia de forma determinante as configurações possíveis das trajetórias [5] e, portanto, as estatísticas relativas a transientes e ciclos que são o objeto de estudo deste trabalho. Assim, damos enfoque à dimensionalidade do meio e estudaremos separadamente caminhadas em meios unidimensionais e em meios com $d = 2$.

Na maioria dos problemas reais de aplicação, condições de contorno abertas são mais plausíveis que condições periódicas. Já as condições periódicas de contorno facilitam um possível estudo analítico. Por isso exploramos aqui ambas as condições de contorno. Todas as estatísticas apresentadas são obtidas da média sobre 1000 realizações (paisagens). Lembramos que de cada paisagem obtém-se N trajetórias, de modo que as estatísticas apresentadas reportam efetivamente médias sobre $1000N$ condições iniciais.

4.1 Sistemas unidimensionais

Os sistemas euclidianos unidimensionais estão no extremo inferior de dimensionalidade e seu estudo tem grande importância pois são os sistemas que permitiram um estudo analítico da caminhada [25]. Devido às restrições geométricas da caminhada em uma linha (Figura 5), as estatísticas da caminhada do turista em uma dimensão apresentam

características peculiares que serão apresentadas a seguir.

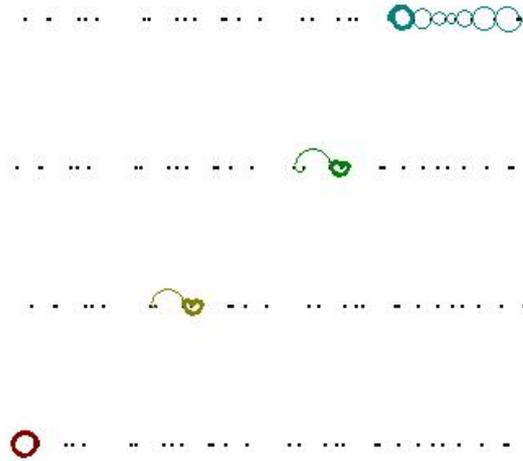


Figura 5: Exemplo de caminhadas do turista em uma dimensão. Ilustramos as trajetórias obtidas a partir de quatro diferentes pontos iniciais com memória 2 e $N = 30$.

4.1.1 Distribuição de tempos de transiente

A Figura 6 mostra a probabilidade de que um caminhante, partindo de um sítio escolhido ao acaso, percorra t passos antes de cair em um atrator. O estudo foi realizado com $N = 64, 256, 512$ e 1024 pontos. Mostramos as distribuições de tempos de transiente para os diferentes tamanhos de rede variando-se o valor de memória μ de 1 a 5. A distribuição de probabilidades de tempos de transiente parece seguir uma função exponencial do tipo:

$$P(t) \propto \exp\left(\frac{-t}{\tau}\right), \quad (4.2)$$

onde τ representa o tempo característico de transiente [26], que aumenta com o aumento do valor da memória.

Para memórias pequenas, a distribuição de tempos de transiente depende fracamente do tamanho do sistema. Na Figura 7, mostramos o colapso das distribuições de tempos de transiente para $N = 64, 256$ e 512 nos casos de $\mu = 1$ e $\mu = 3$.

A distribuição de transientes para t pequeno nos chama a atenção por não ser monotônica. Em detalhe na Figura 8, mostramos os primeiros pontos das distribuições de tempos de transiente de uma rede de $N = 512$ para os diversos valores de μ . Curiosamente, as caminhadas com $t = 1$ são menos prováveis que caminhadas com $t = 0$ (ou seja

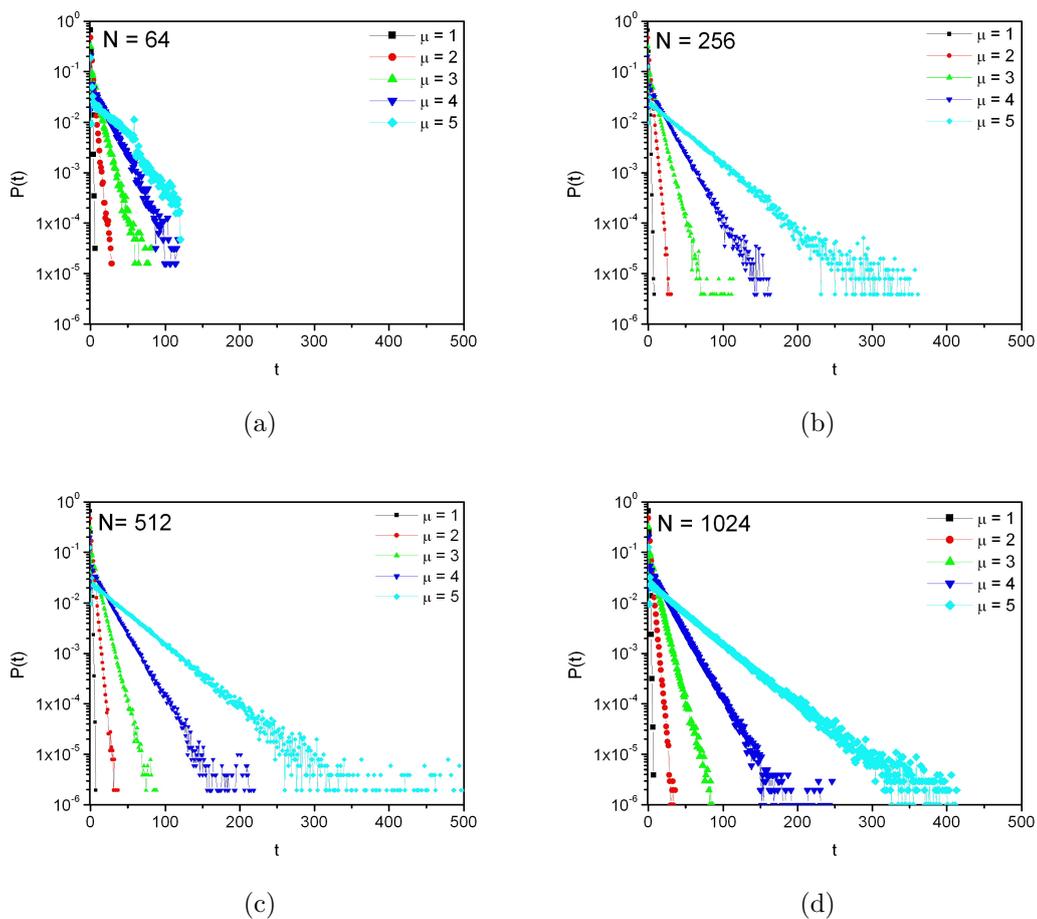


Figura 6: Distribuições de tempos de transiente em função da memória μ para $N = 64, 256, 512$ e 1024 .

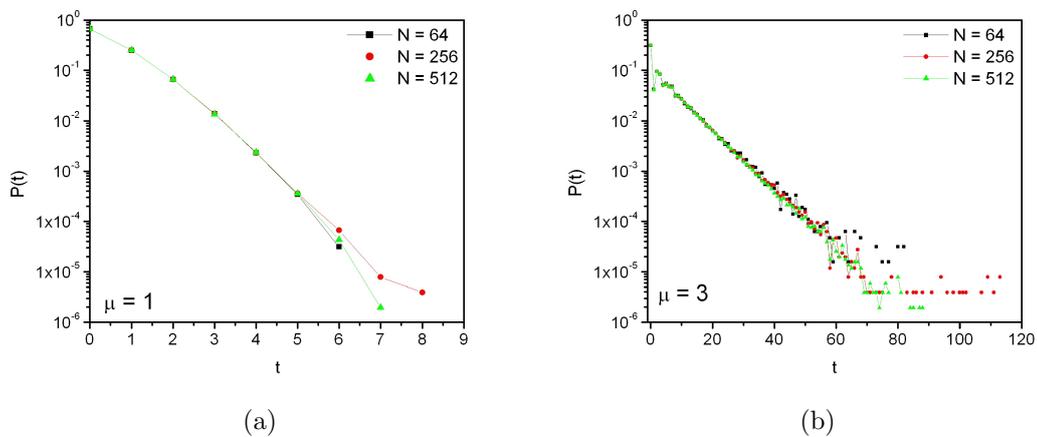


Figura 7: Distribuição de tempos de transiente em função do tamanho do sistema para $\mu = 1$ (Figura 7(a)) e $\mu = 3$ (Figura 7(b)).

o caminhante já inicia a caminhada num atrator) ou com $t = 2$. Este comportamento se repete para sistemas em maiores dimensões. De fato, este comportamento é explicado pela própria natureza determinística da caminhada. O número de sítios a apenas um passo de um atrator é limitado a uma pequena área ao redor deste atrator. Caminhadas com $t = 0$ terão até p possibilidades de condições iniciais para cada atrator (enfatizamos que alguns pontos podem pertencer a atratores diferentes dependendo do ponto em que se inicia a trajetória, ainda que isso aconteça com baixa frequência). Aparentemente, é mais fácil começar a caminhada sobre um atrator que sobre um ponto adjacente a um atrator ($t = 1$). Já caminhadas com tempo de transiente igual a dois são necessariamente mais abundantes que caminhadas com $t = 1$, já que as trajetórias das primeiras sempre sobrepõem as trajetórias dos casos com $t = 1$.

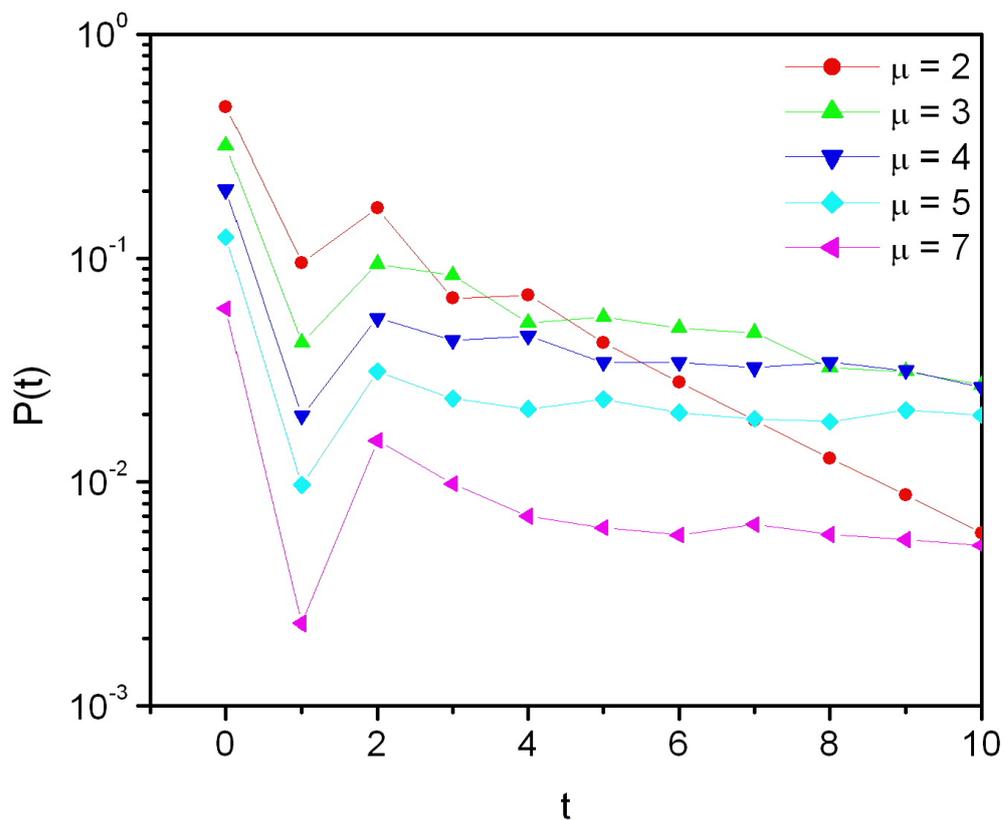


Figura 8: Ampliação dos primeiros tempos de transiente da Figura 6. Note que $t = 1$ é menos provável que os tempos adjacentes para qualquer valor de μ .

Vimos na Figura 7 que a exploração do meio pelo caminhante não depende ou depende fracamente de N para valores de memória pequenos. Por outro lado, a memória do sistema tem papel decisivo na exploração do meio pelo caminhante. Memórias pequenas acarretam

uma baixa exploração do meio e os tempos de transiente são baixos. Com o aumento do parâmetro μ , a probabilidade de ocorrência de tempos de transiente grandes, da ordem de N aumenta e a distribuição $P(t)$ passa a depender mais fortemente do tamanho do sistema. É o que vemos na Figura 9. A partir de um certo valor de μ , há uma mudança de comportamento das trajetórias que se reflete na distribuição de tempos de transiente. Uma questão pertinente aqui é acerca da existência de um valor característico de μ que produza uma transição de fase, ou seja, um valor crítico de memória a partir do qual o sistema percole, e todos os sítios sejam visitados.

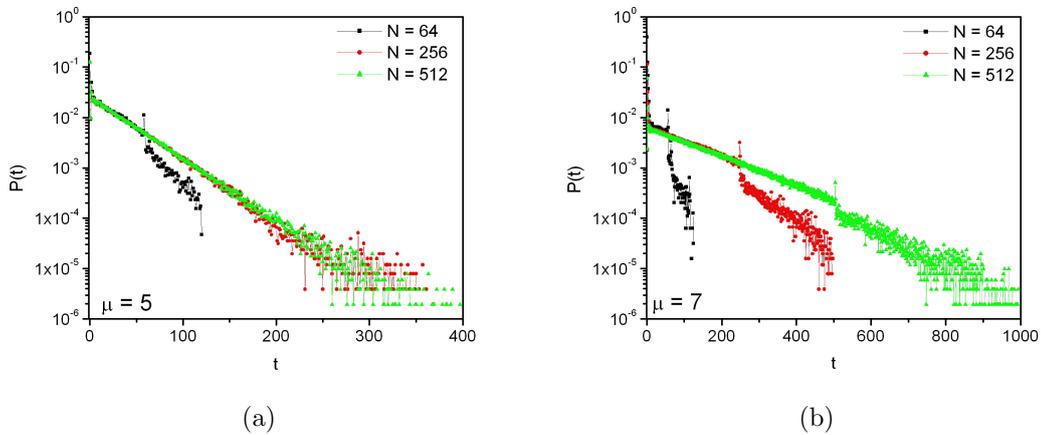


Figura 9: Distribuições de tempos de transiente em função do tamanho do sistema para $\mu = 5$ e $\mu = 7$.

Porém, antes de discutirmos a existência de transição de fase na caminhada do turista, devemos considerar as condições de contorno utilizadas. Condições de contorno abertas são mais plausíveis em algumas das aplicações consideradas nesta tese, por isso é de grande interesse considerar suas propriedades. Contudo, condições de contorno abertas impõem restrições geométricas devido aos chamados efeitos de borda. Os pontos próximos às extremidades do sistema interagem com seus vizinhos de forma diferente dos pontos situados próximo ao centro e isto influencia de maneira substancial na dinâmica do caminhante. Veremos a seguir as características das distribuições de transiente nos casos de condição de contorno periódica e aberta para valores mais altos de memória, em que o caminhante explora todos ou a maior parte dos sítios e os efeitos de borda se tornam mais evidentes. Vale notar que os resultados mostrados até aqui foram obtidos com condições de contorno periódicas, contudo, para as memórias empregadas, não notamos relevância do tipo de condição de contorno para as discussões levantadas.

4.1.1.1 Condição de contorno periódica

A Figura 10 mostra as distribuições de tempos de transiente para $N = 64, 256, 512$ e 1024 em condições de contorno periódicas e valores de μ acima de 5.

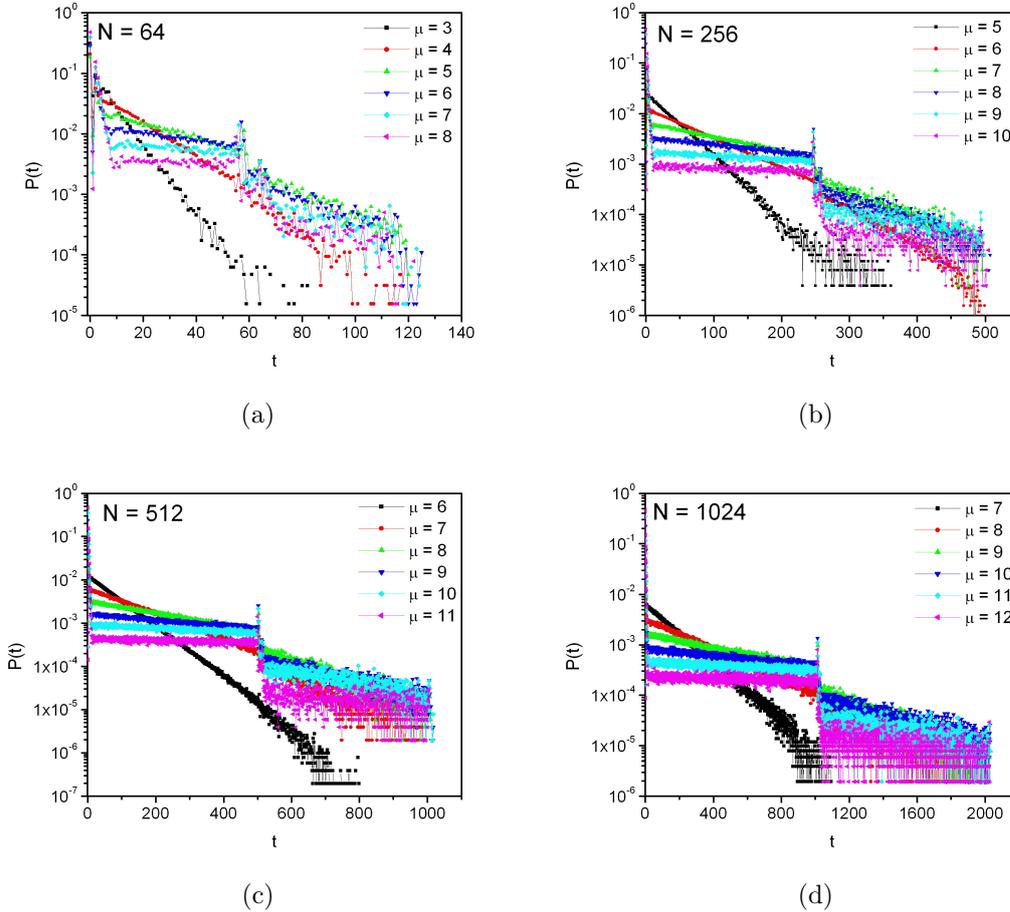


Figura 10: Distribuições de tempos de transiente em função de μ para $N = 64, 256, 512$ e 1024 em condições de contorno periódica. Podemos notar uma mudança de comportamento das curvas de distribuição num valor característico de μ que varia com o tamanho do sistema.

Podemos observar uma mudança súbita de comportamento nas curvas de distribuição de probabilidades a partir de um certo valor de μ característico. Observe que o valor de μ onde ocorre esta mudança varia com N . De fato, Terçariol e colaboradores [27] encontraram um valor de μ crítico a partir do qual o sistema percola, ou seja, o caminhante explora todos os sítios do mapa. Para condições de contorno abertas, o valor encontrado depende de N na forma $\mu = \log_2 N$. Para condições de contorno periódicas, este valor ainda não foi calculado analiticamente. Aqui, observamos esta transição de forma numérica nas curvas de distribuição de tempos de transiente. Na transição, a distribuição de tempos de transiente muda da forma exponencial para uma reta aproximadamente

paralela ao eixo x , onde todos os tempos de transiente se tornam igualmente prováveis até $t = N - (\mu + 1)$. Neste ponto, notamos o surgimento de um pico e, em seguida, uma queda nas probabilidades de ocorrência de t . A Figura 11 mostra esta região em detalhe para $N = 256$ e $N = 512$.

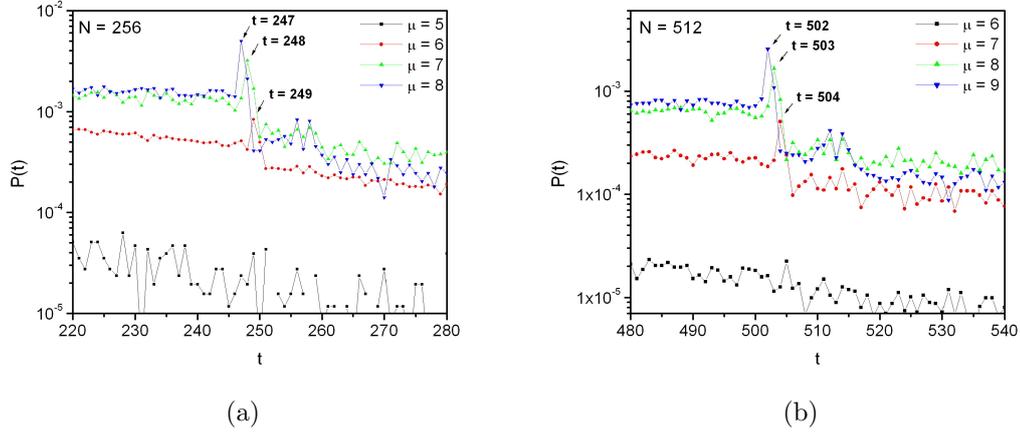


Figura 11: Ampliação das Figuras 10(b) e 10(d) mostrando a região de transição. A transição parece ocorrer em $\mu = 6$ para $N = 2^8$ e $\mu = 7$ para $N = 2^9$. Após a transição, nota-se o aparecimento de um pico em $t = N - (\mu + 1)$. Os valores dos picos estão indicados nas Figuras.

Nesta Figura, vemos claramente a existência de dois comportamentos distintos e uma transição súbita num valor crítico de memória. Com base nestes resultados de simulação, sugerimos que:

$$\mu_c^{(cpc)} = \log_2 \left(\frac{N}{4} \right), \quad (4.3)$$

onde *cpc* designa a condição de contorno.

Observamos que, após a transição, o tempo médio de decaimento (τ) da distribuição diverge (equação 4.2). Assim, definimos uma grandeza α escalada com o valor crítico de μ dado pela equação 4.3 como:

$$\alpha = \frac{\mu}{\mu_c^{cpc}} = \frac{\mu}{\log_2 N - 2} \quad (4.4)$$

e na Figura 12 mostramos como o parâmetro τ se comporta em função da variável $1 - \alpha$.

Observamos que, conforme aumentamos o tamanho do sistema, τ se aproxima da assíntota $1 - \alpha = 0$ o que é característico de sistemas críticos com efeito de tamanho finito [23]. O que a Figura nos mostra é que para $\alpha < 1$, ou seja, para valores de μ menores que a memória crítica, o parâmetro τ é finito. Já para sistemas com valores de

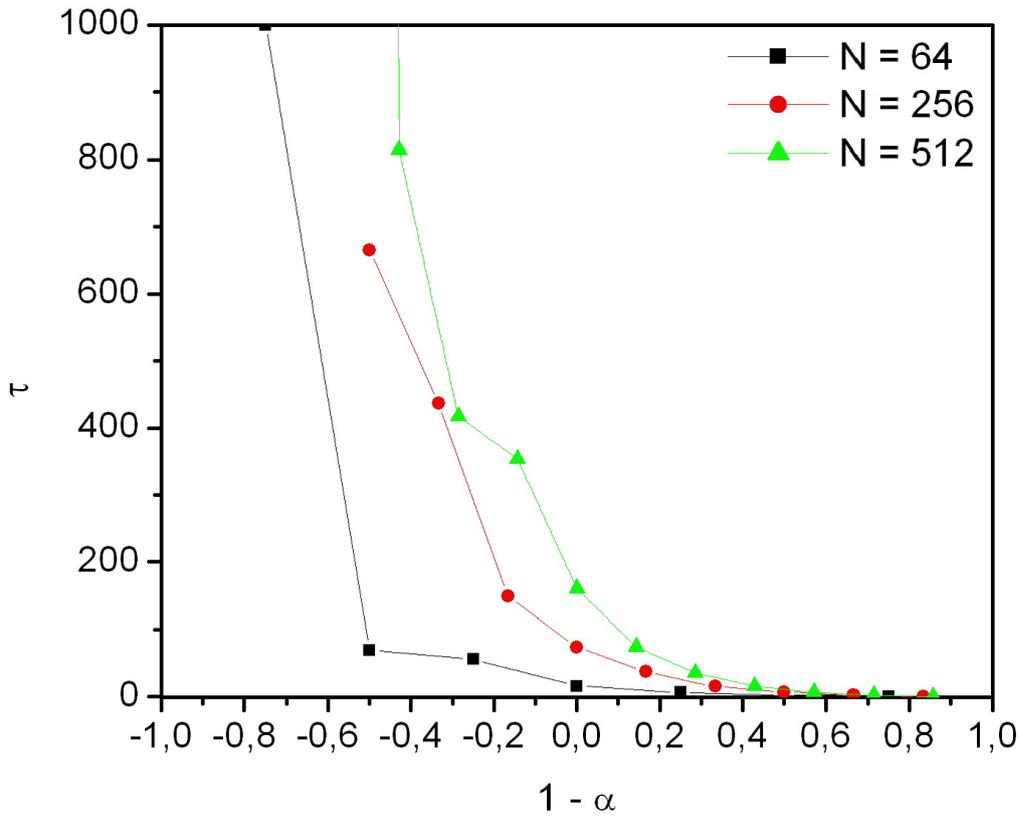


Figura 12: O parâmetro τ (equação 4.2) diverge nos sistemas acima da memória crítica dada pela equação 4.3. Mostramos esta divergência em função do complementar da variável α dada pela equação 4.4. Vemos aqui que, para $\alpha > 1$, ou seja, para valores de memória menores que a memória crítica, o parâmetro τ é finito. Para $\alpha < 1$, valores de memória acima da memória crítica, o parâmetro τ diverge conforme aumentamos N .

memória maiores que a memória crítica, $\alpha > 1$, o parâmetro τ diverge com N . Assim, encontramos um valor de memória crítica μ_c para os sistemas contínuos unidimensionais com condições periódicas de contorno diferente dos resultados obtidos por [27], relativos a sistemas com condições de contorno abertas. Abordaremos este tipo de condição a seguir.

Para tentarmos elucidar o comportamento observado a partir da transição de fase, na Figura 13 mostramos as distribuições de períodos das caminhadas com transientes $t = N - (\mu + 1)$, abaixo e acima deste valor para $N = 256$ e na memória crítica, $\mu = 6$, ou seja:

$$S_{1,6}^{(256)}(t, p), \quad (4.5)$$

com t fixo em 49, 249 e 349.

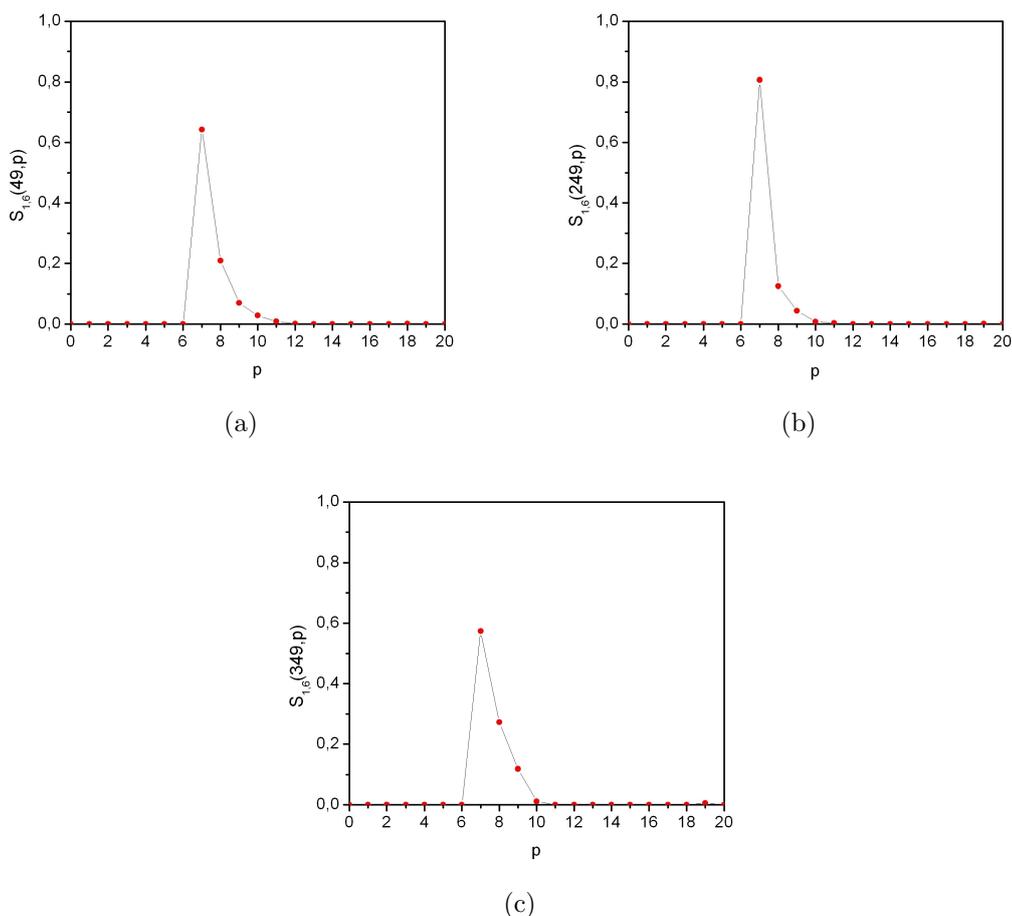


Figura 13: A Figura mostra a distribuição de ciclos das caminhadas com $t = 49$ (Figura 13(a)), $t = 249$ (Figura 13(b)) e $t = 349$ (Figura 13(c)) para $N = 256$ e $\mu = 6$. As curvas foram obtidas a partir da distribuição conjunta, fixando-se t .

É importante observarmos que na distribuição de períodos da caminhada com $t = 49$ há uma porcentagem não desprezível de períodos iguais a N que não é visível na escala da Figura 13(a). Entendemos que, na transição, há a predominância de alguns atratores com períodos específicos, permitidos pela dinâmica. A maior parte dos atratores formados são os de períodos $\mu + 1 \leq p \leq 2\mu$. Veremos a seguir, na seção de distribuição de atratores que estes são os atratores mais frequentes nas caminhadas em meios contínuos. Estas são as caminhadas responsáveis pelo surgimento dos picos em $t = N - (\mu + 1)$. Porém, na transição dá-se também a formação de um atrator com período do tamanho do sistema, ou seja, $p = N$. Isto explica que todos os tempos de transiente sejam igualmente prováveis, pois, partindo-se de qualquer ponto, chega-se ao mesmo atrator.

4.1.1.2 Condição de contorno aberta

Conforme enfatizamos acima, as condições de contorno não exercem influência nas estatísticas estudadas desde que trabalhem com valores de memória abaixo da memória crítica. Na Figura 14 mostramos as distribuições de tempos de transiente nos dois casos para $\mu = 1, 2, 3$ e 5 em um sistema com $N = 512$. Não há distinção entre as curvas.

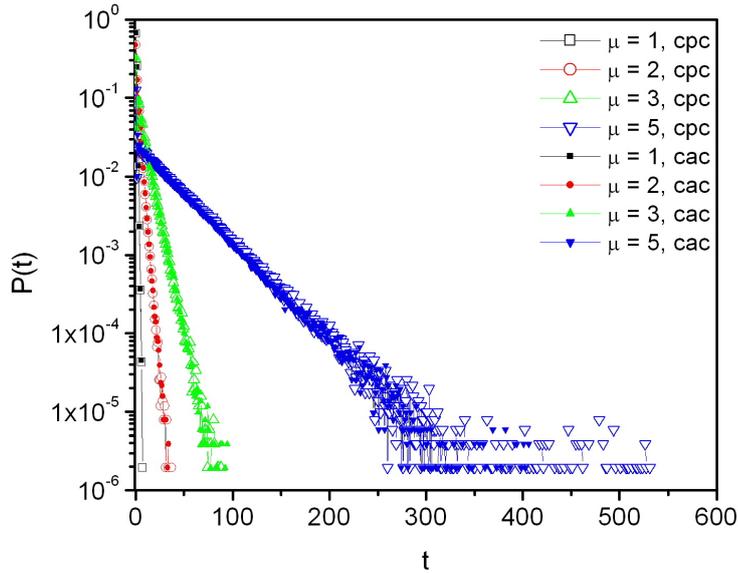


Figura 14: Distribuição de tempos de transiente para $N = 512$ e $\mu = 1, 2, 3$ e 5 nas condições de contorno aberta e fechada.

Na Figura 15 mostramos as distribuições de tempos de transiente para valores de μ acima de 5 nos sistemas com condição de contorno aberta. Novamente, observamos uma transição de fase em um valor de μ característico dependente de N .

De maneira similar aos sistemas com condição de contorno periódica, vemos um aumento no tempo de decaimento das distribuições com o aumento de μ até que o parâmetro τ divirja (horizontalização das curvas de distribuição). Na Figura 16 fica claro que a transição ocorre em valores de μ maiores que os encontrados para as condições de contorno periódicas (equação 4.3). No entanto, nossos resultados não reproduzem os resultados encontrados por [27] para a memória crítica ($\mu_c^{(*)} = \log_2(N)$). Para a condição de contorno aberta, encontramos:

$$\mu_c^{(cac)} = \log_2 \frac{N}{2}, \quad (4.6)$$

em que designamos a condição de contorno pelo índice *cac*.

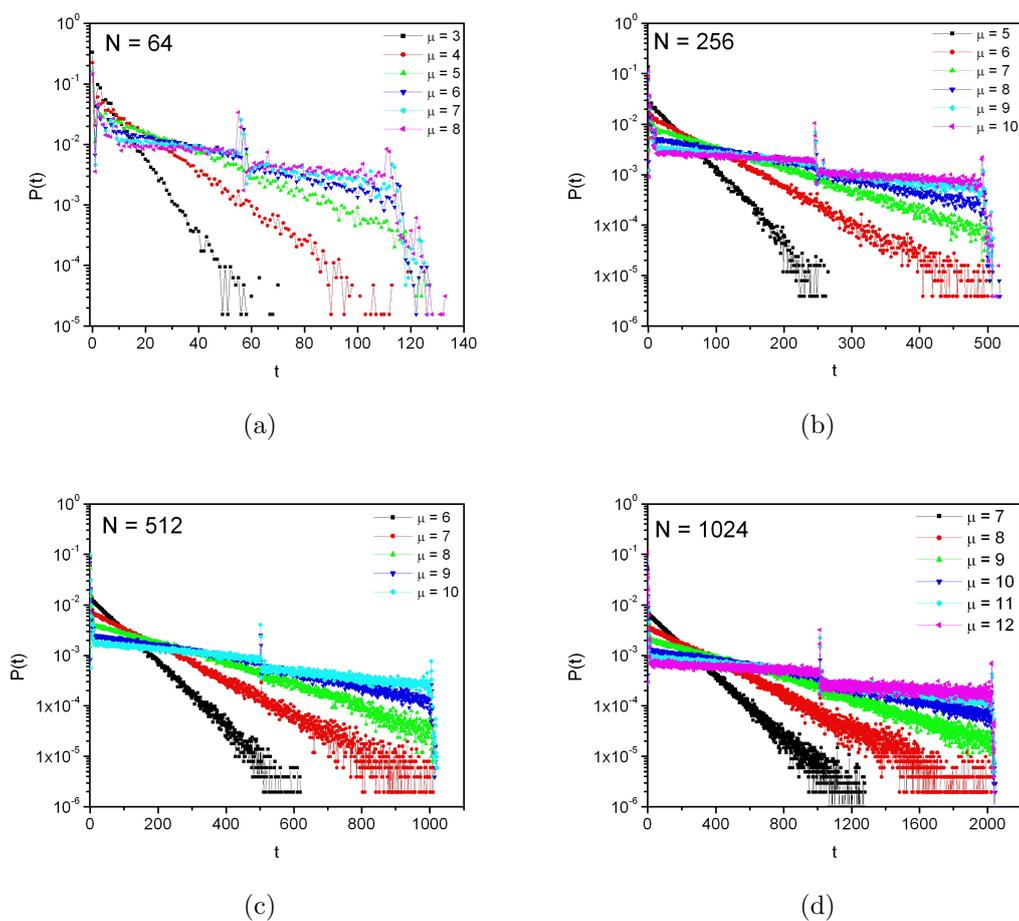


Figura 15: Distribuição de tempos de transiente em função de μ e $N = 64, 256, 512$ e 1024 em condição de contorno aberta.

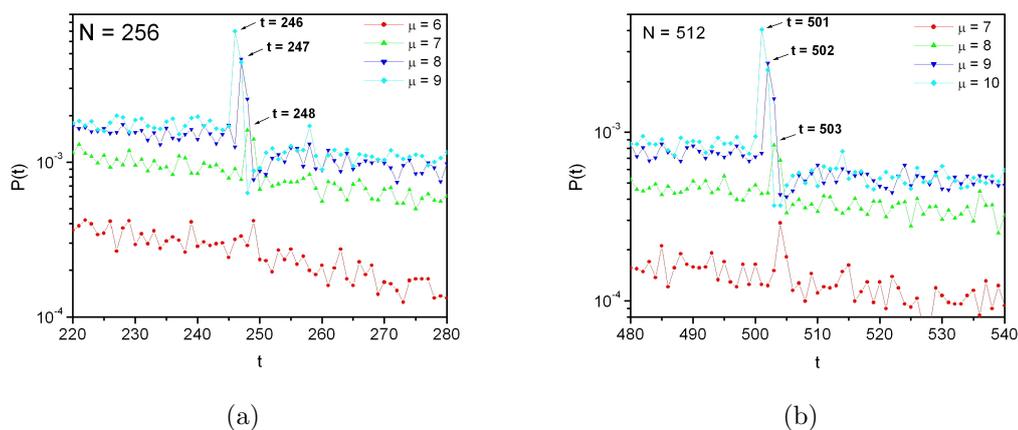


Figura 16: Ampliação das Figuras 15(b) e 15(c) mostrando a região de transição. As curvas representadas em verde indicam a transição. Os valores dos picos em $t = N - (\mu + 1)$ após a transição estão indicados nas Figuras.

É interessante ressaltarmos que os resultados obtidos por [27] foram calculados a partir da probabilidade de um caminhante partir do primeiro sítio do mapa e chegar ao último. No problema abordado neste trabalho, partimos simultaneamente de todos os sítios, assim, um caminhante partindo de um ponto mediano no mapa tem chance de ir para ambas as extremidades. Esta diferença de abordagem pode ter dado origem à diferença encontrada no valor de μ_c que é exatamente a diferença de se considerar o tamanho do sistema dividido por um fator 2, ou seja, a distância média de um ponto até a borda é $N/2$. Neste mesmo raciocínio, entendemos a diferença entre as condições de contorno abertas e periódicas, que também diferem entre si por um fator 2 dividindo o tamanho do sistema. Conjecturamos que:

$$\mu_c^{(cac)} = \mu_c^{(*)} - 1 = \log_2 \left(\frac{N^{(*)}}{2} \right); \quad (4.7)$$

$$\mu_c^{(epc)} = \mu_c^{(cac)} - 1 = \log_2 \left(\frac{N^{(cac)}}{2} \right). \quad (4.8)$$

Certamente, estes valores precisam ser comprovados através de cálculos analíticos nos moldes realizados por [27].

Os picos em $t = N - (\mu + 1)$ (Figura 16) também surgem nestes sistemas e a argumentação apresentada no item anterior também é válida aqui. No entanto, notamos também a presença de um pico em $t = 2N - 2\mu$ para estes sistemas. Nos sistemas com condição de contorno aberta, a formação de trajetórias do tipo “vai-e-vem” é facilitada, já que, uma vez alcançada a borda, o caminhante não tem como prosseguir e é obrigado a voltar. Ao voltar, a restrição da memória faz com que o caminhante salte μ sítios em cada uma das extremidades. Por isso, observamos o aumento da probabilidade de ocorrência de tempos transientes com $2N - 2\mu$ passos.

4.1.2 Distribuição de ciclos

As distribuições de ciclos $P(p)$ para $N = 256$ e $N = 512$ estão apresentadas na Figura 17. A figura mostra a probabilidade de ocorrência de um atrator com período p dentre todos os atratores formados. A dinâmica da caminhada do turista impõe que $p_{min} = \mu + 1$ por isso observamos o deslocamento das curvas para os diferentes valores de μ . A probabilidade de ocorrência de p cai rapidamente para os ciclos no intervalo $\mu < p \leq 2\mu$ seguido de três períodos proibidos. Observamos também que para $\mu = 2$, todos os ciclos com períodos ímpares a partir de $p > 2\mu$ são proibidos. Todas estas características foram discutidas previamente por [26].

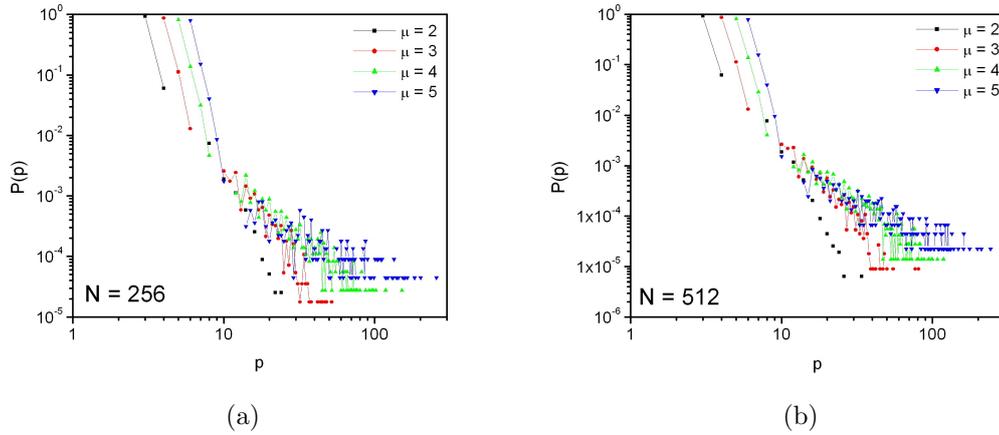


Figura 17: Distribuição de ciclos em função de μ para $N = 256$ e 512 .

A distribuição de ciclos a partir do valor de $p = 2\mu$ torna-se bastante ruidosa, tornando inviável a comparação entre diferentes distribuições. Uma maneira de se contornar este problema é usar a distribuição acumulada complementar. A distribuição acumulada complementar (DAC) pode ser escrita como:

$$P_c(p) = P(X > p) = 1 - \sum_{p_i \leq p} P(p_i) \quad (4.9)$$

onde X é a variável aleatória associada ao período p . $P_c(p)$ define a probabilidade de se encontrar um evento maior que p . Para uma distribuição de probabilidades que segue uma lei de potência, a distribuição acumulada também é uma lei de potência com o expoente subtraído de uma unidade [28]. Assim, mostramos na Figura 18 as distribuições acumuladas complementares para os sistemas mostrados na Figura 17. Podemos perceber claramente que, ao contrário do sugerido em [5, 6], as distribuições de ciclos para sistemas contínuos em uma dimensão não seguem uma lei de potência.

Na Figura 19, comparamos as DAC's de ciclos para $N = 256$ e $N = 512$ e $\mu = 2, 3$ e 4 e mostramos que não há dependência com o tamanho do sistema para valores baixos de memória. Aumentando-se o valor de μ , observamos novamente uma mudança qualitativa nas DAC's com características diferentes para as diferentes condições de contorno. A seguir, mostraremos estas diferenças, abordando as estatísticas de ciclos em ambas as condições para memórias acima de 5.

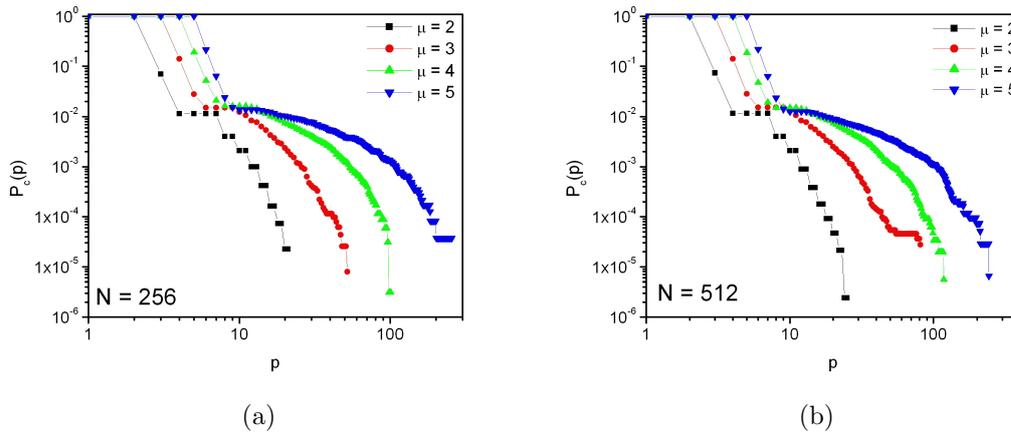


Figura 18: Distribuição acumulada complementar de ciclos para os sistemas mostrados na Figura 17.

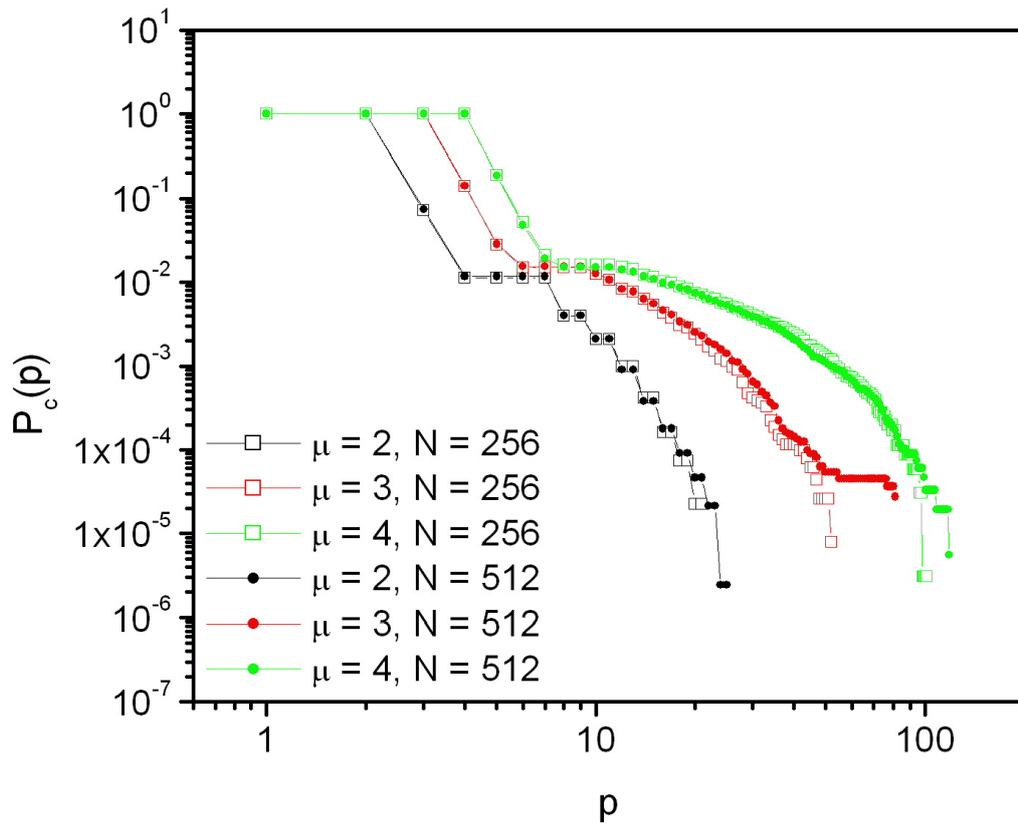


Figura 19: Distribuições acumuladas complementares para $N = 256$ e $N = 512$ e μ variando de 2 a 4.

4.1.2.1 Condição periódica de contorno

A Figura 20 mostra as distribuições de ciclos para $\mu > 5$. As distribuições acumuladas complementares permitem uma distinção nítida das duas fases existentes dependendo do

valor de μ . Podemos confirmar o valor de μ_c dado pela equação 4.3.

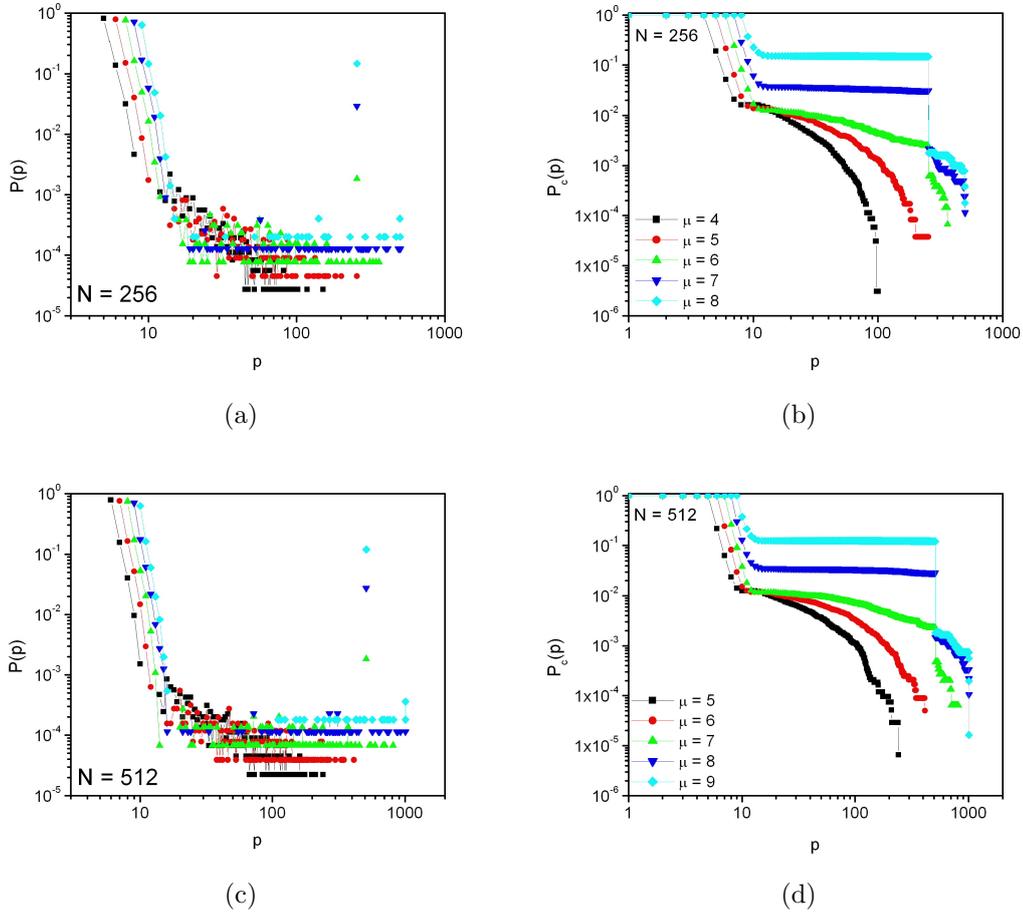


Figura 20: Figuras 20(a) e 20(c): distribuição de ciclos para $N = 256$ e $N = 512$ respectivamente em condições de contorno periódicas. Figuras 20(b) e 20(d) mostram as DAC's para os mesmos sistemas anteriores. Note que as DAC's permitem a distinção entre as fases antes e depois da transição.

A partir da transição, observamos o surgimento de um pico em $p = N$ nas Figuras 20(b) e 20(d) que corresponde à queda abrupta ocorrida no mesmo ponto nas Figuras 20(a) e 20(b). Este comportamento corrobora nossa argumentação sobre o possível cenário das caminhadas na transição em condições de contorno periódicas.

4.1.2.2 Condição aberta de contorno

Na Figura 21 mostramos as distribuições de ciclos para $N = 256$ e $N = 512$ e suas respectivas distribuições acumuladas complementares para $\mu > 5$. Para as condições abertas de contorno, a transição não é tão clara quanto em condições periódicas. No entanto, podemos identificar o valor de memória em que ocorre $p = 2N$ coincidindo

com o valor dado pela equação 4.6. Conforme discutimos na seção 4.1.1.2, nas condições abertas de contorno, trajetórias do tipo “vai-e-vem” são mais fáceis de ocorrer, por isto, aqui a transição se dá quando ocorrem ciclos de período igual a $2N$.

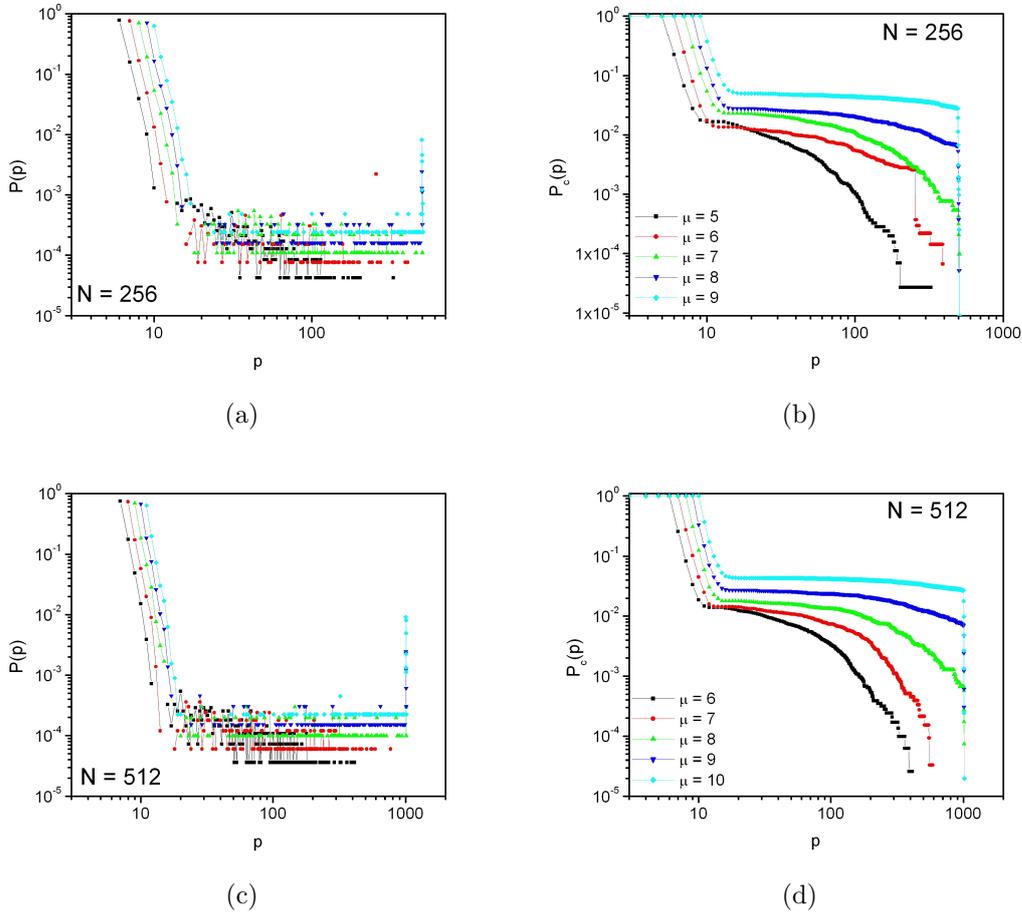


Figura 21: As Figuras 21(a) e 21(c) trazem respectivamente as distribuições de ciclos para $N = 256$ e $N = 512$ em condições de contorno abertas. As Figuras 20(b) e 20(d) mostram as DAC’s para os mesmos sistemas.

Todo este cenário descrito revela que nas condições periódicas de contorno, a caminhada do turista pode assumir dois comportamentos qualitativamente distintos. O primeiro, com memória baixa, em que o movimento do caminhante é localizado, ou seja, vemos o aparecimento de diversos pequenos ciclos espalhados pela paisagem. Chamaremos esta de fase *I*. O segundo tipo de comportamento, ou fase *II*, é caracterizado pela presença de um ciclo que circula toda a rede, ou seja, $p = N$. No limite $N \rightarrow \infty$, podemos definir um valor μ_c tal que a probabilidade de aparecer um ciclo infinito é não nula. As Figuras 22(a) e 22(b) ilustram estes dois comportamentos. Esta transição pode ser comparada à transição oscilação-rotação do pêndulo simples. Com pouca energia, o pêndulo oscila em torno de um ponto médio. Conforme se aumenta a energia, o ângulo

de oscilação aumenta até que, a partir de uma energia crítica, ele passa a rotacionar em torno de seu eixo indefinidamente. Neste caso, consideramos a memória μ análoga à energia aplicada ao sistema (Figura 22(c)). Quanto maior o valor de μ , maior a probabilidade de ocorrência de atratores com $p > N$. Nas Figuras 20(b) e 20(d), notamos uma cauda em $p > N$ que representam as condições iniciais (ou realizações) que encontram-se na primeira fase, mesmo com memória $\mu > \mu_c$. Com o aumento de μ , cada vez menos realizações encontram-se em fase de oscilação e, por isso, as caudas diminuem.

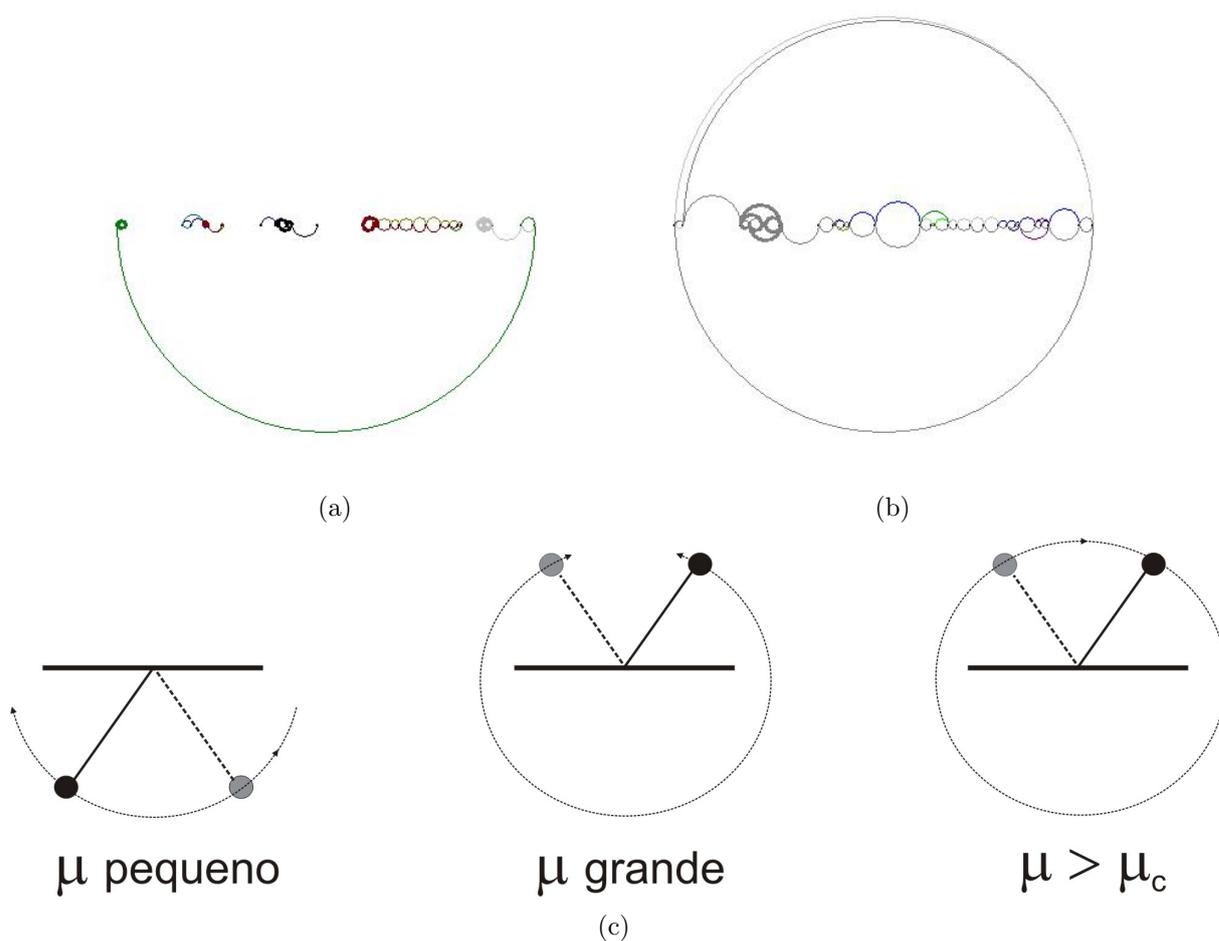


Figura 22: Figuras 22(a) e 22(b): caminhadas em meio unidimensional ilustrando a fase localizada (Figura 22(a)) e delocalizada (Figura 22(b)). Figura 22(c): o pêndulo simples mostrando a analogia entre as fases localizada e delocalizada da caminhada em meio unidimensional.

O mesmo não se pode observar nas Figuras 21(b) e 21(d), já que não existem atratores com rotação. Nesta condição, a nova fase é caracterizada pelo movimento de “vai-e-vem” (período $p = 2N$).

4.2 Caminhadas no espaço bidimensional

Caminhadas deterministas em sistemas com $d > 1$ são de muito mais difícil análise que em sistemas unidimensionais. No entanto, como veremos a seguir, oferecem aplicações mais realistas e mais interessantes. Nesta seção consideramos as características da caminhada do turista em meios aleatórios com $d = 2$.

Na Figura 23 ilustramos um exemplo de caminhadas em um meio contínuo aleatório de duas dimensões com memória $\mu = 5$. A figura mostra quatro diferentes trajetórias com a parte transiente em vermelho e o atrator em preto. Note que as trajetórias podem conter polígonos e laços sem, contudo configurar a formação de um atrator.

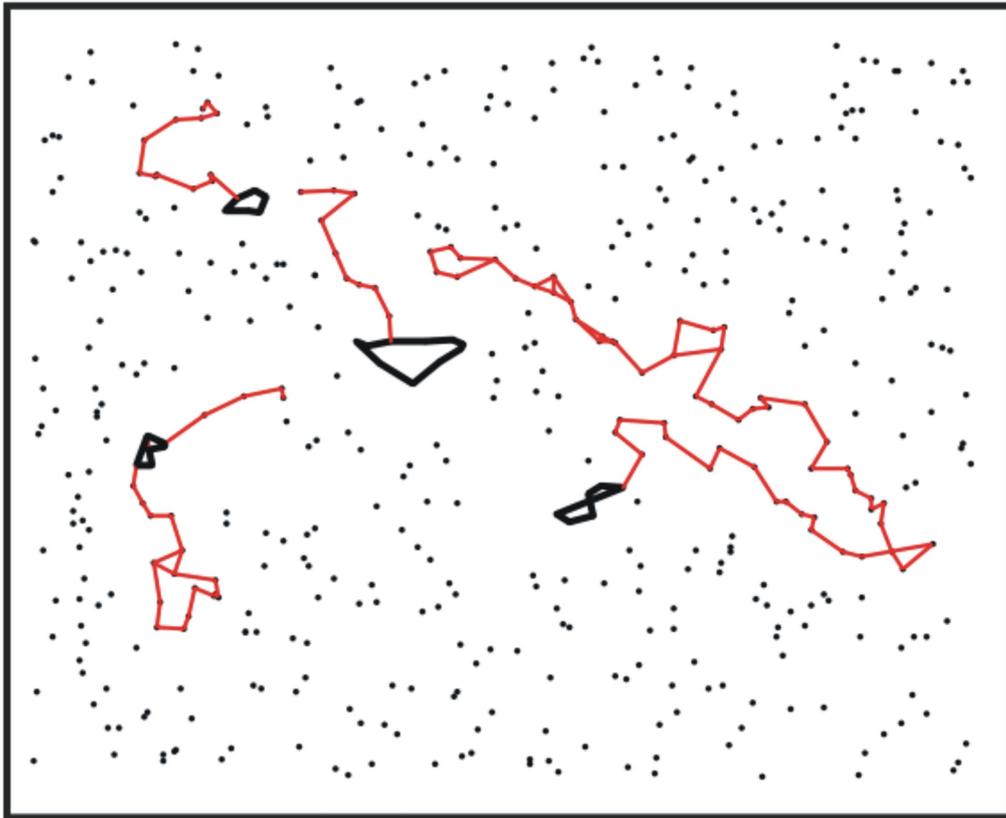


Figura 23: Exemplos de caminhadas do turista em um meio bidimensional de $N = 500$ e memória $\mu = 5$. As trajetórias de caminhadas partindo de quatro pontos diferentes estão ilustradas com as partes transitientes em vermelho e os atratores em preto.

Em uma dada paisagem, conforme se aumenta o valor de μ tem-se um aumento do período mínimo $p_{min} = \mu + 1$ e novos atratores podem se formar da fusão de atratores anteriores (Figura 24). Note que alguns atratores somem para um dado valor de μ e

reaparecem em seguida.

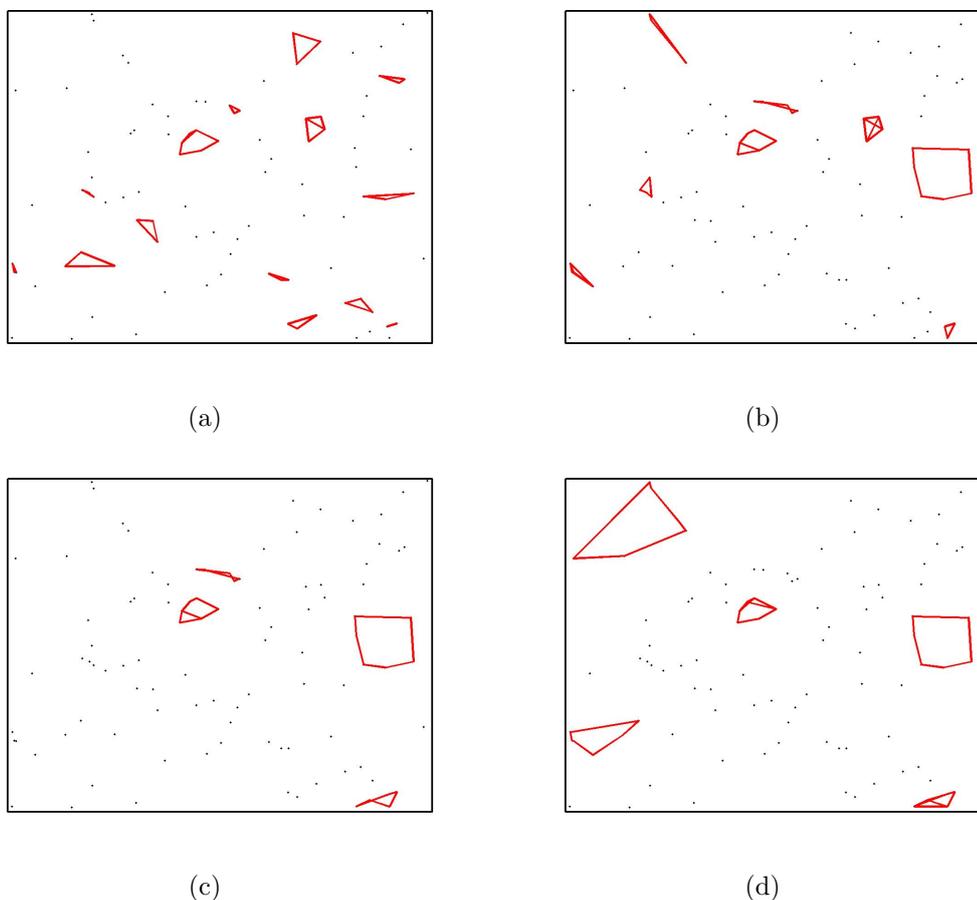


Figura 24: Caminhada numa paisagem fixa, variando-se o valor de μ . A Figura mostra apenas os atratores para cada valor de μ . Figura 24(a): $\mu = 2$; Figura 24(b): $\mu = 3$; Figura 24(c): $\mu = 4$; Figura 24(d): $\mu = 5$.

A seguir apresentamos as estatísticas de tempos de transiente e de ciclos atratores para os sistemas bidimensionais.

4.2.1 Distribuição de tempos de transiente

Na Figura 25, vemos a distribuição de tempos de transiente para um sistema de 512 pontos em duas dimensões. Para memórias baixas, a distribuição de transientes obedece a uma função exponencial e o tempo característico τ de cada curva aumenta com o aumento da memória, de forma semelhante aos sistemas em uma dimensão. Conforme se aumenta a memória, as distribuições começam a apresentar uma curvatura e sua cauda é melhor ajustada por uma exponencial estendida (equação 4.10):

$$P(t) \propto \exp\left(\frac{-t}{\tau}\right)^q, \quad (4.10)$$

com $q < 1$.

Apesar das curvas de tempos de transiente apresentarem uma mudança de comportamento a partir de $\mu = 7$, não é possível evidenciar uma transição de fase até a memória estudada ($\mu = 25$).

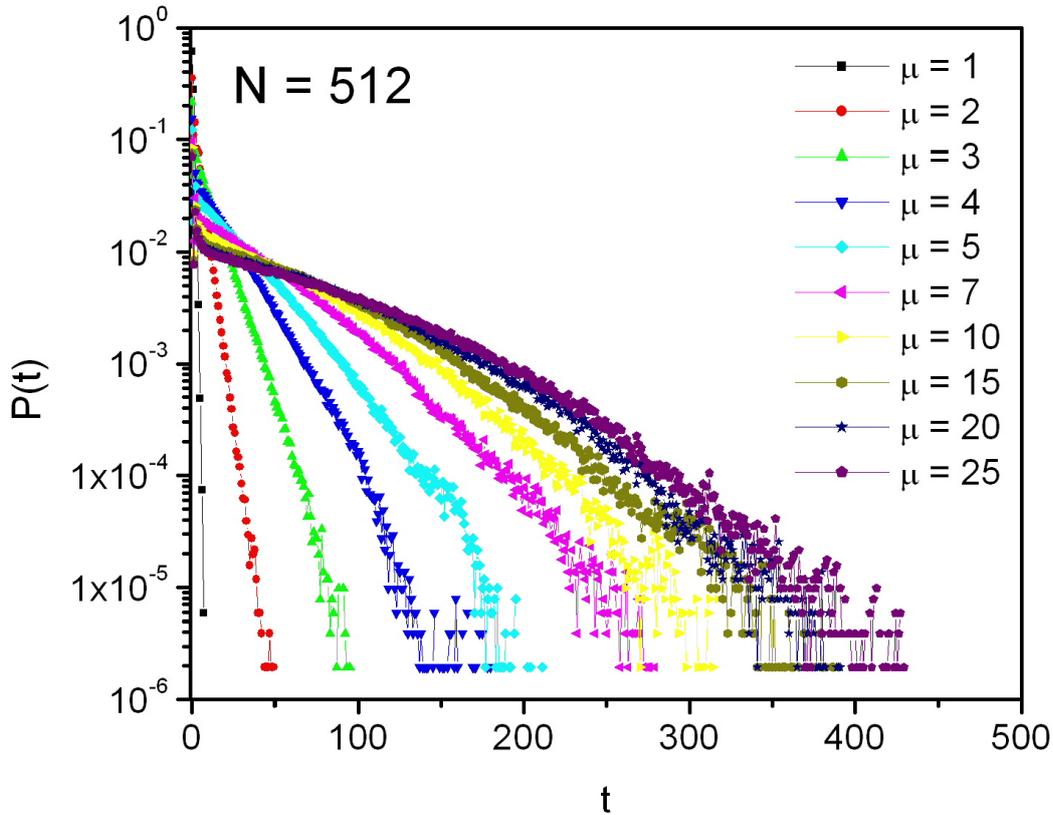


Figura 25: Distribuição de tempos de transiente em função de μ para um sistema de $N = 512$.

Repetimos este experimento com diferentes tamanhos de sistema e, para memórias pequenas, a distribuição de transientes não varia com N (Figura 26). Porém, para memórias mais altas, com maiores probabilidades de ocorrência de transientes grandes, vemos que as curvas de $P(t)$ diferem fortemente. A Figura 27 ilustra como o tempo médio de transiente

$$\langle t \rangle = \sum_{t=0}^{\infty} tP(t). \quad (4.11)$$

varia com o tamanho do sistema.

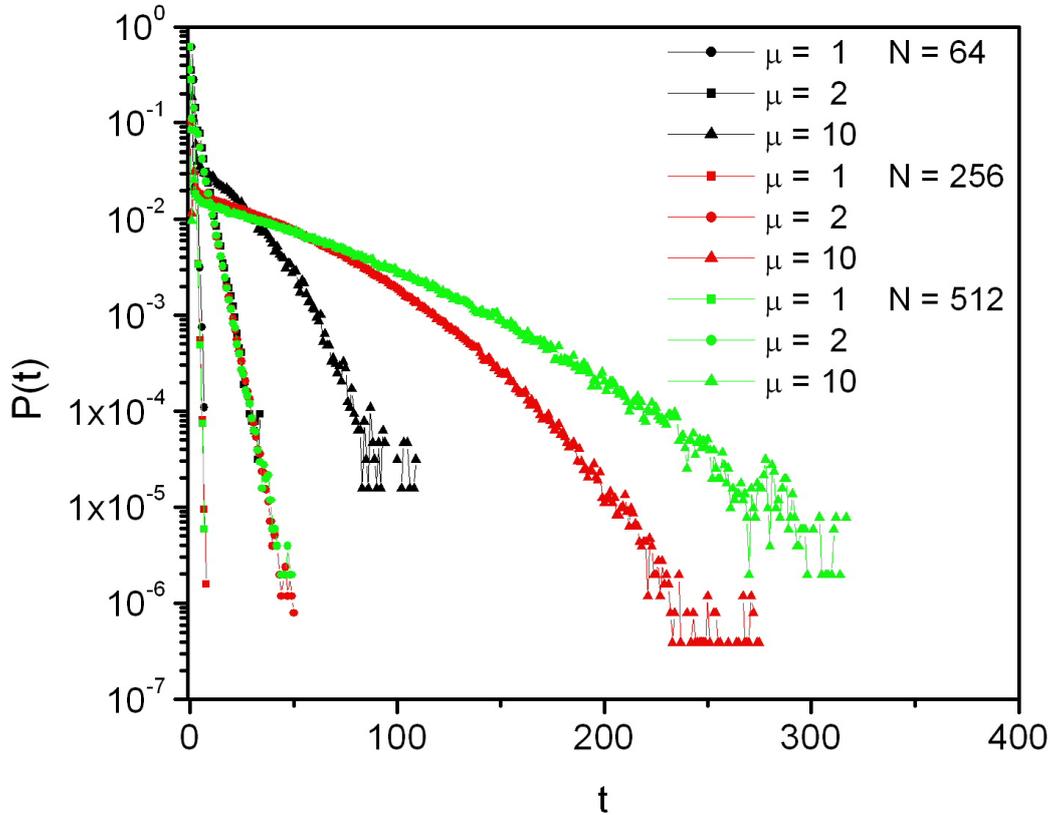


Figura 26: Distribuição de tempos de transiente em função de N para valores de $\mu = 1, 2$ e 10 .

Corroborando o resultado anterior, vemos que para $\mu = 2$, o tempo médio de transiente não varia com N , já que a exploração do meio é pequena. Já para $\mu = 10$, vemos um aumento inicial de $\langle t \rangle$ em função de N e uma posterior tendência à saturação.

Na Figura 28 investigamos como $\langle t \rangle$ varia em função de μ . O tempo médio de transiente cresce rapidamente até aproximadamente o valor de $\mu = 8$. A partir daí $\langle t \rangle$ passa a crescer mais devagar com μ , indicando uma saturação devido a N finito. Todos estes resultados sugerem a existência de uma transição de fase dependente de μ como encontrado em sistemas unidimensionais, porém em valores de memória provavelmente da ordem de N . De fato, estudos em andamento [29] indicam uma transição de fase próxima a $\mu = N/2$. No entanto, o estudo desta transição foge ao escopo deste trabalho e não será investigado em detalhe.

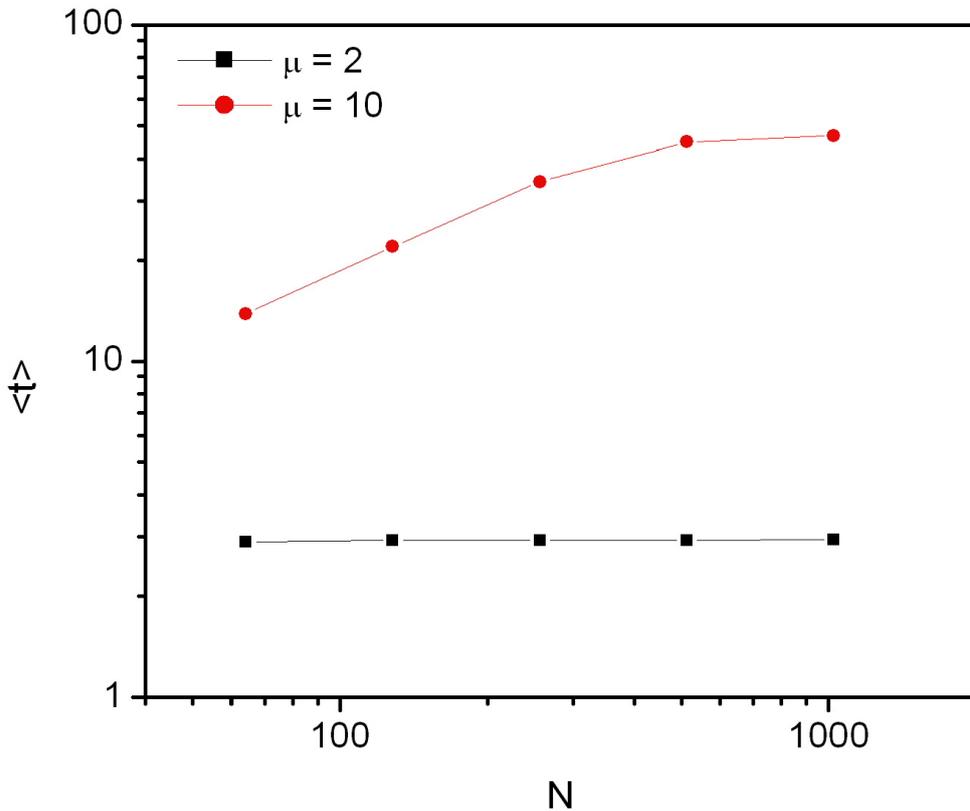


Figura 27: Tempo de transiente médio em função do tamanho N do sistema para $\mu = 2$ e $\mu = 10$.

4.2.2 Distribuição de ciclos

A distribuição de ciclos para um sistema com $N = 512$ está mostrada na Figura 29(a). Observamos o deslocamento das curvas para os diferentes valores de memória devido à restrição $p \geq \mu + 1$. Para $\mu = 1$, todos os ciclos têm período $p = 2$ e a densidade de ciclos pode ser obtida analiticamente para qualquer dimensão d [3]. Para $\mu = 2$, a probabilidade de ocorrência de ciclos com períodos p é uma função oscilatória a partir de $p = 7$. Os ciclos de períodos pares são mais prováveis que os ciclos de períodos ímpares vizinhos. Este comportamento foi observado em redes booleanas com valores de conectividade baixos [30, 31].

As curvas não indicam claramente a presença de leis de potência. Isto fica mais claro nas distribuições acumuladas complementares mostradas na Figura 29(b).

Verificamos como o período médio dos atratores varia com as grandezas N e μ (Fi-

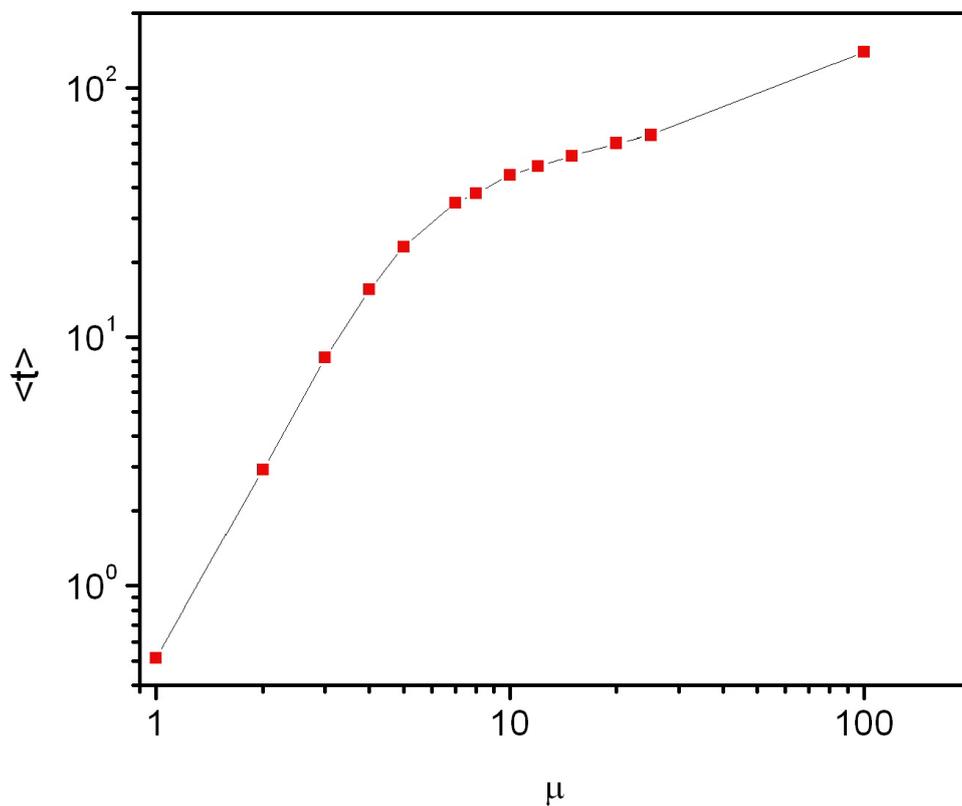


Figura 28: Tempo de transiente médio em função de μ para $N = 512$.

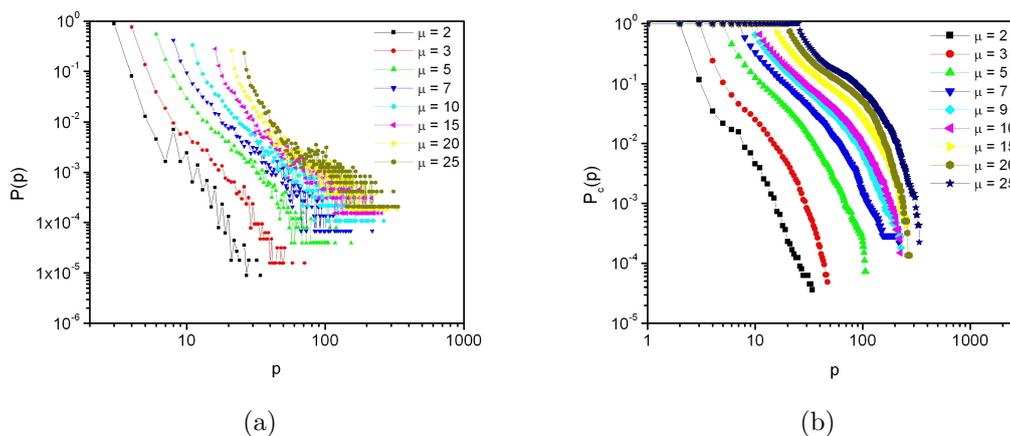


Figura 29: Distribuição de ciclos para um sistema bidimensional de $N = 512$ em função de μ . A Figura 29(b) mostra as respectivas distribuições acumuladas complementares.

gura 30). Definimos o período médio $\langle p \rangle$ pela equação:

$$\langle p \rangle = \sum_{p=0}^{\infty} pP(p). \quad (4.12)$$

Na Figura 30(a), vemos que $\langle p \rangle$ praticamente não varia para ambos os valores de μ na faixa de valores de N mostrada. Vale observar que a densidade de ciclos não varia com N nestes valores de memória (dados não mostrados). Por outro lado, verificamos um crescimento quase linear de $\langle p \rangle$ com μ . Isto também é compatível com um cenário com uma transição de fase quando $\mu_c = \alpha_c N$.

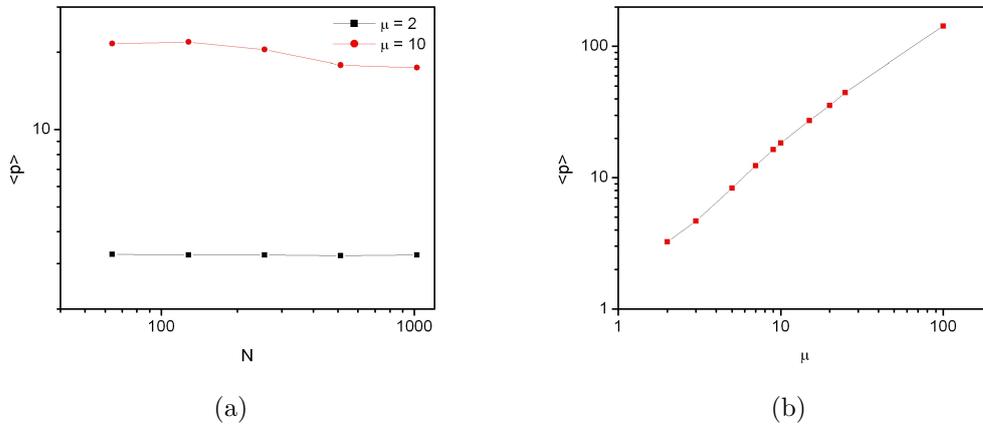


Figura 30: Figura 30(a): Período médio de atratores em função de N para $\mu = 2$ e $\mu = 10$. Figura 30(b): Período médio de atratores em função de μ para $N = 512$.

Uma questão interessante a se abordar é se o tempo de transiente tem correlação com o tamanho do atrator atingido. A Figura 31 mostra que não há correlação entre o tempo de transiente e o período do atrator. No entanto, é interessante notar que o gráfico é limitado superiormente pela reta $p + t = C$ e que C é uma constante que depende do valor de μ .

A soma $p + t$ representa o total da trajetória percorrida pelo caminhante e nos dá idéia da proporção de sítios explorados pela caminhada. Entender como esta grandeza depende do valor de μ nos permite prever quando o sistema irá percolar e, portanto prever se podemos esperar uma transição de fase em um valor finito de μ para o dado sistema. Definimos portanto a grandeza $\langle p + t \rangle$ calculada sobre todas as caminhadas realizadas para um sistema de $N = 512$. Esta grandeza está mostrada em função de μ na Figura 32.

A variação de $\langle p + t \rangle$ com μ não é linear e o tamanho das trajetórias cresce vagarosamente com μ . Notamos que até o valor de memória investigado ($\mu = 100$) não houve percolação. No entanto, podemos inferir que o valor de memória em que $\langle p + t \rangle$ se apro-

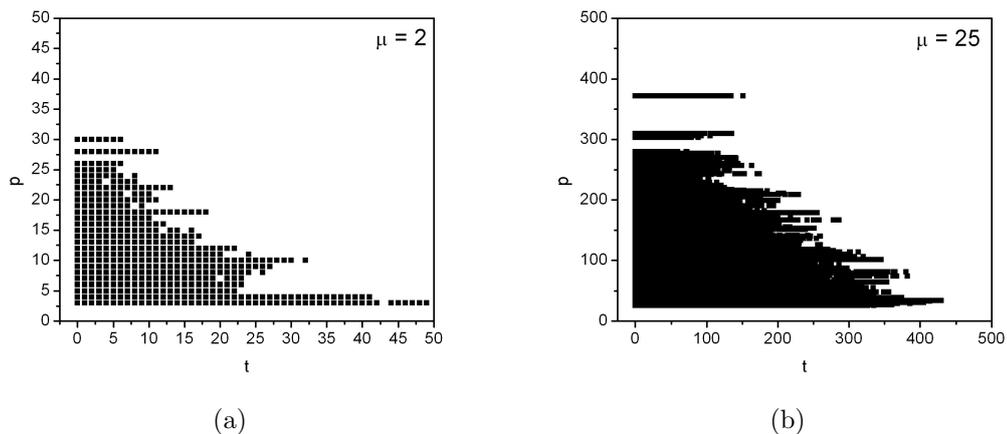


Figura 31: A Figura mostra os tempos de transiente contra os períodos de atrator obtidos para cada trajetória percorrida num sistema de $N = 512$ e $\mu = 2$ (Figura pxt2) e $\mu = 25$ (Figura 31(b)). Note que as figuras estão mostradas em escalas diferentes.

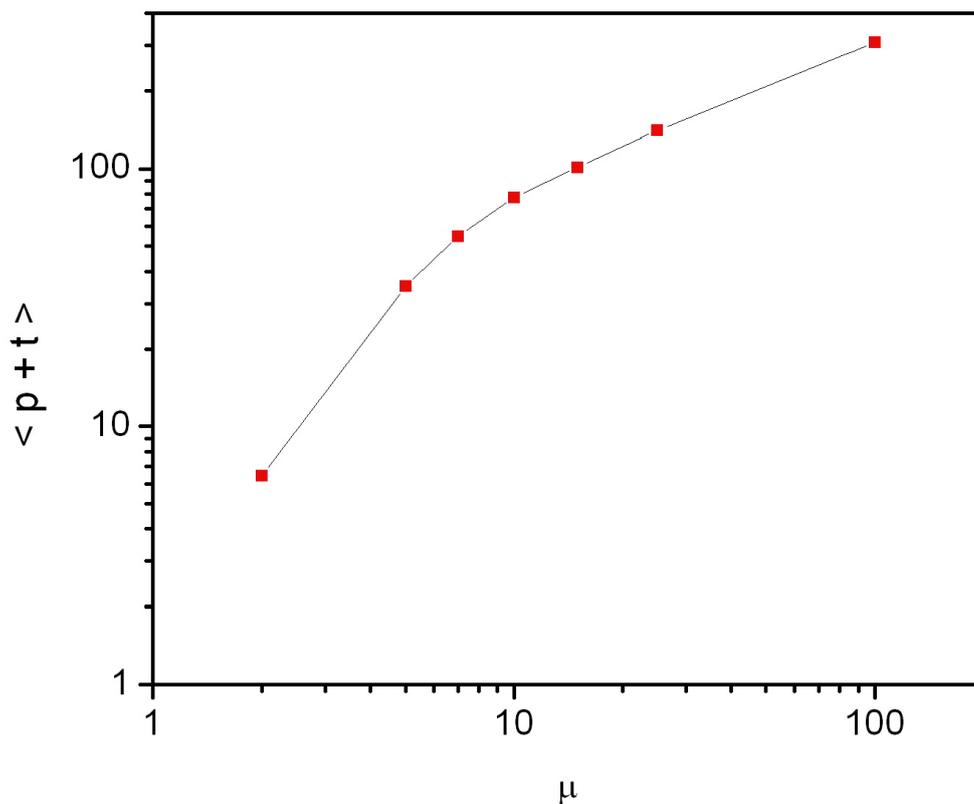


Figura 32: Média do total de pontos percorridos por trajetória num sistema de $N = 512$ pontos em função de μ .

xima de N é da ordem de N , compatível com $\mu_c \approx N/2$ conforme citamos acima [29]. Para aplicações em exploração de dados e busca de agrupamentos naturais que consideramos nesta tese, um valor de μ_c da ordem de N representa um obstáculo para o tratamento de conjuntos de dados muito grandes, já que é necessário atingir uma memória grande para que se explore todo o conjunto. Contudo, conjuntos de dados reais, freqüentemente apresentam um padrão de organização entre os pontos e isto modifica a dinâmica da caminhada. Conjuntos com padrões de organização permitem uma difusão mais rápida do caminhante e esta propriedade será explorada no Capítulo 6 que apresenta um algoritmo de agrupamento automático de dados multidimensionais baseado na caminhada do turista.

4.2.3 Condição aberta de contorno

As distribuições de tempos de transiente com condições de contorno abertas se comportam de maneira similar às com condições de contorno periódicas. As curvas apresentam inicialmente (para memórias pequenas) um comportamento exponencial e, à medida que se aumenta o valor de μ , observa-se um comportamento do tipo de exponencial estendida.

A Figura 34 mostra as distribuições de tempos de transiente para $N = 512$ em ambas as condições de contorno e valores de $\mu = 1, 2$ e 10 . Para valores baixos de memória, com baixa exploração do meio, não é possível notarmos efeitos de borda. Entretanto, já com $\mu = 10$, notamos uma queda na probabilidade de ocorrência de tempos de transiente longos para a condição de contorno aberta conforme esperado.

Nas distribuições acumuladas complementares de ciclos (Figura 34), vemos o mesmo efeito. Para $\mu = 10$, podemos notar diferenças entre as duas curvas, com uma diminuição na probabilidade de ciclos longos para a condição de contorno aberta. Podemos esperar que, em condições de contorno abertas, o caminhante demore mais para explorar todo o meio em comparação com os sistemas em condições de contorno fechadas, assim como verificamos em sistemas unidimensionais.

4.3 Conclusões

Apresentamos as estatísticas de tempos de transiente e de ciclos para sistemas em meios contínuos unidimensionais e bidimensionais em condições de contorno periódicas e abertas. Uma das questões mais pertinentes no estudo de sistemas dinâmicos é acerca da ocorrência de transição de fase. A caminhada determinista do turista representa um

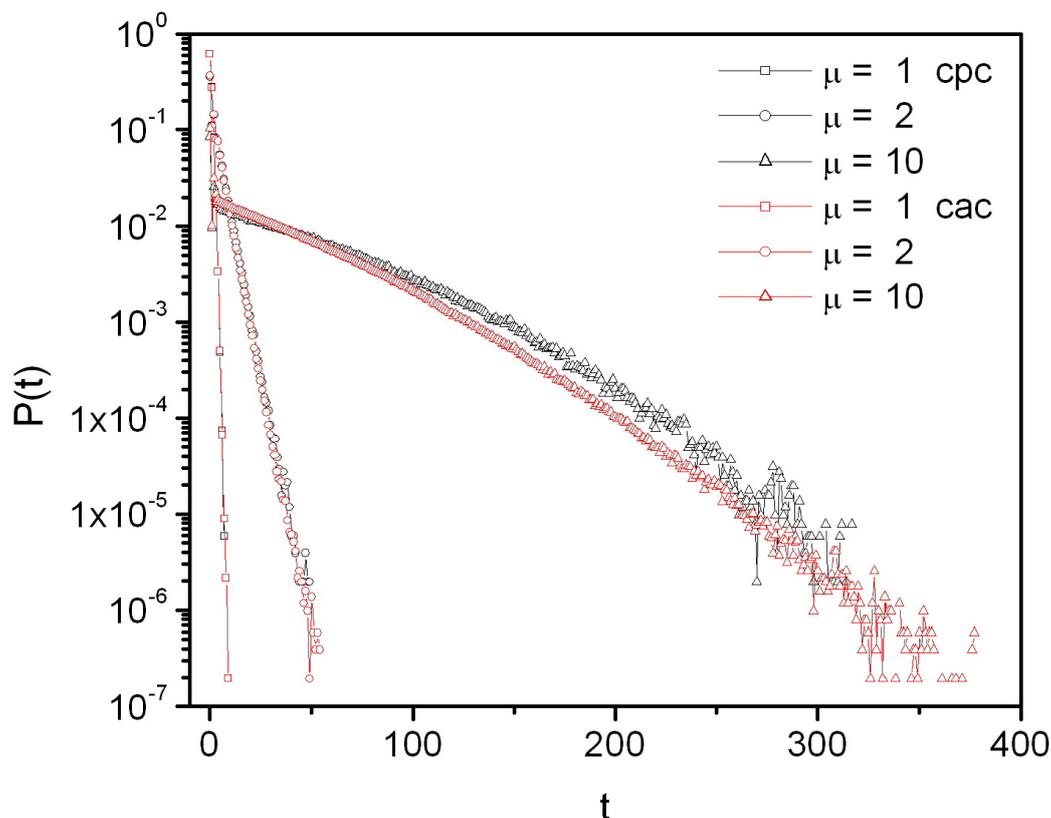


Figura 33: Comparação entre as condições de contorno aberta (cac) e periódica (cpc) para $N = 512$ e $\mu = 1$, $\mu = 2$ e $\mu = 10$.

problema interessante da classe dos sistemas dinâmicos e suas propriedades como sistema crítico vêm sendo investigadas [27, 32].

Neste capítulo, apresentamos uma importante contribuição para o entendimento da caminhada do turista como sistema crítico. Conforme mencionamos acima, [27] realizaram o estudo da transição de fase em uma dimensão calculando a probabilidade de percolação do sistema, ou seja, a probabilidade do caminhante chegar ao último sítio do mapa, partindo do primeiro sítio e encontraram o valor crítico $\mu_c = \log_2 N$. Aqui, utilizamos as estatísticas de tempos de transiente e de ciclos para mostrar a transição e entender as mudanças na dinâmica da caminhada que levam às diferentes propriedades apresentadas em cada comportamento (antes e depois da transição). No caso de condições periódicas de contorno, temos uma transição oscilação-rotação com $\mu_c = \log_2 \frac{N}{4}$. No caso da caminhada com condições abertas, temos uma transição oscilação localizada ($p < 2N$) - oscilação “vai-e-vem” ($p = 2N$). É claro que estes comportamentos correspondem todos a uma

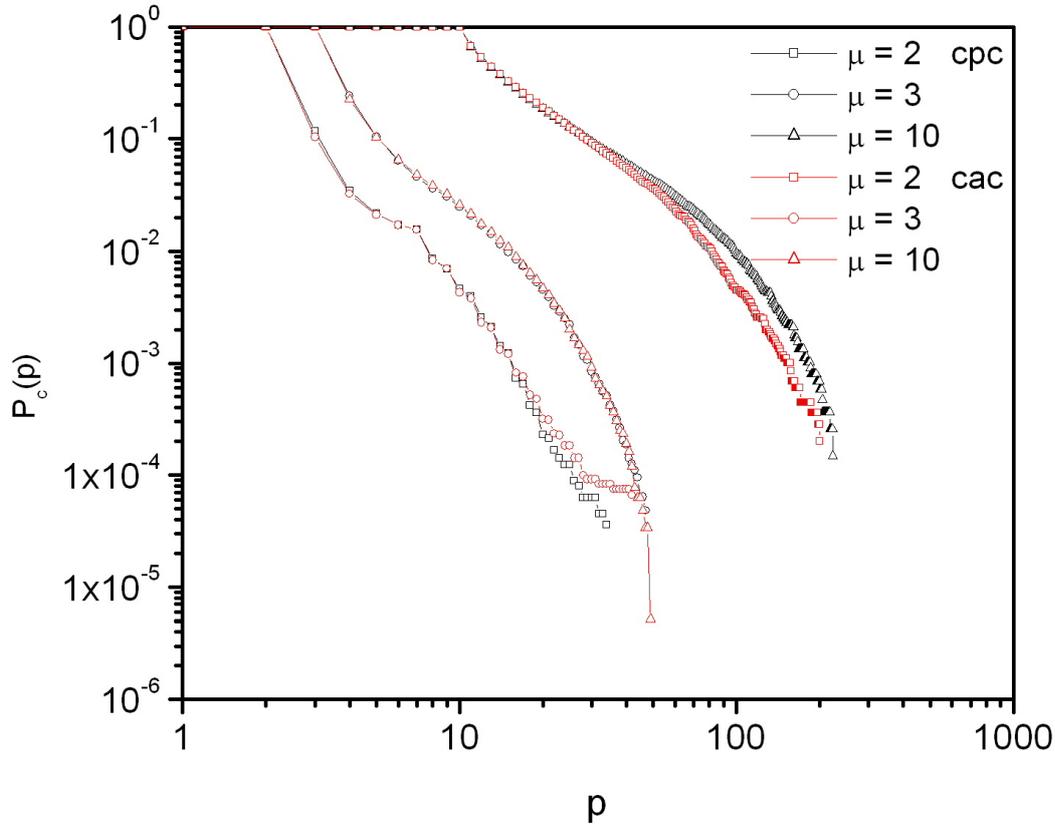


Figura 34: Comparação das distribuições acumuladas complementares de ciclos em condições de contorno aberta e periódica para $N = 512$ e $\mu = 2, 3$ e 10 .

transição do tipo localização-delocalização no limite $N \rightarrow \infty$, em que o caminhante segue sempre em frente (não retorna) quando:

$$\alpha = \lim_{N \rightarrow \infty} \frac{\mu}{\log_2(CN)} \quad (4.13)$$

(em que C é uma constante dependente das condições de contorno) for maior que $\alpha_c = 1$.

Um segundo ponto importante levantado neste capítulo e feito pela primeira vez no estudo da caminhada do turista é a representação das distribuições de ciclos pela distribuição acumulada complementar. A DAC foi utilizada por [31] em redes booleanas aleatórias para representar este tipo de distribuições. Da mesma maneira que mostramos aqui, o autor utiliza a distribuição acumulada complementar para mostrar de forma qualitativa a transição de fase nestas redes. Também pudemos demonstrar, utilizando as propriedades das distribuições acumuladas, que as distribuições de ciclos dos sistemas abordados aqui não seguem lei de potência, diferentemente do que foi sugerido em [6].

Acreditamos que tais leis de potência apareçam apenas quando $\mu = \mu_c$. Nos trabalhos anteriores [5], a existência de leis de potência quando $d \rightarrow \infty$ e as caudas longas das distribuições foram tomadas como evidência a favor de leis de potência com d finito. Nossos resultados, especialmente o gráfico de $\langle p \rangle$ contra N , sugerem que, se μ não cresce com N , não ocorrerão leis de potência. Este é mais um exemplo onde sistemas quasi-críticos foram tomados como genericamente críticos devido às limitações dos resultados numéricos [33].

Em meios bidimensionais, mostramos os comportamentos das distribuições de tempos de transiente e de período de atrator em função do tamanho do sistema e da memória em ambas as condições de contorno. Mostramos que as grandezas $\langle t \rangle$ e $\langle p \rangle$ não crescem com N , mas apenas com μ . A presença de transição de fase não foi estudada aqui, no entanto, mostramos evidências de que ela exista em um valor de memória da ordem de N . Assim, para valores de memória abaixo da transição, que acreditamos seja $\mu_c \approx \frac{N}{2}$, as trajetórias obtidas são finitas, ou seja, não escalam com N . A partir de $\mu = \mu_c$, a grandeza $\langle t + p \rangle \approx N$. Esta observação terá papel importante na avaliação do algoritmo proposto de agrupamento de dados multidimensionais apresentado no Capítulo 6.

5 *A caminhada do turista em redes regulares*

Neste Capítulo, mostraremos como a caminhada do turista se comporta em redes regulares em ambientes aleatórios e em ambientes estruturados. Consideremos uma rede regular quadrada de lado L e $N = L^2$ sítios. Consideramos na dinâmica somente os primeiros e segundos vizinhos conforme ilustrado na Figura 35. Cada sítio recebe um valor de energia E_i . Neste experimento, testamos três regras de caminhada. Nelas o caminhante move-se sempre (i) na direção de menor gradiente de energia, (ii) na direção de maior gradiente de energia e (iii) para o sítio de menor energia na vizinhança considerada.

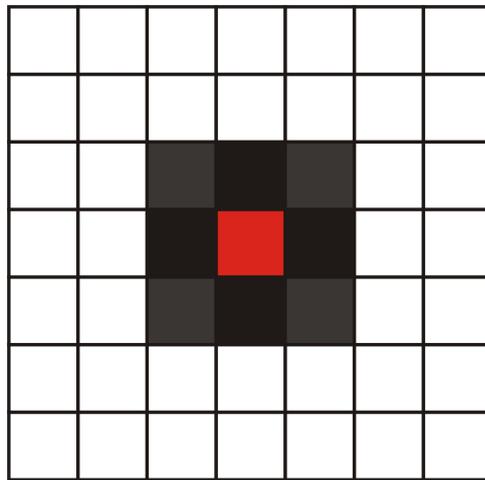


Figura 35: Exemplo de rede regular quadrada. O sítio destacado em vermelho apresenta quatro primeiros vizinhos, destacados em cinza escuro e quatro segundos vizinhos destacados em tom mais claro.

Em caso de empate, o sítio escolhido é o primeiro sítio na direção horária. A Figura 36 ilustra um exemplo de caminhada do turista em uma rede regular usando esta regra de desempate. A regra de desempate pode influenciar nas trajetórias obtidas por caminhadas em redes regulares [13]. Como consequência, as estatísticas estudadas também são influenciadas pela regra escolhida. Nesta tese estudamos apenas a regra descrita.

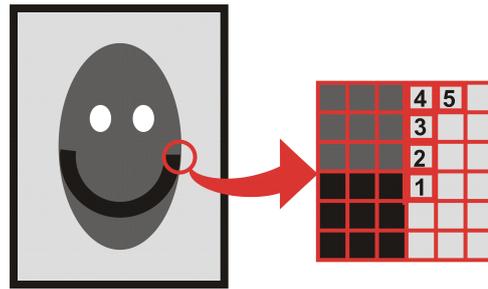


Figura 36: Diagrama mostrando a regra de desempate utilizada nas redes regulares. Neste exemplo, o caminhante inicia o movimento no sítio 1. Pela regra de caminhada utilizada neste exemplo (regra de caminhada (i)), há 5 sítios possíveis para o próximo passo de iteração. O sítio escolhido entre estes será o primeiro sítio que encontrar-se disponível saindo em posição norte e movendo-se em direção horária, neste caso, o sítio de número 2. A mesma regra vale para os próximos 3 passos ilustrados.

Neste sistema consideramos condições de contorno abertas. Esta condição foi escolhida para o tratamento de imagens por ser mais plausível para os fins apresentados neste trabalho. Assim, o número de vizinhos de um dado sítio pode variar de 3, nos cantos da rede, até 8, número máximo permitido neste experimento. Com o aumento da memória μ , a auto-repulsão aumenta e o caminhante é obrigado a percorrer trajetórias mais longas. Imaginando a caminhada realizada sobre a tabela de vizinhanças (Capítulo 2), é fácil perceber que, para memórias acima de 3, algumas caminhadas encontrarão um beco sem saída retornando eventualmente a um sítio cujos únicos possíveis vizinhos já tenham sido todos visitados dentro dos μ passos anteriores. Neste caso, a caminhada é interrompida e o valor -1 é retornado tanto para tempo de transiente como para período de atrator. Vale notar que neste cenário, μ pode assumir valores maiores que 8, no entanto o número de caminhadas resultantes em becos sem saída abrangeria uma proporção muito grande do número total de caminhadas impossibilitando um tratamento estatístico satisfatório. O gráfico mostrado na Figura 37 traz as proporções de caminhadas interrompidas em função de L e de μ para redes aleatórias. Observe que conforme se aumenta o tamanho da rede, a proporção de caminhadas não terminadas torna-se irrelevante.

Realizamos a caminhada do turista em redes aleatórias e em redes que chamaremos estruturadas, ou seja, que exibem um padrão de organização na distribuição dos valores dos sítios. Este último caso é ilustrado com imagens, onde cada sítio é representado por um pixel com um valor inteiro dentro da faixa $[0, 255]$ (256 tons de cinza).

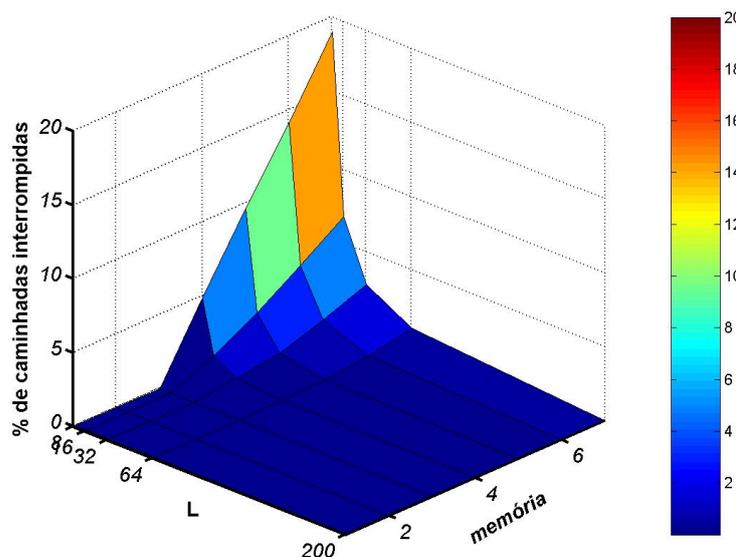


Figura 37: Proporção de trajetórias interrompidas em função de L e μ .

5.1 Redes aleatórias

Nas redes aleatórias, os valores dos sítios são sorteados aleatoriamente dentro do conjunto dos números reais.

Variamos o tamanho das redes, a memória e as regras de caminhadas. Os valores apresentados são a média sobre 1000 realizações.

5.1.1 Regra de caminhada (i) - caminhante se move na direção de menor gradiente de energia

Neste caso, partindo de um dado sítio, o caminhante escolhe, dentro de sua vizinhança o sítio que apresente o menor gradiente de energia com relação ao sítio atual, respeitando a disponibilidade dos sítios dada pela janela de memória. A regra pode ser escrita como:

$$x(t+1) = \min_i (|x(t) - x_i(t)|), \quad (5.1)$$

onde x representa a energia da posição atual e x_i , a energia de seu i -ésimo vizinho. Esta regra foi escolhida para o tratamento de texturas que será apresentado na Seção 5.3. Com esta regra, o caminhante tende a permanecer dentro de uma mesma faixa de valores de energia ou tonalidade. Assim, numa imagem, onde existem padrões espaciais de tonali-

dades, o caminhante encontra uma faixa de tonalidade e permanece ali até que encontre um atrator ou seja forçado a sair.

Apresentaremos a seguir as estatísticas de tempos de transiente e de atrator para esta regra.

5.1.1.1 Distribuição de tempos de transiente

A Figura 38 apresenta a distribuição de tempos de transiente em função do lado L das redes para μ variando de 1 a 7. Utilizamos redes com $L = 8, 16, 32, 64$ e 200. O valor de $L = 200$ foi escolhido para possibilitar a comparação com as imagens que será apresentada na Seção 5.2.

A distribuição de tempos de transiente para redes regulares, assim como em sistemas contínuos, segue uma função exponencial para $\mu > 1$. Vemos que estas distribuições dependem fortemente do tamanho da rede mesmo em memórias pequenas, onde a exploração do meio é baixa. Para $\mu = 2$, o efeito do tamanho deixa de ser importante a partir de $L = 32$. Com o aumento do valor de μ , vemos que as curvas de $P(t)$ deixam de ser exponenciais favorecendo tempos de transiente mais curtos para redes com $L < 200$.

A Figura 39 mostra as distribuições de tempos de transiente para uma rede pequena, de lado $L = 8$ (Figura 39(a)) e uma rede grande, de lado $L = 200$ (Figura 39(b)) em função de μ . Novamente, percebemos o efeito do tamanho da rede bastante pronunciado na rede de $L = 8$. Podemos notar o encurtamento dos tempos de transiente conforme μ cresce. Este efeito não é, no entanto, observado para $L = 200$. A probabilidade de transientes longos cresce conforme aumentamos o valor do parâmetro μ . Concluimos que o comportamento geral de $P(t)$ é exponencial e que desvios em relação a isto se devem a efeitos de tamanho finito.

Investigamos como o tempo médio $\langle t \rangle$ de transiente de um sistema varia com o tamanho da rede. Definimos $\langle t \rangle$ na equação 4.11. Se esta grandeza for proporcional a L , o problema torna-se computacionalmente mais caro para sistemas grandes. A Figura 40(a) mostra como $\langle t \rangle$ varia em função de L para diversos valores de μ . Não foi possível determinar a função que relaciona as duas variáveis. Vemos, no entanto, que $\langle t \rangle$ aumenta inicialmente com L , mas parece convergir para um valor assintótico para redes muito grandes. Acreditamos que o tempo médio de transiente para estes sistemas seja finito, ou seja, não diverge quando $L \rightarrow \infty$.

Na Figura 40(b), mostramos o tempo médio de transiente em função de μ . Vemos

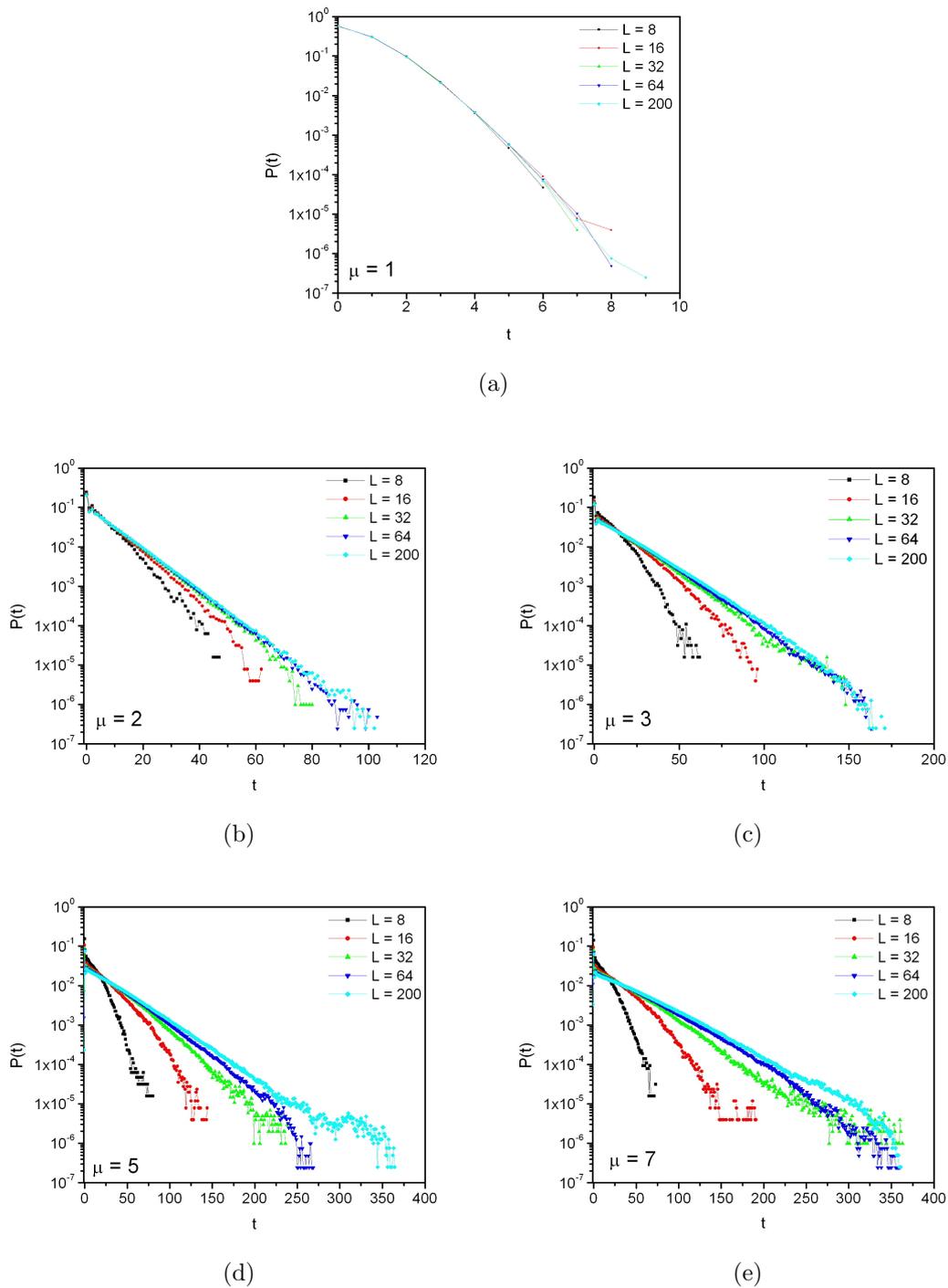


Figura 38: Distribuições de tempos de transiente em função de L para $\mu = 1, 2, 3, 5, 6$ e 7 .

que $\langle t \rangle$ aumenta linearmente com μ . Vale lembrar que em sistemas contínuos em duas dimensões, também observamos um aumento aproximadamente linear para memórias baixas e, em seguida, uma tendência à saturação. Lembramos que no sistema abordado aqui, não é conveniente do ponto de vista estatístico estudar memórias maiores que sete, devido ao grande número de trajetórias terminadas em becos sem saída.

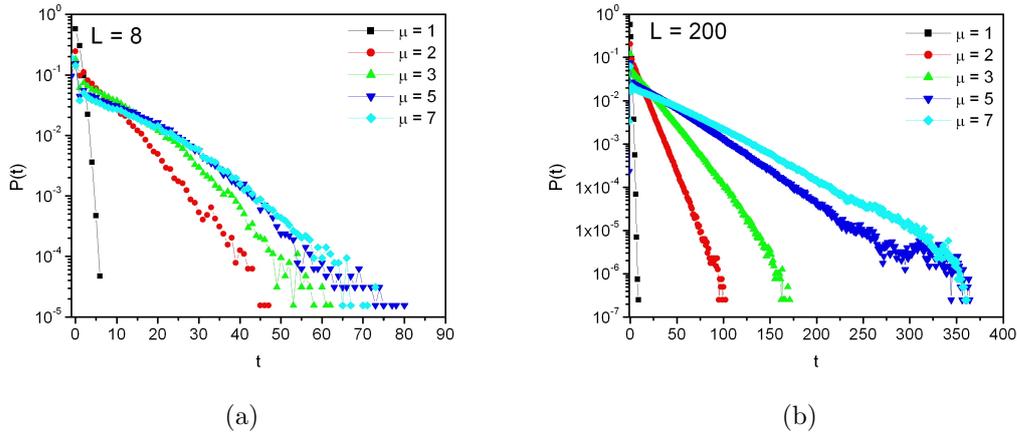


Figura 39: Distribuições de tempo de transiente em função de μ para $L = 8$ e $L = 200$.

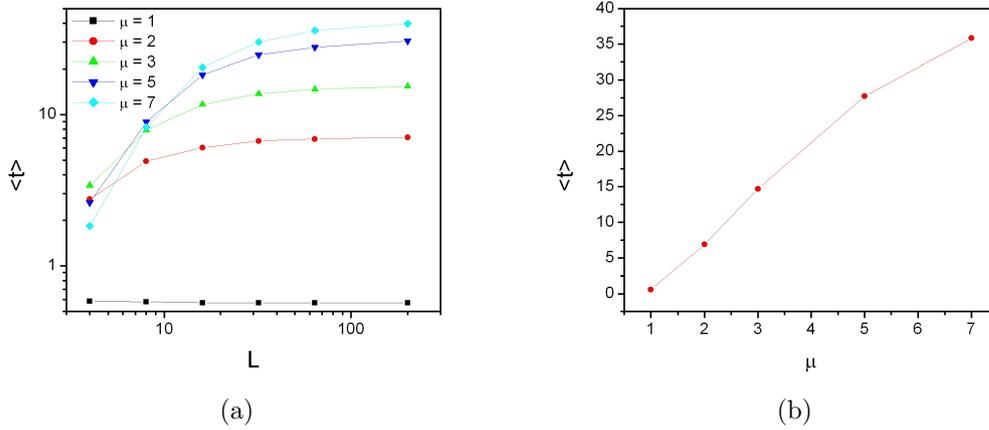


Figura 40: A Figura 40(a) mostra o tempo médio de transiente em função do tamanho da rede para os diversos valores de μ . Vemos um aumento inicial e uma tendência à saturação para redes grandes. A Figura 40(b) mostra a variação de $\langle t \rangle$ em função de μ . O tempo médio de transiente nestes sistemas parece variar linearmente com μ até a memória estudada.

5.1.1.2 Distribuição de ciclos

Apresentamos agora as estatísticas de ciclos das redes quadradas. A probabilidade de um caminhante encontrar um atrator de período p , partindo de um sítio qualquer do mapa é designada por $P_A(p)$ (seção 3.4). Nas Figuras 41 e 42 mostramos a distribuição $P_A(p)$ em função de L para diferentes valores de μ . Ao lado direito, mostramos a distribuição acumulada complementar. Esta figura nos dá a probabilidade de encontrarmos um ciclo de período maior que p , partindo de um sítio qualquer do mapa. Definimos esta grandeza na equação 4.9.

Para $\mu = 2$, observamos a prevalência de períodos pares sobre os ímpares para $p > 7$

como em sistemas contínuos. Não há evidência de lei de potência na distribuição de períodos e podemos notar que as curvas não dependem fortemente do tamanho da rede. Nas distribuições acumuladas (gráficos à direita das Figuras 41 e 42), as curvas tornam-se mais lisas, o que nos permite perceber mais claramente que o desvio apresentado nas curvas de redes pequenas deve-se a um efeito de borda. Este desvio torna-se mais pronunciado com o aumento de μ , mas as curvas continuam a convergir para L grande.

Na Figura 43, mostramos as distribuições $P_A(p)$ em função de μ para uma rede de $L = 64$. O aumento de μ desloca a curva para a direita, ou seja, para períodos maiores. Note que, apesar de trabalharmos com memórias pequenas, $\mu \ll N$, o caminhante explora grande parte do meio e encontra períodos até quase uma ordem de grandeza maiores que a restrição mínima $\mu + 1$.

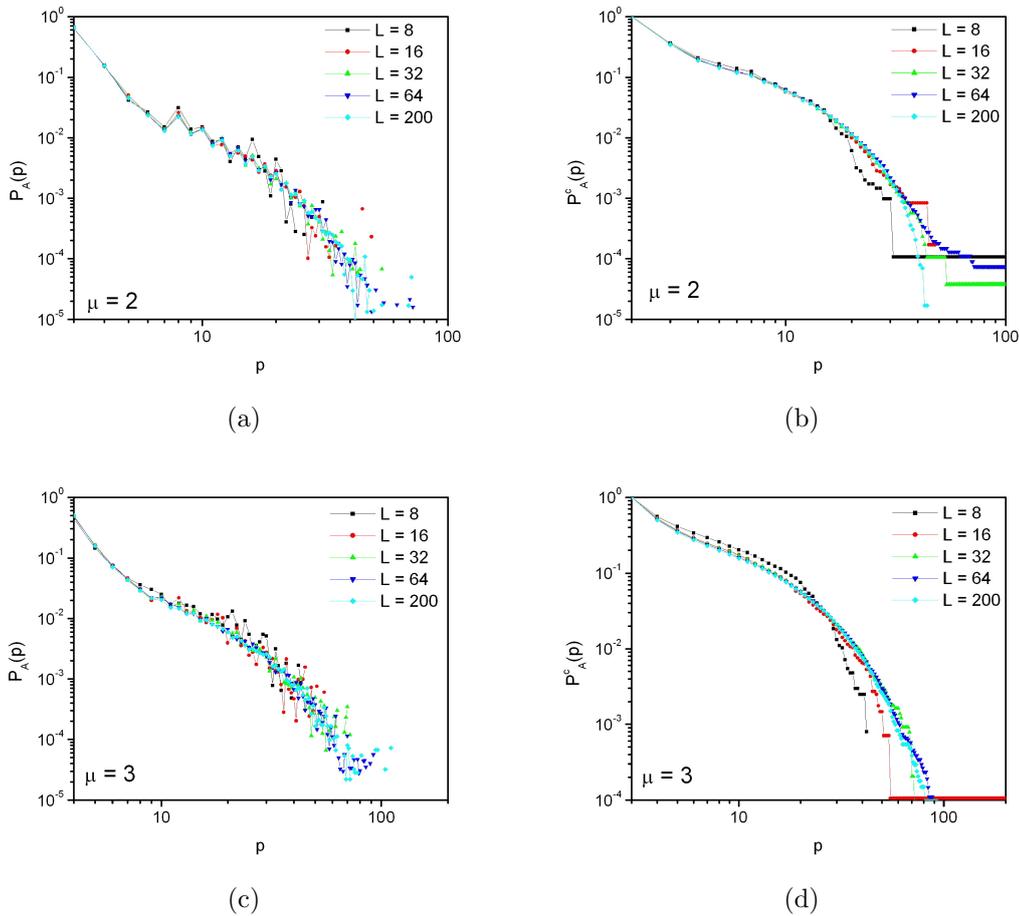


Figura 41: $P_A(p)$ em função de L para $\mu = 2$ e 3. As Figuras 41(b), 41(d) mostram as distribuições acumuladas complementares desta mesma grandeza, definidas na equação 4.9.

A Figura 44(a) mostra a probabilidade $P(p)$ de formação de um atrator de período p em função de μ . Conforme já discutido anteriormente, esta grandeza é conceitualmente

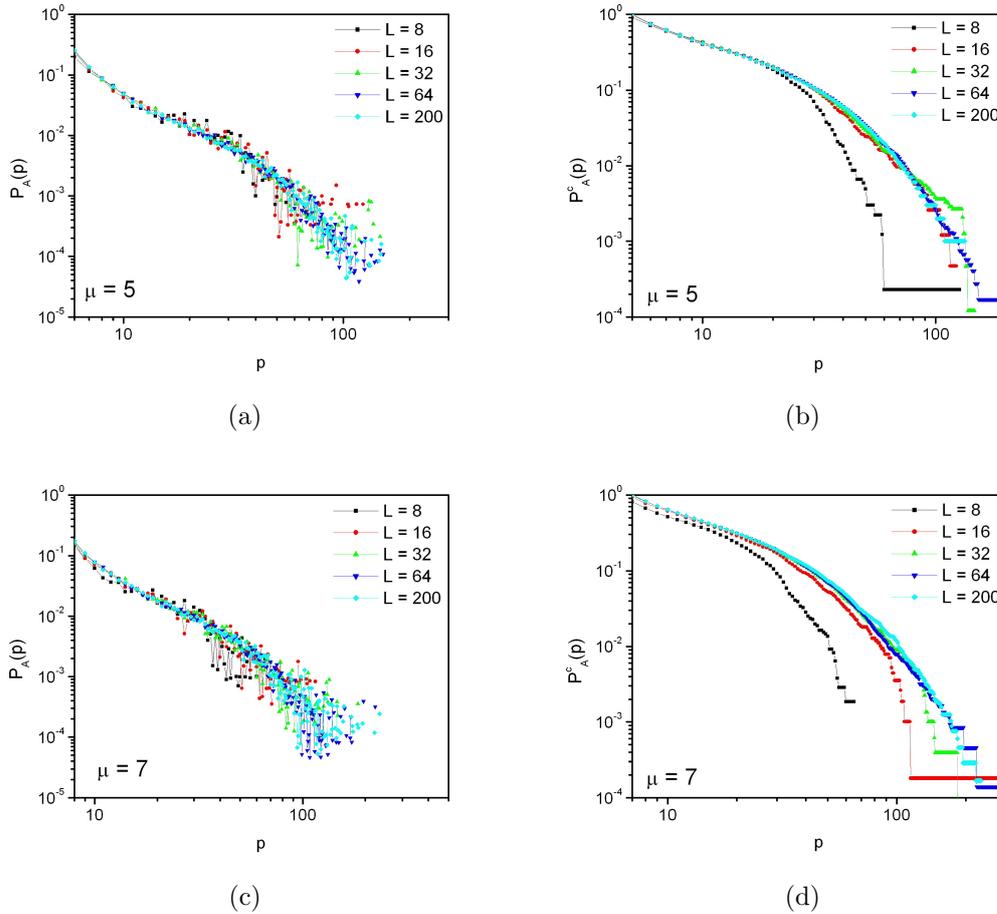


Figura 42: $P_A(p)$ em função de L para $\mu = 5$ e 7 . As Figuras 42(b), 42(d) mostram as distribuições acumuladas complementares desta mesma grandeza, definidas na equação 4.9.

diferente da grandeza apresentada nas Figuras acima ($P_A(p)$), embora as curvas se comportem de maneira muito semelhante. A probabilidade $P(p)$ nos dá informação sobre quantos ciclos diferentes com período p são formados no mapa explorado (Seção 3.4). Conforme p aumenta, diminui-se o número de ciclos formados. Na comparação entre as duas grandezas (Figura 44(b)), vemos um comportamento quase idêntico em ambas as curvas. Enfatizamos que a distribuição $P_A(p)$ é computacionalmente mais barata de se calcular que a distribuição $P(p)$ já que não envolve a comparação entre os ciclos. Dado que as duas distribuições mantêm as mesmas propriedades, em muitas aplicações pode ser vantajoso se utilizar $P_A(p)$ em lugar de $P(p)$.

Na Figura 45(a) vemos como o período médio varia em função do tamanho da rede para os diversos valores de memória. De fato, o período médio não tem relação com o tamanho da rede mas sim, novamente, com o parâmetro de memória. Vale notar que a queda no início das curvas de $\langle p \rangle \times L$ para $\mu \geq 3$ deve-se às caminhadas interrompidas

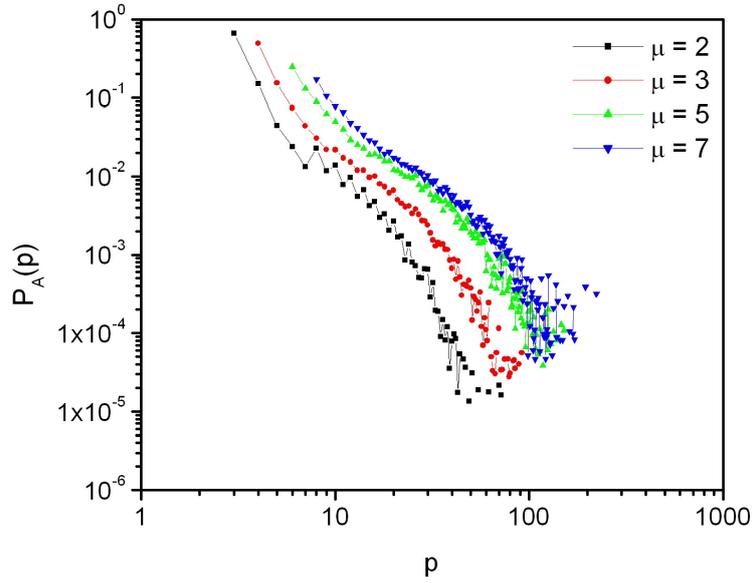


Figura 43: $P_A(p)$ em função de μ para uma rede de $L = 64$.

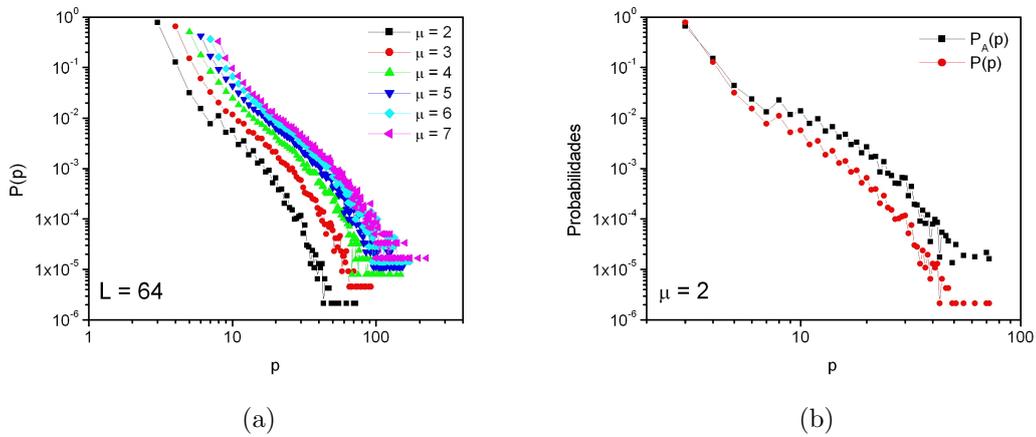


Figura 44: Figura 44(a): Distribuição de ciclos com período p em função de μ para uma rede de $L = 64$. Figura 44(b): Comparação entre as distribuições $P_A(p)$ e $P(p)$.

que diminuem a contagem de ciclos e representam uma proporção relevante para redes pequenas (Figura 37).

O fato de $\langle p \rangle$ não depender de L significa que as caminhadas em redes regulares apresentam um período de atrator típico, ou seja, desde que se trabalhe em redes suficientemente grandes, de modo a anular qualquer efeito de borda, quando $L \rightarrow \infty$, o período médio dos atratores continua finito.

A variação de $\langle p \rangle$ em função de μ está mostrada na Figura 45(b). Vemos uma variação

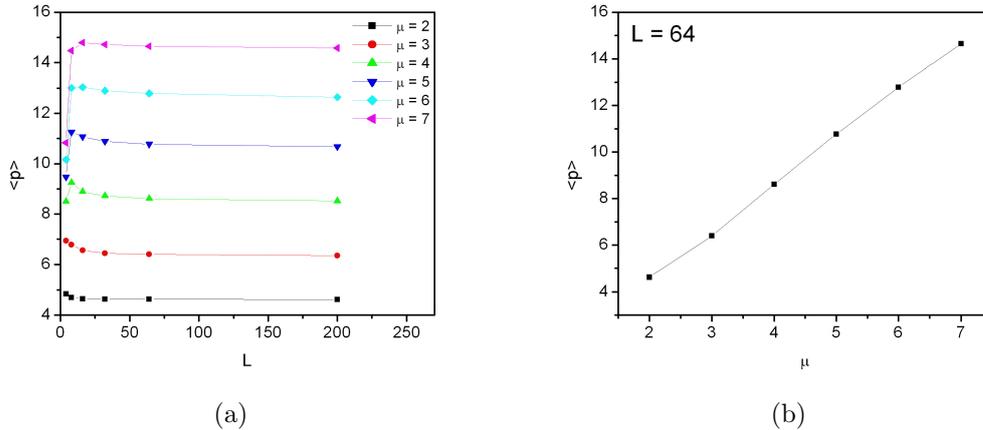


Figura 45: A Figura 45(a) mostra como $\langle p \rangle$ varia em função do tamanho da rede para os diversos valores de μ . As curvas indicam uma independência entre as duas grandezas. A queda observada em redes pequenas para $\mu > 3$ é devida ao grande número de trajetórias interrompidas observadas nestes sistemas. A Figura 40(b) mostra a variação de $\langle p \rangle$ em função de μ . O período médio destes sistemas aumenta linearmente com μ até a memória estudada.

linear do período médio em função da memória, assim como foi mostrado em sistemas contínuos (Seção 4.2).

5.1.1.3 Transiente versus período

Na Seção 4.2 abordamos a questão de uma possível correlação entre os tempos de transiente e os períodos dos atratores formados em uma dada caminhada. Não encontramos correlação nos sistemas contínuos. Nos sistemas abordados nesta seção, também não encontramos correlação entre as grandezas. A Figura 46 mostra o tempo de transiente médio $\langle t \rangle$ em função do período p . Vemos que não há variação significativa no tamanho médio do transiente com relação aos períodos numa faixa de até 3 ordens de grandeza para os diversos tamanhos de rede ou memórias.

A Figura 47 corrobora este resultado. Aqui, plotamos o tempo de transiente contra o período do atrator para cada caminhada. Novamente não há correlação entre as grandezas t e p e a trajetória percorrida $t + p$ depende do valor de μ como mostramos na Seção 4.2.

5.1.2 Regra de caminhada (ii) - caminhante se move na direção de maior gradiente de energia

Apresentamos nesta seção, as estatísticas obtidas para a segunda regra de caminhada. Aqui, o caminhante move-se sempre para o sítio de sua vizinhança que apresentar maior

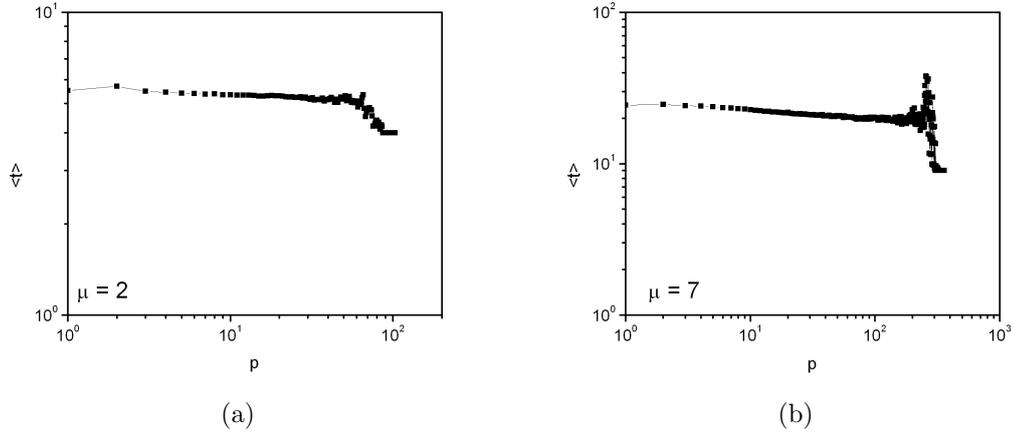


Figura 46: Tempo médio de transiente em função do período do atrator, para uma rede de $L = 64$ e $\mu = 2$ (Figura 46(a)) e $\mu = 7$ (Figura 46(b)). Não há correlação entre tempo médio de transiente e período.

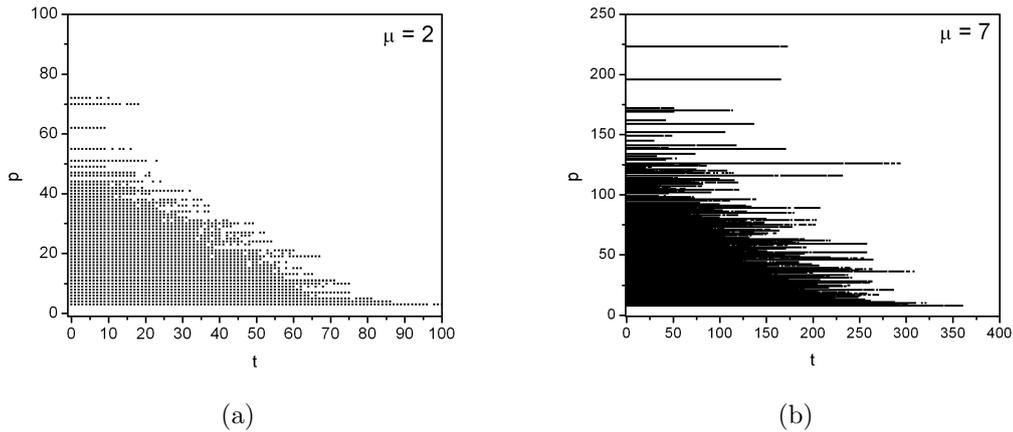


Figura 47: A Figura mostra o período do atrator contra o tempo de transiente de cada trajetória para as caminhadas realizadas em uma rede de $L = 64$ e $\mu = 2$ (Figura 47(a)) e $\mu = 7$ (Figura 47(b)). As curvas são limitadas superiormente pela reta $p + t = c$, e c depende do valor de μ .

gradiente de energia com relação ao sítio em que ele se encontra. Esta regra pode ser escrita como:

$$x(t+1) = \max_i |x(t) - x_i(t)|. \quad (5.2)$$

Vale a pena notar que nesta regra de caminhada, se encararmos os valores dos sítios como cores de pixels, o caminhante ficará alternando entre os extremos de cores de uma dada região conforme o exemplo da Figura 48.

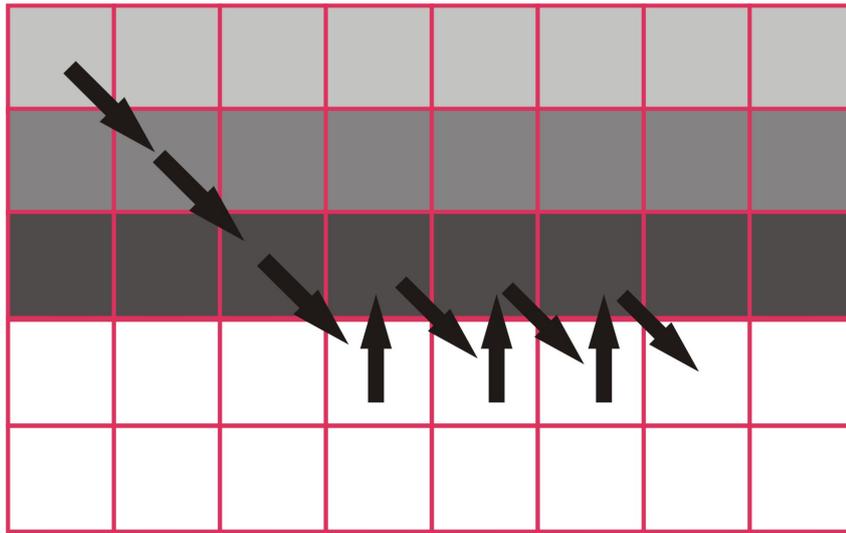


Figura 48: Exemplo de caminhada seguindo a regra (ii). Partindo de um sítio qualquer, o caminhante busca as regiões de maior contraste e passa a alternar entre os extremos.

A Figura 49(a) mostra a distribuição de transientes para uma rede com $L = 64$ e diversos valores de μ . Esta regra de caminhada gera uma condição curiosa de simetria que resulta na coincidência das distribuições com μ e $\mu + 1$, para $\mu > 1$. O mesmo fenômeno é observado para as distribuições $P_A(p)$ (Figura 49(b)).

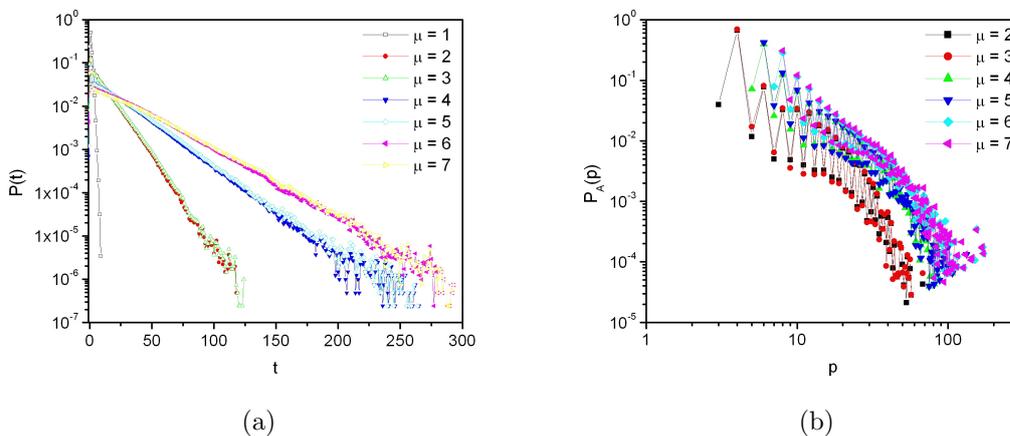


Figura 49: Distribuições $P(t)$ (Figura 49(a)) e $P_A(p)$ (Figura 49(b)) para a regra de caminhada (ii). Vemos o colapso das distribuições para μ e $\mu + 1$.

Observamos que a formação de períodos ímpares nesta regra de caminhada é muito menos provável que a de períodos pares. Podemos considerar como uma boa aproximação, que dentro das regiões de vizinhança de cada vértice da rede, o caminhante move-se alternando entre mínimos e máximos a cada passo (Figura 50). Assim, a probabilidade de que se forme um ciclo que contenha um “canto”, ou seja, três ou mais vértices adjacentes é

pequena. Por isso, as trajetórias das caminhadas com μ e $\mu+1$ para μ par são praticamente idênticas gerando este interessante padrão de distribuições.

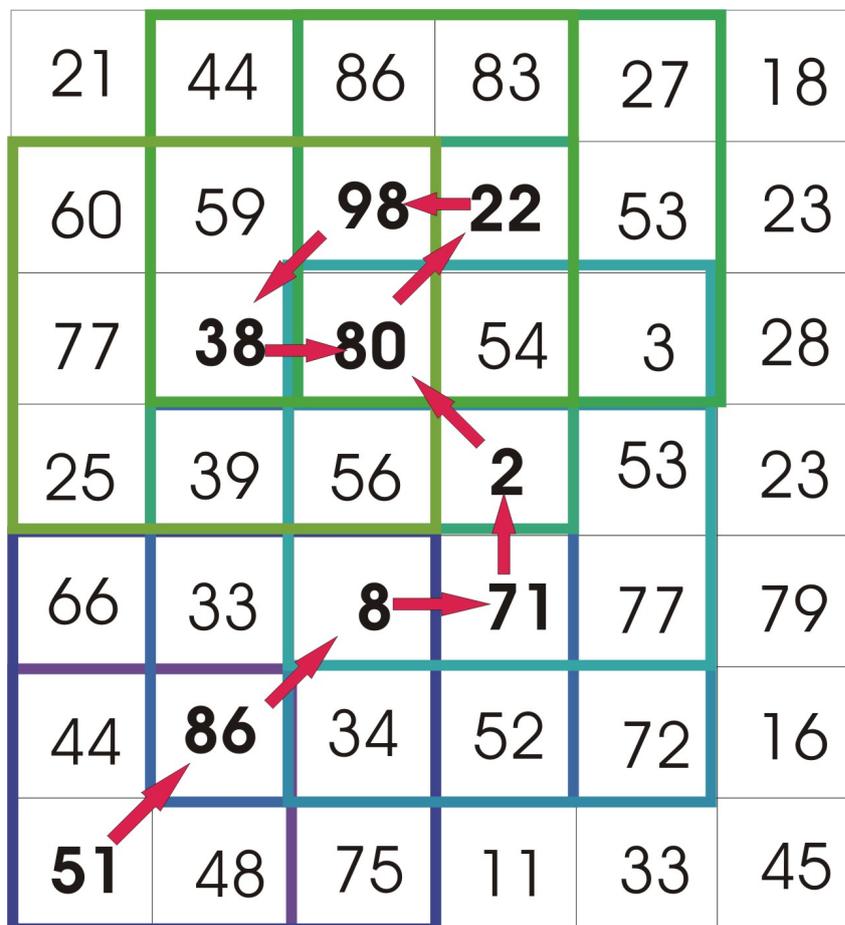


Figura 50: Exemplo de caminhada seguindo a regra (ii). Partindo do sítio com valor 51, as flechas indicam a seqüência de passos até o atrator. A trajetória é a mesma para $\mu = 2$ ou $\mu = 3$.

5.1.3 Regra de caminhada (iii) - caminhante se move na direção do sítio de menor energia

A regra de caminhada apresentada nesta seção é a de que o caminhante move-se sempre para o seu vizinho com menor valor de energia, respeitada a restrição da janela de memória. Nesta regra, o caminhante move-se em direção às regiões de tonalidade mais clara, onde permanece preso. Escrevemos esta regra como:

$$x(t+1) = \min_i(x_i(t)). \quad (5.3)$$

Na Figura 51, vemos as distribuições $P(t)$ (Figura 51(a)) e $P_A(p)$ (Figura 51(b)) para

esta regra. O comportamento observado é semelhante ao observado em sistemas contínuos e na primeira regra de caminhada em redes regulares.

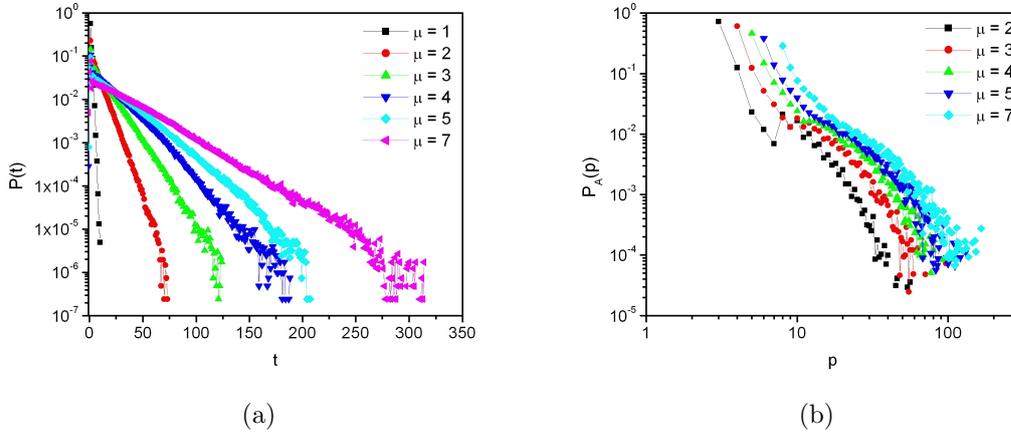


Figura 51: Distribuições de $P(t)$ (Figura 51(a)) e $P_A(p)$ (Figura 51(b)) para a regra de caminhada (iii).

É interessante compararmos as distribuições obtidas para os diferentes sistemas. Na Figura 52, mostramos as distribuições de tempo de transiente (Figura 52(a)) e de $P_A(p)$ (Figura 52(b)) para redes com $L = 32$, $\mu = 2$ e as três regras de caminhada. Mostramos também, na mesma figura as distribuições para um sistema em meio contínuo com $N = 1024$ e mesmo valor de μ .

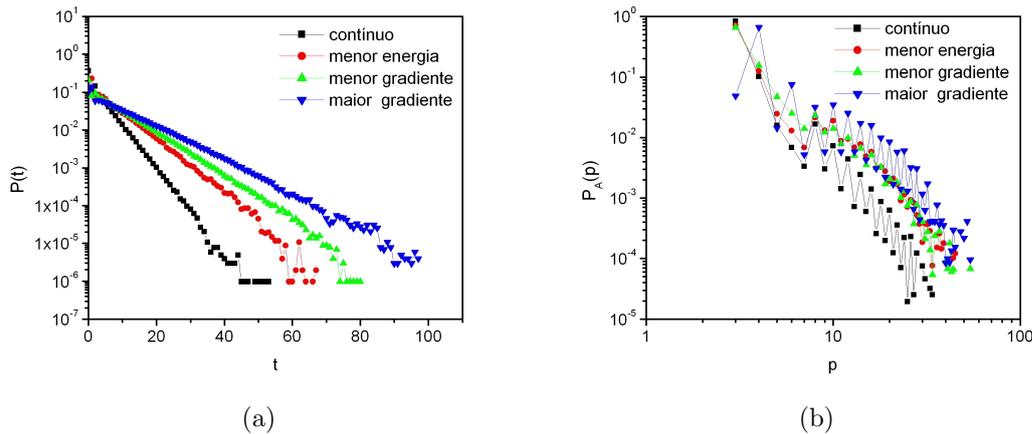


Figura 52: Comparação entre as distribuições de tempos de transiente e $P_A(p)$ para as três diferentes regras de caminhada e um sistema contínuo bidimensional. As redes regulares possuem $L = 32$ e o sistema contínuo, $N = 1024$. $\mu = 2$.

O que vemos é que os tempos de transiente aumentam conforme se muda de sistema na seguinte ordem: contínuo bidimensional, regular com regra de movimentação de menor

energia, regular com regra de movimentação de menor gradiente de energia, regular com regra de movimentação de maior gradiente de energia. O mesmo se repete para as bacias de atração.

5.2 Redes estruturadas - texturas

Nesta seção iremos explorar a caminhada do turista em meios regulares que chamamos estruturados, ou seja, que apresentam um padrão de organização espacial na distribuição de valores dos sítios. Utilizaremos alguns exemplos de texturas retirados de um conjunto de imagens comumente utilizado para a avaliação de técnicas de tratamento de imagens e texturas [34]. Texturas podem ser definidas para os nossos propósitos como combinações de padrões visuais repetidos com uma frequência regular numa escala microscópica. Aqui, o caminhante move-se sobre os pixels e os valores dos sítios são dados por um tom de cinza dentro da faixa $[0, 255]$.

Neste experimento, utilizamos a primeira regra de caminhada (regra (i)), em que o caminhante move-se sempre na direção de menor gradiente de energia. Com esta regra, o caminhante tende a encontrar regiões de mesma tonalidade e aí permanecer. Esta característica permite a detecção de bordas em figuras. A Figura 53 exemplifica este comportamento.

O que vemos na Figura 53(a) é um fantoma, uma figura típica constituída de formas geométricas que se interpõem, preenchidas com tonalidades contrastantes. Estas figuras são comumente utilizadas em avaliação de técnicas de reconstrução de imagens. Aplicamos a caminhada do turista nesta figura e plotamos todas as trajetórias obtidas na Figura 53(b) com μ variando de um a sete. A Figura reproduz claramente a imagem da Figura 53(a) e distingue todas as bordas, mesmo as de regiões de intersecção entre as elipses do fantoma. Como a caminhada é realizada partindo-se de todos os sítios, as trajetórias cobrem toda a figura porém, trajetórias iniciadas em regiões de tonalidades diferentes não se inter cruzam devido à regra de caminhada escolhida, revelando as bordas presentes na imagem.

A seguir, apresentaremos as estatísticas de tempos de transiente e de ciclos para as texturas ilustradas na Figura 54. Amostramos cada imagem em nove regiões não superpostas de lado $L = 200$. As estatísticas apresentadas são a média entre as nove amostras. A Figura 55 mostra as distribuições de tempos de transiente para $\mu = 2$. Vemos aqui que cada tipo de textura apresenta uma curva característica. Algumas texturas apresentam uma acentuado desvio da forma exponencial como as curvas das texturas

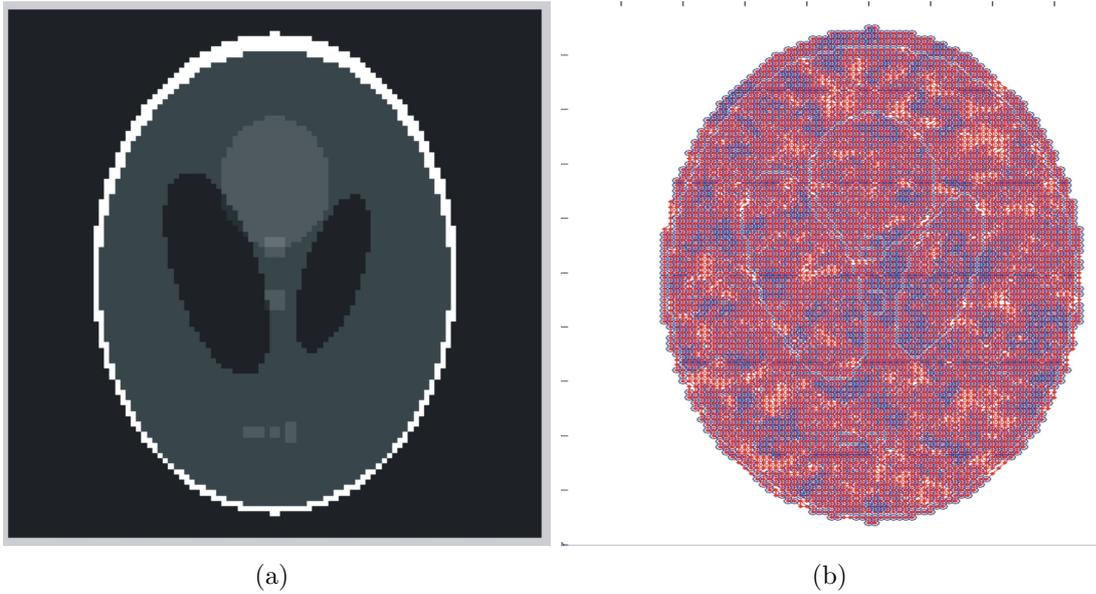


Figura 53: A Figura 53(a) mostra uma figura geométrica utilizada para avaliação de técnicas de reconstrução de imagens chamada fantoma. A caminhada do turista foi aplicada a esta figura e as trajetórias obtidas (trajetória transiente + atrator) estão mostradas na Figura 53(b). As caminhadas foram rodadas com μ variando de um a sete e as trajetórias foram sobrepostas.

(a), (e) e (f) que apresentam um desvio significativo para transientes mais longos. Esta característica de especificidade das curvas de distribuição de transientes será explorada na proposição de uma nova metodologia de análise de imagens (Seção 5.3).

A Figura 56(a) mostra a distribuição $P_A(p)$ para as texturas ilustradas (Figura 54) e $\mu = 2$. Não é possível uma distinção clara das diferentes imagens através das curvas de $P_A(p)$. Na distribuição acumulada complementar mostrada na Figura 56(b), entretanto, torna-se possível distinguir entre alguns grupos de texturas. No entanto, há uma grande sobreposição das curvas. Não pudemos identificar quais características das texturas poderiam ocasionar a sobreposição das curvas de $P_A(p)$.

Mostramos na Figura 57 como as distribuições de transiente e de $P_A(p)$ variam em função de μ . Para efeito de comparação, escolhemos duas das texturas da Figura 54 que julgamos interessantes com este fim: (d) e (e). A textura (d) apresenta um padrão altamente repetitivo e organizado. Ao contrário, a textura (e) apresenta um alto grau de desorganização, ainda que com um padrão bastante distinto do ruído (Figura 58).

Notamos uma diferença drástica nas curvas de distribuição de transiente para cada uma das imagens. Podemos perceber porém, o aumento da probabilidade de transientes longos com o aumento de μ e percebemos que as características de cada imagem se mantêm nas diferentes memórias.

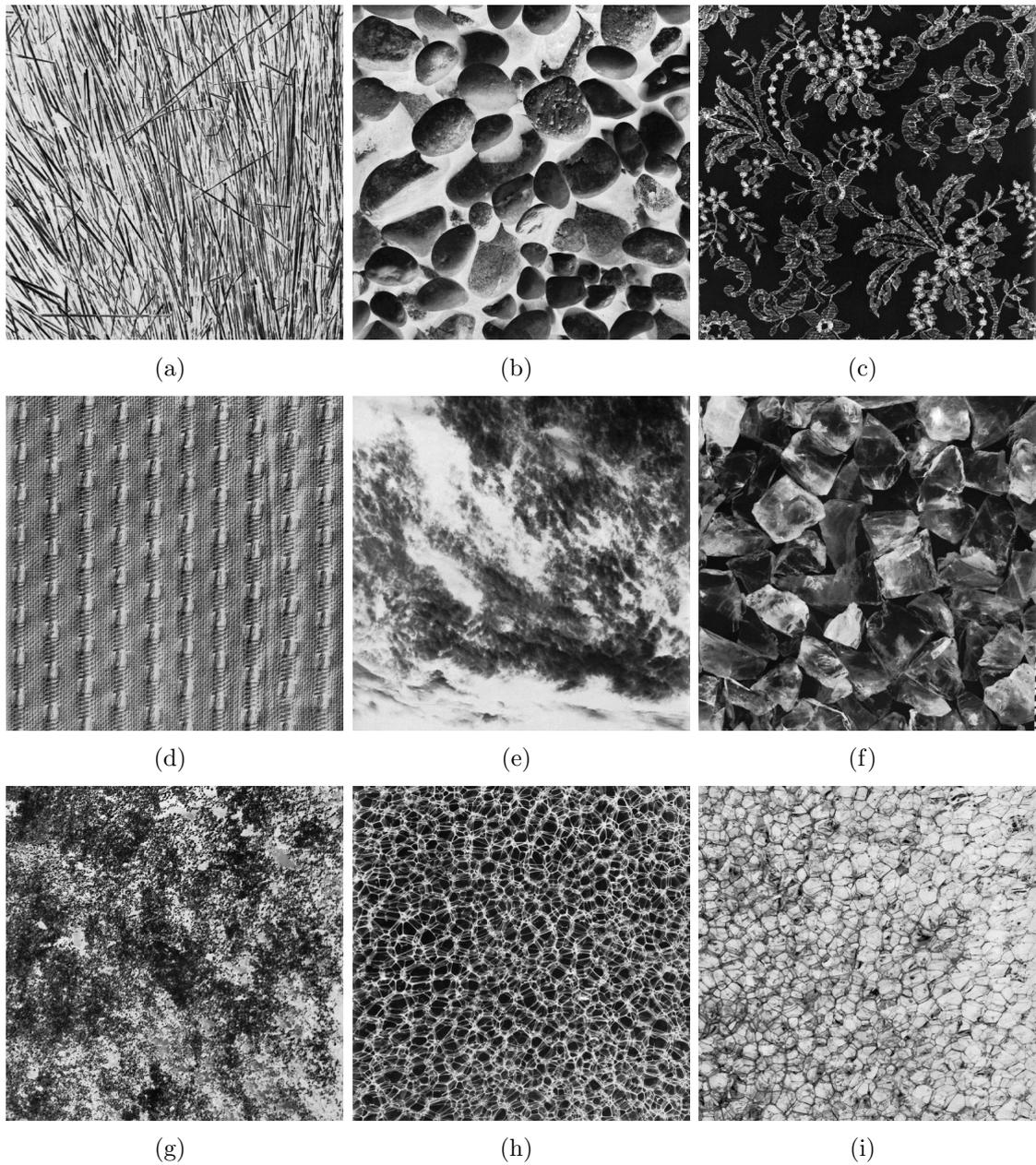


Figura 54: Texturas utilizadas nas estatísticas apresentadas.

Já para as curvas de $P_A(p)$, vemos que não há diferença notável entre as estatísticas de cada textura para cada valor de μ , como já havíamos observado na Figura 56. Notamos um deslocamento das curvas para a direita com o aumento de μ , o que é característico destes sistemas. Note que estes sistemas apresentam períodos maiores que dois para $\mu = 1$. Isto deve-se à regra de desempate aplicada. No caso de empate, bastante frequentes em imagens, o caminhante pode escolher mover-se para outro vizinho ao invés de voltar para o sítio de onde veio. Vale observar que a probabilidade de empates em redes aleatórias tende a zero para valores dentro do conjunto de números reais, por isto não observamos

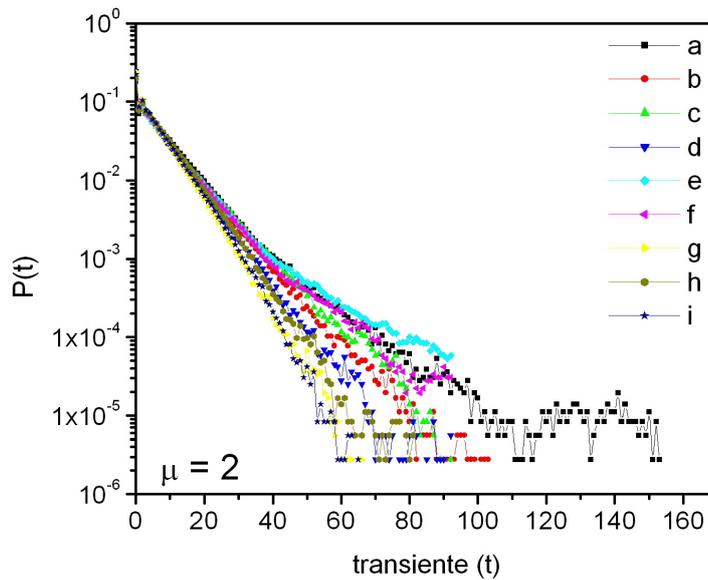


Figura 55: Distribuição de tempos de transiente das texturas da Figura 54.

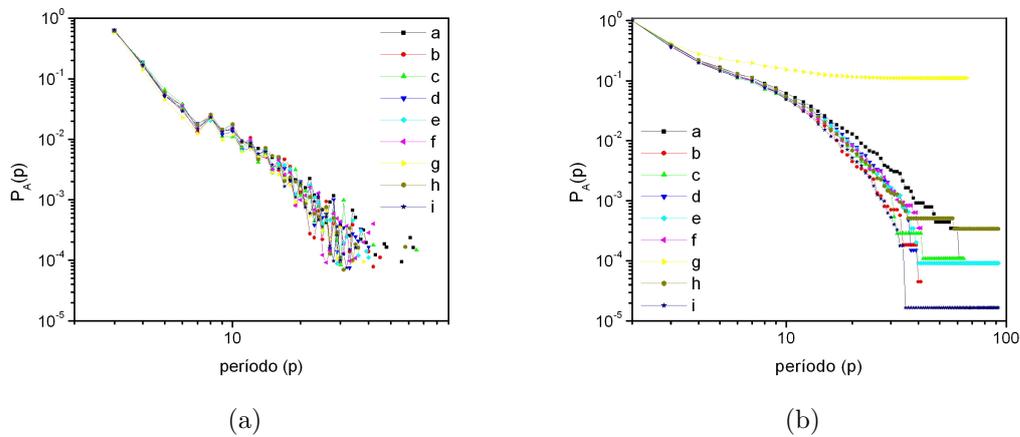


Figura 56: Figura 56(a): $P_A(p)$ das texturas para $\mu = 2$. Figura 56(b): distribuição acumulada complementar de $P_A(p)$ para os mesmos sistemas.

este fenômeno nestas redes.

É interessante saber como as distribuições de tempos de transiente e de $P_A(p)$ em ambientes estruturados se comparam às distribuições relativas a ambientes aleatórios. Na Figura 59, mostramos as distribuições para as texturas (a), (d) e (e) e comparamos as distribuições obtidas com o ruído, uma rede de valores aleatórios sorteados dentro da faixa $[0, 255]$ e de lado $L = 200$. A Figura 58 mostra uma imagem deste tipo.

Podemos observar como as curvas das texturas desviam da forma exponencial das

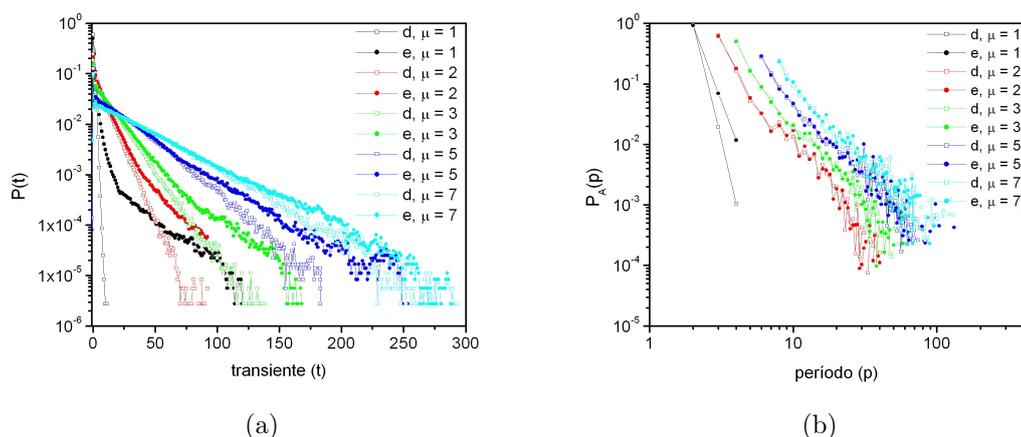


Figura 57: Distribuições de tempos de transiente (Figura 57(a)) e de bacias de atração (Figura 57(b)) para as texturas (d) e (e) (Figura 54) variando-se o valor de μ .

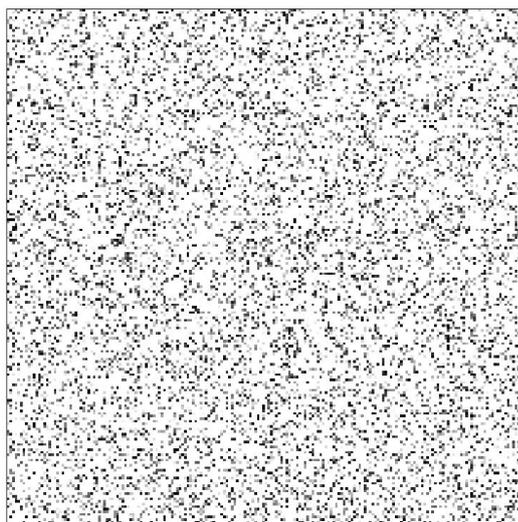


Figura 58: Imagem gerada de uma rede aleatória com valores de pixels sorteados dentro da faixa $[0, 255]$.

distribuições de uma rede aleatória. A textura (d) apresenta tempos de transiente mais curtos com relação ao caso aleatório, enquanto que as texturas (a) e (e) apresentam um desvio acentuado para tempos de transiente mais longos. Cogitamos que a entropia com relação à posição de cada imagem apresente correlação com as distribuições das trajetórias. A entropia é um parâmetro que quantifica o grau de ordem de um sistema, em outras palavras, a presença ou não de um padrão espacial. O caso aleatório situa-se no extremo superior de entropia. Estruturas com alto grau de organização espacial representam o

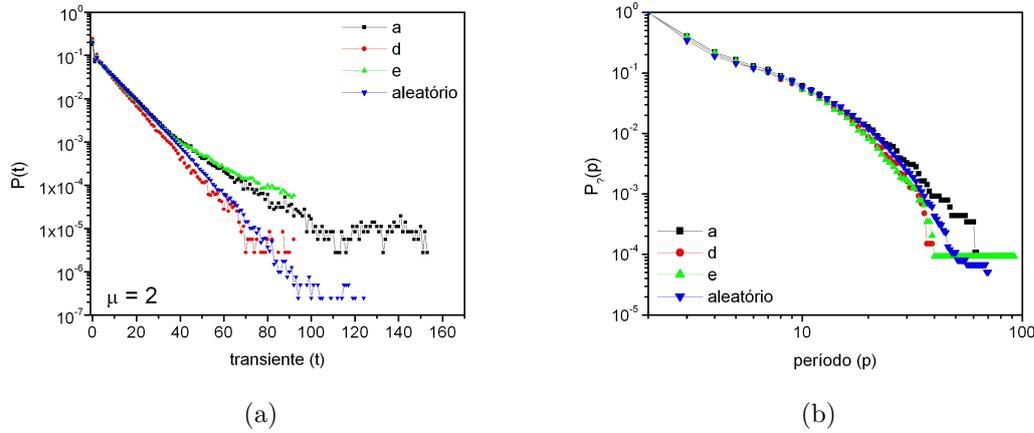


Figura 59: Distribuições de tempos de transiente (Figura 59(a)) e $P_A(p)$ (Figura 59(b)) para as texturas (a), (d) e (e) (Figura 54). No mesmo gráfico estão as distribuições de tempo de transiente e $P_A(p)$ para uma rede aleatória de lado $L = 200$ e com valores dos pixels sorteados dentro da faixa $[0, 255]$.

extremo inferior desta grandeza. A textura (a) apresenta um padrão altamente organizado e, por isso, uma entropia com relação à posição baixa. As demais texturas encontram-se num nível intermediário entre aquela e o ruído. Em [35], mostramos que a caminhada do turista é uma ferramenta capaz de encontrar padrões embutidos em um conjunto de dados (esta metodologia é abordada no Capítulo 6). Estamos propondo aqui que as distribuições de tempos de transiente e de ciclos reflitam de alguma forma este padrão de organização apresentado nas texturas.

5.3 A Caminhada do turista como metodologia de análise de imagens

5.3.1 Motivação

A análise de imagens é um campo da inteligência artificial e da visão computacional responsável pela extração de informação relevante de imagens. A textura é um atributo visual importante utilizado em análise de imagens e possui uma ampla gama de aplicações, tais como: ajuda ao diagnóstico médico, sensoriamento remoto, análise de imagens geológicas e microscópicas.

Embora não haja uma especificação formal de análise de texturas, podemos definí-la como a classificação e segmentação de características extraídas dos padrões visuais com respeito a forma, densidade e direção das regularidades. As texturas podem ser definidas

de acordo com o tamanho do conjunto de pixels analisado. A maioria das técnicas usadas em estudos de extração de informação de imagens concentra-se na análise das micro-texturas [36, 37]. O número de métodos aplicados às macro-texturas ainda é restrito devido à dificuldade inerente à análise [38].

Apresentamos aqui um novo método de caracterização de texturas baseado na caminhada do turista. Exploramos aqui as propriedades das distribuições conjuntas de tempos de transiente e período de atrator obtidas com as imagens e construímos as curvas de assinatura de cada textura. Mostramos que este método é eficiente na exploração da imagem em todas as escalas simultaneamente.

5.3.2 Experimentos

Utilizamos imagens extraídas do livro de Brodatz [34], referência em análise de imagens e realizamos dois experimentos de exploração de texturas com a caminhada do turista. No primeiro, usamos seis tipos diferentes de texturas, cada uma com quatro amostras. Este experimento, com um baixo número de imagens, foi realizado para ilustrar algumas características do método que julgamos interessantes. No segundo experimento, o método da caminhada do turista foi aplicado a 37 diferentes tipos de texturas, cada uma com nove amostras. Neste experimento, pretendemos mostrar o alto potencial do método para analisar e caracterizar imagens. A regra de caminhada utilizada foi a de mover-se para o sítio na vizinhança com menor gradiente de energia.

Cada imagem apresenta 200×200 pixels com 256 tons de cinza. Após extraídas as assinaturas das texturas, uma técnica de análise discriminante foi empregada. A técnica empregada foi a Análise Discriminante Flexível (ADF) [39].

5.3.3 Análise Discriminante Flexível (ADF)

Uma técnica de análise discriminante foi empregada para a classificação dos vetores de dados extraídos com a caminhada do turista. Esta análise foi implementada com o pacote do sistema R 2.1.1 [40] para análise discriminante flexível. A análise discriminante é usada para determinar quais variáveis discriminam entre dois ou mais grupos presentes num conjunto de dados. É comumente usada para a classificação de dados e redução da dimensionalidade. A análise discriminante linear é uma técnica padrão no contexto de análise de dados multidimensionais. Entretanto, o método não funciona muito bem quando as classes de objetos se relacionam através de limites não lineares. Para contornar

esta limitação, foi proposta a técnica de análise discriminante flexível que usa ajustes não paramétricos para conseguir uma classificação mais flexível dos dados.

5.3.4 Resultados

A Figura 60 ilustra as seis classes de texturas utilizadas no primeiro experimento.

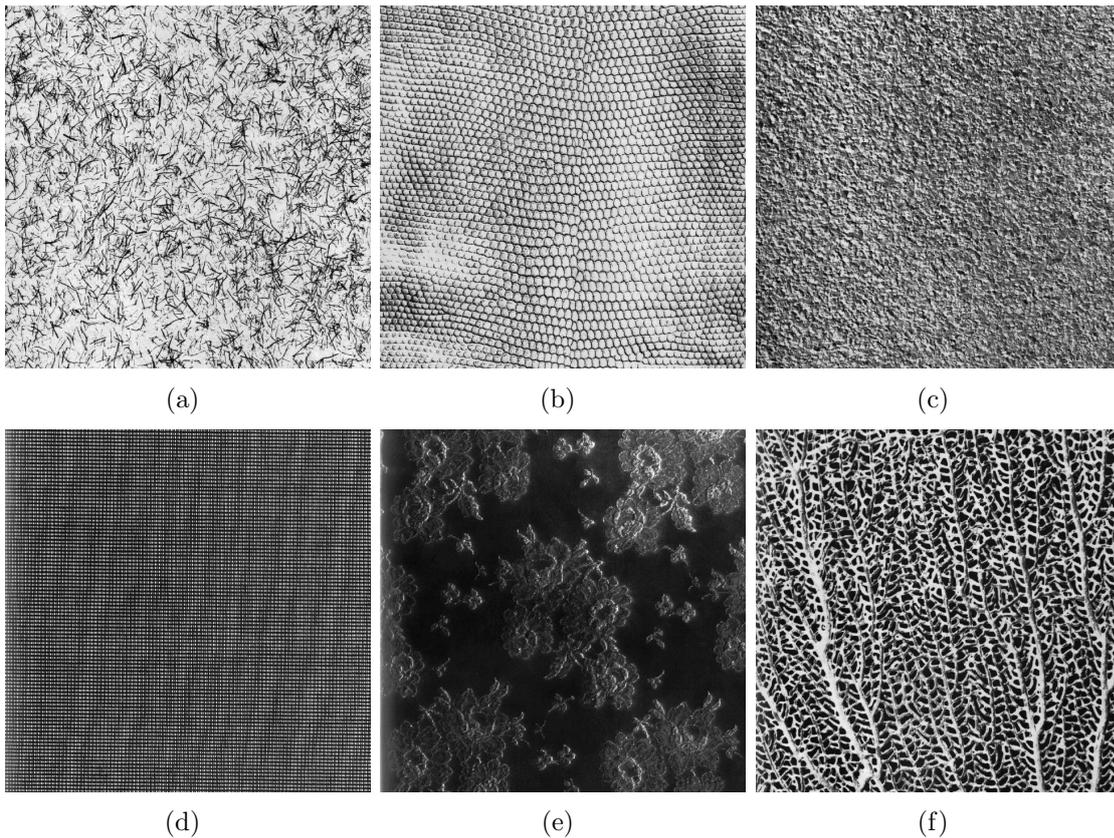


Figura 60: Texturas utilizadas no experimento 1.

Os resultados das distribuições de tempos de transiente para as texturas (b) e (c) estão mostrados na Figura 61 para valores de μ variando de 2 a 10. Vemos que as distribuições de tempos de transiente são fortemente dependentes do valor de μ e que elas se aproximam de uma curva assintótica em $\mu = 10$. Note que valores de memória altos não são convenientes neste estudo devido às trajetórias em beco sem saída discutidas no início deste Capítulo. Apesar de as distribuições dependerem do valor de μ , a Figura 61 mostra que a variação entre as curvas das diferentes texturas ocorre simultaneamente para todos os valores de memória, significando que é possível se distinguir entre as texturas independentemente do valor de μ adotado.

A Figura 62 mostra seis gráficos de superfície das distribuições conjuntas para a

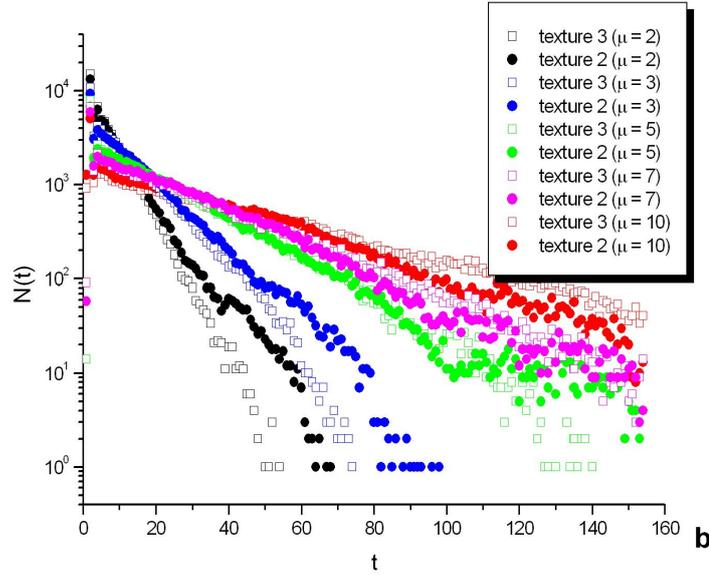


Figura 61: Distribuição de tempos de transiente das texturas 2 e 3 para diferentes valores de μ .

caminhada do turista com $\mu = 7$ correspondentes a cada uma das classes de texturas. Pode-se observar que cada distribuição apresenta um padrão bastante diferente. Isto ilustra o grande potencial da caminhada do turista em fornecer informações para a análise de imagens. Estas distribuições conjuntas foram usadas para caracterizar e classificar as imagens usadas neste experimento. A análise discriminante foi realizada como se segue: um vetor de dados foi construído para cada imagem utilizando-se as frequências do primeiro e segundo tempos de transiente da matriz conjunta. Um vetor de características foi criado, então concatenando-se os valores armazenados na primeira e segunda coluna da matriz conjunta, conforme a equação 5.4.

$$V_{\mu}(m) = [S_{\mu}(0, \mu + 1), S_{\mu}(1, \mu + 1), S_{\mu}(0, \mu + 2), S_{\mu}(1, \mu + 2), \dots, S_{\mu}(0, \mu + m), S_{\mu}(1, \mu + m)] \quad (5.4)$$

A ADF foi aplicada a estes vetores. Os resultados para o primeiro experimento com dois discriminantes estão mostrados na Figura 63 para os diversos valores de μ . Para todos os valores de memória testados, a taxa de acerto na classificação foi de 100%. Além de

classificar os padrões corretamente para os diferentes valores de μ , os resultados demonstram uma outra característica importante da caminhada do turista. Podemos observar que, para cada diferente valor de μ , os padrões de agrupamento dos discriminantes são completamente diferentes. Isto indica que, para cada valor de μ , a caminhada do turista extrai diferentes informações das imagens. Esta propriedade pode ser usada para melhorar a capacidade de classificação do método. Poderia-se compor vetores de dados maiores, formados pela concatenação de curvas de assinaturas (os tempos de transiente) obtidas com diferentes memórias, o que aglutinaria informações de diferentes escalas da imagem em um único vetor de dados.

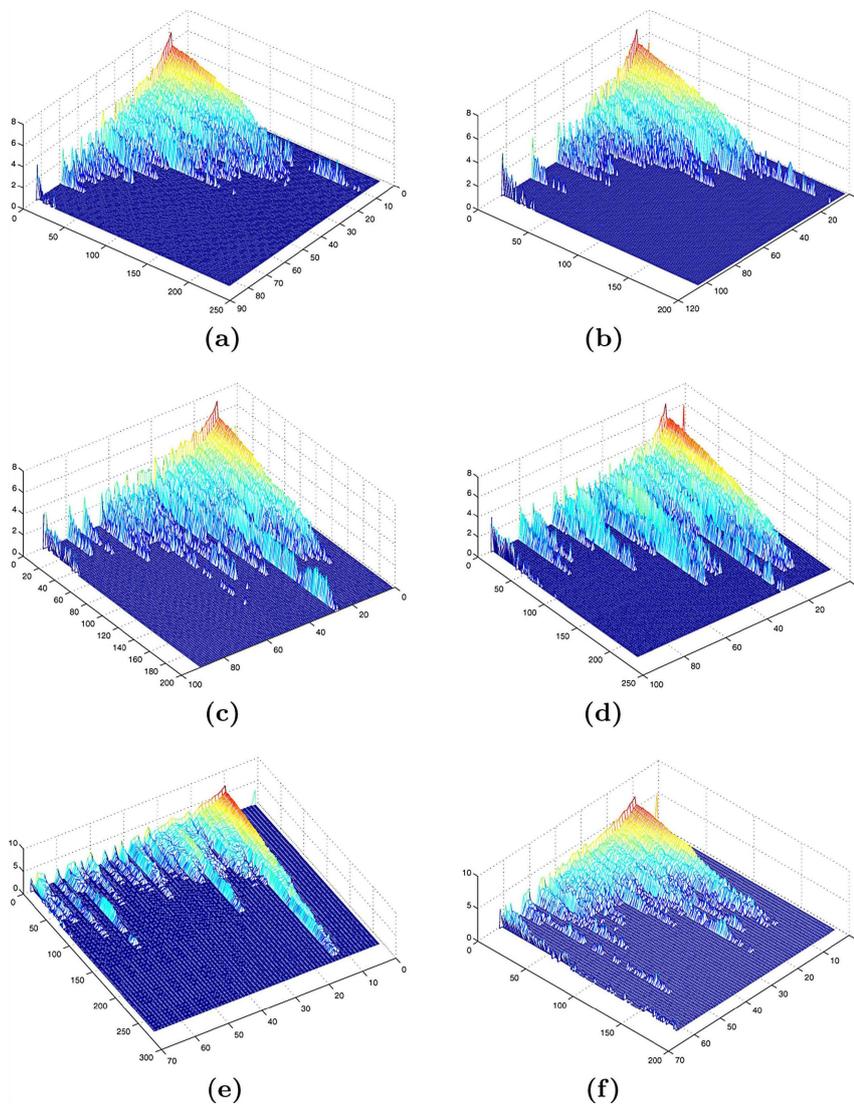


Figura 62: Gráficos de superfície das distribuições conjuntas das texturas ilustradas na Figura 60. Os painéis de (a) a (f) correspondem respectivamente às Figuras 60(a) a 60(f).

O segundo conjunto de experimentos foi realizado com a intenção de testar a caminhada do turista para um grande número de classes. Como no primeiro experimento, foi

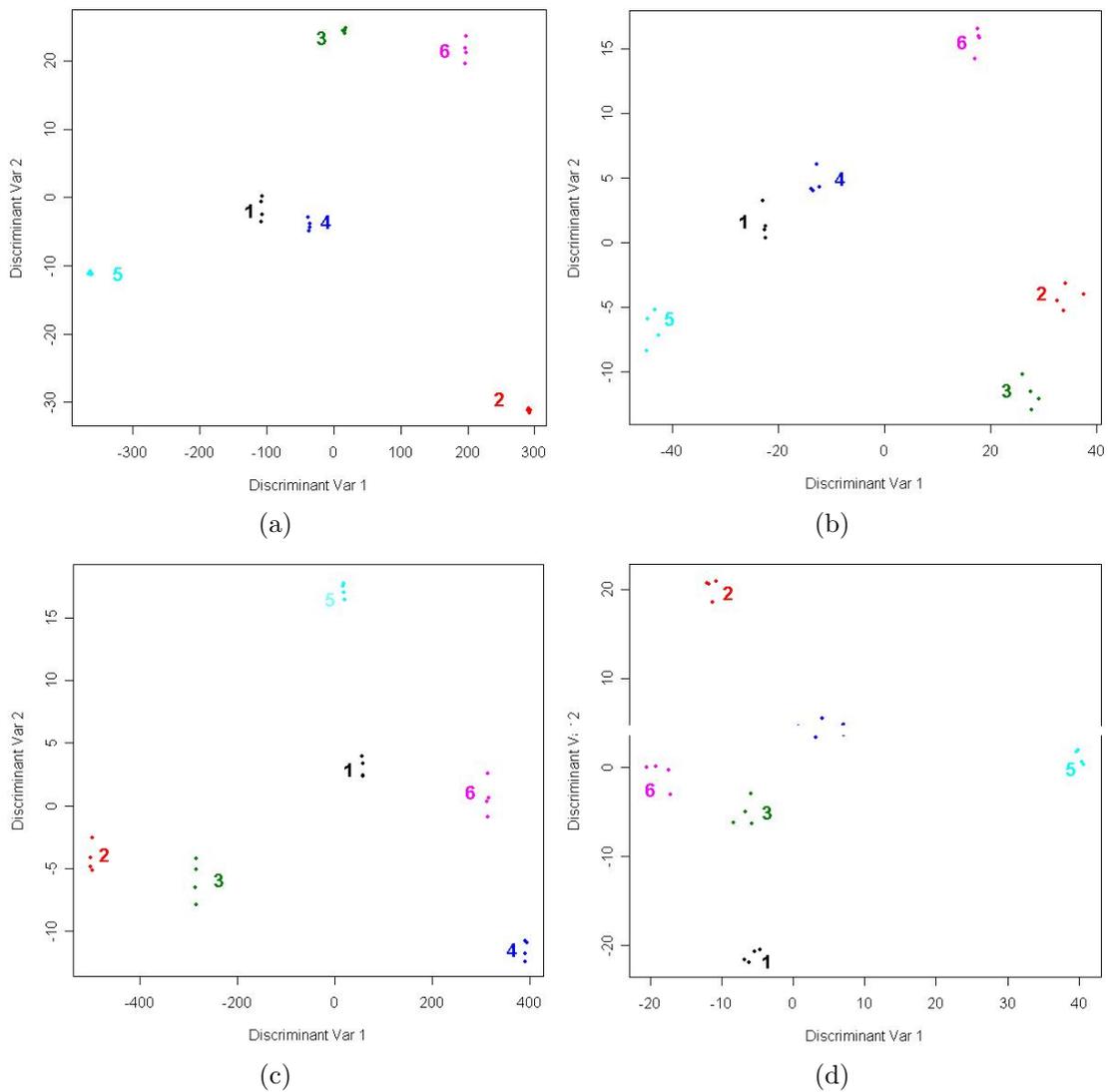


Figura 63: Resultado das análises discriminantes flexíveis do experimento 1 para os valores $\mu = 2$ (Figura 63(a)), $\mu = 3$ (Figura 63(b)), $\mu = 7$ (Figura 63(c)) e $\mu = 10$ (Figura 63(d)).

criado o vetor de dados conforme a equação 5.4 com $\mu = 2$ para cada classe de texturas e a ADF foi realizada sobre estes vetores. 18 discriminantes foram utilizados para a classificação e a taxa de acertos foi de 90.4%. Ressaltamos que a taxa de acertos pode ser melhorada combinando-se resultados para diferentes valores de μ na construção do vetor de dados ou ainda, utilizando-se algum outro método de classificação como redes neurais artificiais [41, 42].

Observamos que, como são realizadas N caminhadas sobre cada imagem e a determinação do transiente envolve $\mu(\tau + 2\langle p \rangle)$, com τ e $\langle p \rangle$ finitos, concluímos que a eficiência do algoritmo é $O(N)$.

5.4 Conclusões

Neste capítulo exploramos as propriedades da caminhada do turista em redes regulares. Até o presente momento a caminhada do turista havia sido explorada apenas em sistemas euclidianos no espaço real, ou seja, os sítios podem assumir qualquer posição dentro do hipervolume tratado. Caminhadas em espaços discretos como redes regulares ou grafos são amplamente exploradas na literatura. Apresentamos aqui, as propriedades estatísticas da caminhada do turista em redes quadradas em função do tamanho das redes, do valor de memória e da regra de caminhada. Mostramos que as estatísticas de tempo de transiente e período de atrator são altamente dependentes da regra de caminhada e que, de maneira geral, as caminhadas em redes apresentam tempos de transiente e períodos mais altos que no espaço contínuo.

Mostramos também que as estatísticas das trajetórias da caminhada do turista refletem a estrutura embutida nos dados. As distribuições de tempos de transiente de caminhadas em redes estruturadas - texturas - apresentam desvios significativos da forma exponencial obtida para redes aleatórias e este desvio é distinto para cada classe de texturas. Esta propriedade motivou o estudo da aplicação da caminhada do turista como método de análise de imagens utilizando as distribuições conjuntas para a extração de assinaturas de imagens.

Os métodos mais comumente utilizados de extração de assinaturas de imagens e texturas abordam escalas definidas de distribuição de pixels. As distribuições de tempo de transiente e período de atrator de um conjunto de dados (no caso, uma imagem) apresenta uma faixa ampla de exploração, capturando detalhes da organização dos pixels das escalas micro à macro e a curva resultante deste método é estritamente relacionada à configuração do conjunto de dados.

Nós mostramos que a distribuição conjunta da caminhada do turista é uma ferramenta eficiente para a classificação de texturas. Os resultados aqui apresentados mostram o grande potencial da caminhada do turista como metodologia de análise de texturas. Estes resultados foram publicados em revista internacional indexada especializada na área [43].

6 A caminhada do turista como algoritmo de reconhecimento de padrões

Com os avanços tecnológicos recentes, a quantidade de dados produzidos diariamente nos centros de pesquisa aumenta dramaticamente em função do tempo. Técnicas automáticas para explorar e identificar correlações e padrões nesta enorme quantidade de dados brutos, frequentemente multidimensionais se tornam essenciais. Neste capítulo apresentamos uma ferramenta baseada na caminhada do turista para o reconhecimento de padrões e agrupamento automático de dados.

Na maioria dos casos, não há conhecimento prévio sobre o conjunto de dados, sendo indispensável o uso de métodos de agrupamento não supervisionados. Tais métodos particionam o conjunto de dados minimizando as dissimilaridades dentro dos agrupamentos e maximizando-as entre os agrupamentos de forma dirigida somente pelo próprio conjunto de dados. Assim, a eficiência do algoritmo reside em grande parte na interpretação correta das relações entre os elementos do conjunto. O procedimento prático para se obter esta relação se faz definindo-se uma medida de distância conveniente que converta-a em valores numéricos. Portanto, agrupar dados frequentemente envolve algumas considerações implícitas acerca da estrutura do conjunto de forma a melhor ajustá-los no espaço apropriado.

O senso comum nos estudos de mineração de dados é o de que não há um procedimento ótimo. Cada problema demanda uma solução ótima e cada técnica utilizada gera diferentes resultados e carrega diferentes limitações. Assim, métodos fáceis de se implementar, rápidos e robustos são ainda bem-vindos. Devido a sua fácil implementação e visualização intuitiva, métodos hierárquicos estão entre os mais populares procedimentos adotados [42, 44, 45]. Eles organizam os dados em uma estrutura hierárquica de acordo com alguma matriz de proximidades. Grupos vão sendo formados por um processo de divisão ou aglomeração de pares dirigido pelas distâncias mínimas. O resultado é uma

árvore binária hierárquica.

Estamos propondo aqui o uso de um algoritmo baseado na caminhada determinista do turista para identificar padrões em conjuntos de dados de maneira hierárquica. Na seção 6.1, fazemos uma breve revisão do método hierárquico mais comumente empregado, o método conhecido como *single linkage* e apresentamos o algoritmo da caminhada do turista como um método de agrupamento automático e o método desenvolvido para a visualização dos resultados - uma árvore geral. Para avaliarmos o desempenho e ilustrar da maneira mais clara as propriedades do método, criamos dois conjuntos de dados artificiais. Um conjunto de dados reais usado como padrão na avaliação de novas técnicas também é exposto. A descrição destes conjuntos é dada na seção 6.2. Na seção 6.3 apresentamos e discutimos os resultados dos métodos abordados comparando suas principais características.

6.1 Algoritmos de agrupamentos de dados

A seguir, discutimos brevemente o método hierárquico tradicional e introduzimos um novo método baseado na caminhada do turista.

6.1.1 Métodos hierárquicos

Os métodos hierárquicos tradicionais podem ser classificados em duas categorias: os métodos aglomerativos e os métodos divisivos, correspondendo a estratégias *de baixo para cima* e *de cima para baixo* respectivamente [46, 47, 45]. Os métodos aglomerativos têm sido mais comumente utilizados e serão o foco deste trabalho.

Nos métodos aglomerativos, cada elemento é associado a um agrupamento e a matriz de distâncias é calculada. A métrica utilizada depende da conveniência do problema. A característica essencial deste algoritmo é que os dois agrupamentos mais próximos nesta matriz são colapsados resultando em um único ponto. A cada passo de iteração, a matriz de distâncias é atualizada e o processo continua até que haja apenas um agrupamento contendo todos os pontos do conjunto. A Figura 64 ilustra este procedimento.

O resultado de um algoritmo hierárquico é uma árvore binária, representando agrupamentos de padrões e níveis de similaridade onde os grupos são definidos. A árvore é definida em um par de eixos cartesianos e pode-se definir a distância entre dois dados grupos pelo valor correspondente no eixo y . Uma árvore hierárquica é uma ferramenta muito

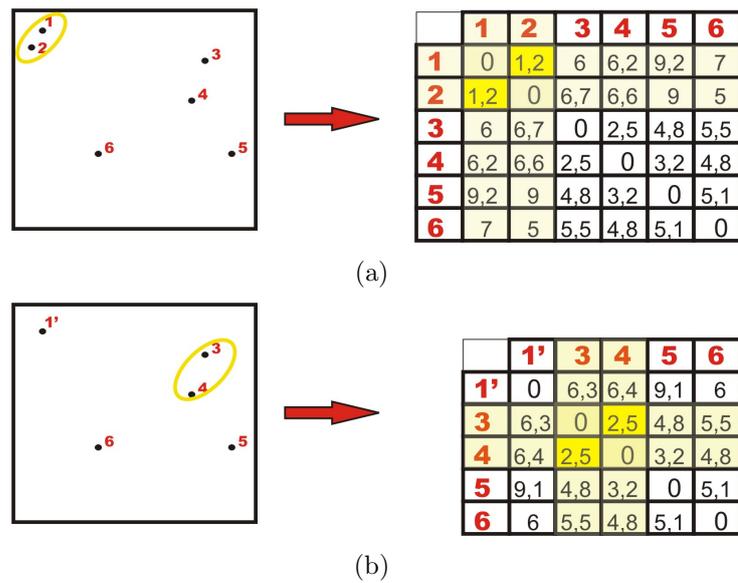


Figura 64: Exemplo de um conjunto de dados e a matriz de distâncias associada a ele. No método hierárquico aglomerativo tradicional, a cada passo de iteração o par mais próximo é identificado. Este par é então colapsado em um único ponto e as linhas e colunas da matriz correspondentes a eles são atualizadas (Figura 64(b)). Este processo se repete até que haja apenas um elemento na matriz. Neste exemplo, as distâncias foram atualizadas baseando-se na média entre as distâncias de cada ponto original.

útil para abstrações acerca da estrutura do conjunto de dados, especialmente se relações hierárquicas estão realmente presentes entre os dados. O resultado final dos agrupamentos pode ser obtido cortando-se os ramos da árvore em diferentes níveis de dissimilaridade de acordo com o problema considerado (Figura 65).

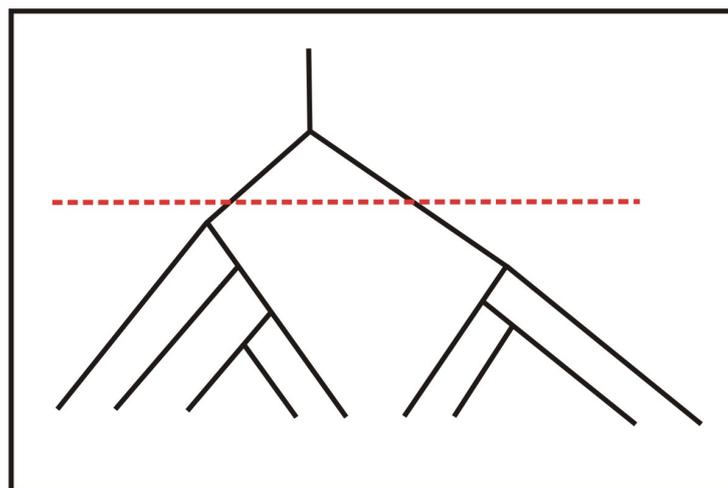


Figura 65: Exemplo de uma árvore hierárquica binária com dois agrupamentos identificados pela linha pontilhada.

A definição de proximidade para um par de agrupamentos depende da natureza dos

dados. Um dos algoritmos mais utilizados é o chamado *single linkage*. Os artigos [46, 47] trazem uma revisão sobre as principais medidas de similaridade entre pares de grupos de objetos. Para o método de *single linkage*, a distância entre dois agrupamentos é determinada pela menor distância entre os objetos de cada agrupamento. Entretanto, todos os algoritmos agem da mesma maneira, coalescendo os pares de objetos a cada nível da hierarquia, impondo uma estrutura de pares ao conjunto de dados.

A complexidade deste algoritmo é da ordem de N^2 , considerando-se a matriz de distâncias como dado fornecido ao algoritmo.

6.1.2 Método baseado na caminhada do turista

Conforme já discutimos nas seções anteriores, a dinâmica determinista da caminhada do turista impõe que o caminhante movimente-se por pontos espalhados em um espaço d -dimensional respeitando a regra de se deslocar sempre para o ponto mais próximo ao ponto onde ele se encontra que não tenha sido visitado nos últimos μ passos. Todos os movimentos do caminhante são realizados sobre uma tabela de vizinhanças que desconsidera a distância entre os pontos mas mantém os ordenamentos de vizinhança dentro da janela de memória considerada.

No contexto de identificação de agrupamentos, pode-se considerar cada atrator formado num dado valor do parâmetro μ como um agrupamento natural no conjunto de dados. O aumento de μ leva a um aumento da auto-repulsão e, conseqüentemente do período médio e os agrupamentos tendem a coalescer. Este processo é ilustrado na Figura 66. Note que os agrupamentos apresentam pelo menos $\mu + 1$ elementos em contraste com as interações binárias dos modelos tradicionais.

A visualização do processo de agrupamentos pelo algoritmo da caminhada do turista é uma árvore geral, similar às geradas nos métodos hierárquicos tradicionais, porém não restrita a qualquer estrutura binária ou n -ária. Os níveis de hierarquia representam a janela de memória em que o agrupamento se formou ao invés de uma medida de dissimilaridade.

O algoritmo para a construção da árvore está ilustrado na Figura 67. A Figura 67(a) mostra um mapa bidimensional com 20 pontos distribuídos aleatoriamente. Na construção da árvore (Figura 67(b)), pontos que pertencem a um mesmo atrator são considerados como um agrupamento no nível de hierarquia em questão, ou seja, no respectivo valor de μ . Para $\mu = 0$, cada ponto representa um atrator. Assim, tem-se N agrupamentos no primeiro nível hierárquico. Para $\mu = 1$, pares de vizinhos mutuamente mais próximos

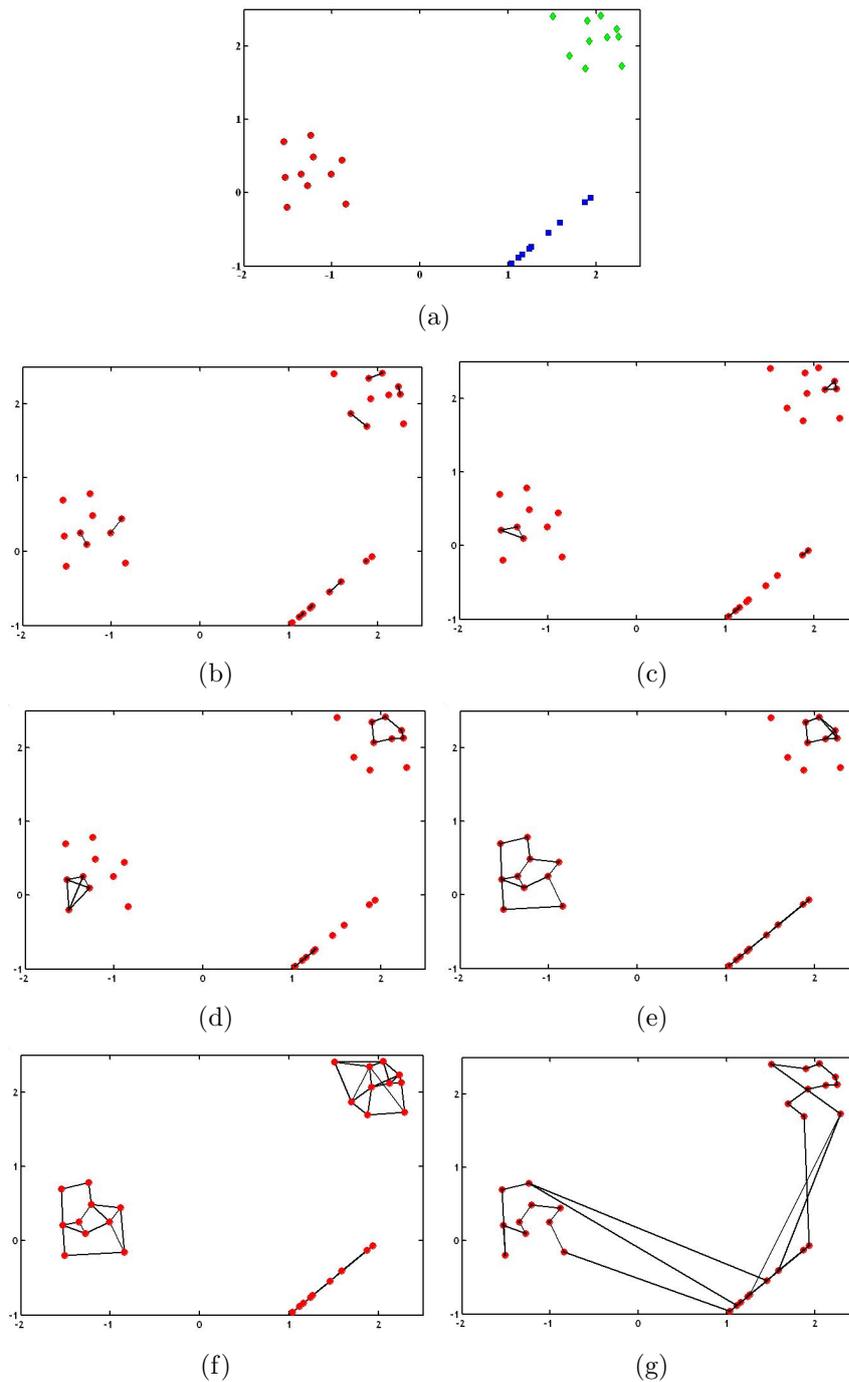


Figura 66: Exemplo de um conjunto de dados bidimensional com três agrupamentos bem definidos (Figura 66(a)). As Figuras 66(b) a 66(g) mostram os atratores formados no conjunto para os seguintes valores de memória $\mu = 1, 2, 3, 5, 7$ e 15 , respectivamente.

são os novos atratores. Note que este passo não é equivalente ao algoritmo de *single linkage* já que este último apenas considera um único par de pontos mutuamente mais próximos a cada passo de iteração. À medida em que os valores de memória aumentam, novos atratores são encontrados. Como enfatizamos na Seção 4.2, estes novos atratores são formados independentemente dos resultados obtidos nos passos anteriores. Cada novo

atrator pode conter uma parte ou todo de um atrator precedente. Se esta sobreposição ocorre, os agrupamentos são colapsados e forma-se um nó na árvore hierárquica. Este processo de aninhamento prossegue até que todo o conjunto de dados esteja contido em um único agrupamento. É interessante enfatizar que este agrupamento único ocorre geralmente para valores de memória muito menores que $N - 1$ para dados organizados, ou seja, com um padrão implícito (o Capítulo 4 discute a percolação em sistemas contínuos).

A seguir, discutimos a complexidade do algoritmo proposto. Para a construção da árvore, incrementamos o valor de μ até que todo o conjunto de pontos esteja contido em um único grupo. Ao valor de μ em que isto acontece chamaremos μ_{max} . Para cada valor de μ incrementado, N caminhadas são realizadas. Uma caminhada termina quando se determina o atrator. O número médio de passos até que se determine o atrator é $(\tau + \langle p \rangle + \mu)$ (ver Seção 3.3 para detalhes sobre a implementação do algoritmo da caminhada do turista). Assim, podemos calcular que, para se determinar todos os atratores de um conjunto de dados de N pontos incrementando o valor de μ até μ_{max} , são necessárias $[\mu_{max}N(\tau + \langle p \rangle + \mu)\mu]$ operações. Note que multiplicamos a expressão por μ . Isto se deve à restrição imposta pela janela de memória que obriga, no pior dos casos a verificação de μ pontos até que se encontre um vizinho disponível. Como μ e $(\tau + \langle p \rangle + \mu)$ são valores finitos, podemos afirmar que para:

$$\begin{cases} \mu_{max} < \mu_c \longrightarrow O(N^2); \\ \mu_{max} = \mu_c \longrightarrow O(N^4). \end{cases} \quad (6.1)$$

Conforme discutimos na Seção 4.2, acreditamos que em sistemas bidimensionais, a memória crítica seja da ordem de N . Neste caso, $(\tau + \langle p \rangle + \mu)$ e μ também ficam da ordem de N , por isso, para casos em que $\mu_{max} = \mu_c$, a complexidade do algoritmo chega a N^4 .

Entretanto, para cada valor de μ incrementado necessitamos fazer as comparações entre os atratores encontrados para estabelecer os nós da árvore. Note que esta comparação se dá entre os atratores formados em um dado valor de μ e entre estes e os atratores formados em $\mu - 1$. Assim, devemos também considerar o número de operações para esta etapa do algoritmo como:

$$(n_\mu p_\mu + n_{\mu-1} p_{\mu-1}) n_\mu p_\mu \quad (6.2)$$

onde n_μ é o número de atratores formados para um dado valor de μ e p_μ é o período médio

$$\begin{cases} \mu_{max} < \mu_c \longrightarrow O(N^2); \\ \mu_{max} = \mu_c \longrightarrow O(N^5). \end{cases} \quad (6.3)$$

6.2 Os conjuntos de dados

Para os experimentos numéricos, criamos dois conjuntos de dados artificiais que pudessem ressaltar as principais características do algoritmo proposto. Também estudamos o conjunto de dados reais conhecido como *Íris* [48]. Os conjuntos são descritos a seguir.

A construção das tabelas de vizinhança (para a caminhada do turista) e das matrizes de distâncias (para o *single linkage*) foi baseada na distância Euclideana entre os vetores de dados para os três conjuntos testados. Empates foram resolvidos arbitrariamente, dando-se prioridade ao vetor que primeiro aparecer no conjunto de dados.

1. Este conjunto de dados consiste de 33 elementos em espaço bidimensional organizados em quatro agrupamentos bem definidos e um elemento equidistante dos outros quatro grupos. Os dados são mostrados na Figura 68.

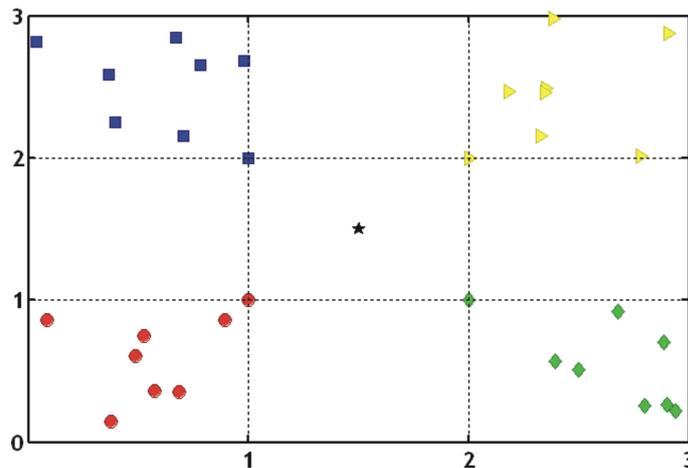


Figura 68: Conjunto de dados 1. 33 elementos organizados em quatro grupos de 8 elementos cada equidistantes de um ponto central.

2. Este conjunto de dados consiste de 100 elementos em espaço bidimensional organizados em dois agrupamentos. Os dois agrupamentos são equivalentes mas representados em escalas diferentes. O maior representa uma ampliação de dez vezes do menor grupo. O conjunto está mostrado na Figura 69.

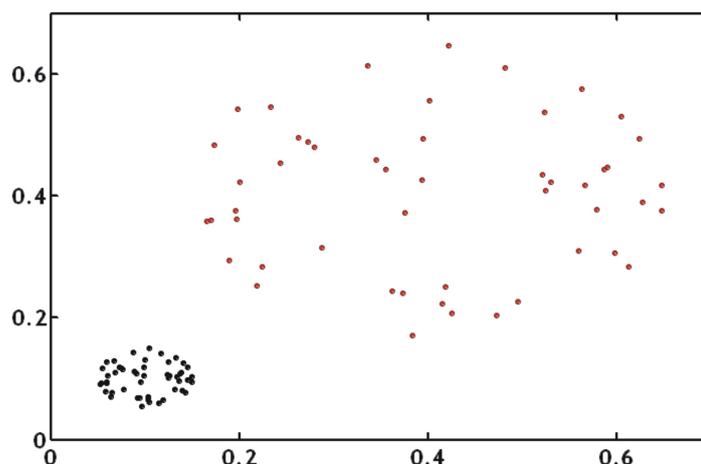


Figura 69: Conjunto de dados 2. O conjunto consiste de 100 elementos dispostos em dois grupos de 50 elementos cada. Os dois grupos são idênticos na organização dos pontos, diferenciados entre si apenas pela escala.

3. O conjunto de dados Íris ([48]) é um conjunto de dados real comumente utilizado como *benchmark* em mineração de dados. Consiste de 150 amostras de flores de íris com as respectivas medidas de comprimento e largura de pétalas e sépalas, imersos, portanto num espaço de 4 dimensões. Estas flores estão divididas em três subespécies ou categorias (Figura 70). Para permitir a visualização do conjunto, os dados estão representados nas duas componentes principais na Figura 70. A análise de componentes principais é uma análise estatística multivariada típica que reduz a dimensionalidade do conjunto de dados. Tecnicamente, consiste de uma transformação linear ortogonal que projeta os dados em um novo sistema de coordenadas de tal forma que a maior variância se dê na primeira coordenada (chamada componente principal), a segunda maior variância, na segunda coordenada e assim por diante.

6.3 Resultados

Ambos os algoritmos abordados foram aplicados aos conjuntos de dados 1 – 3 resultando nos dendrogramas mostrados nas Figuras 71 a 73, respectivamente.

A Figura 71 mostra uma grande similaridade na estrutura dos agrupamentos obtidas com ambos os métodos. Entretanto, conforme discutido anteriormente, o método de *single linkage* impõe uma estrutura hierárquica binária aos agrupamentos, enquanto que o

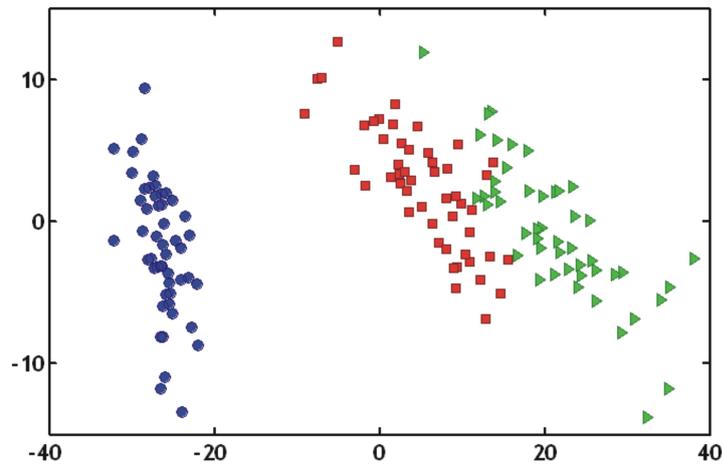


Figura 70: Conjunto de dados 3. Os dados originais estão num espaço de $d = 4$. A Figura mostra a projeção dos dados nas duas componentes principais. As cores representam as três diferentes categorias para este conjunto de dados.

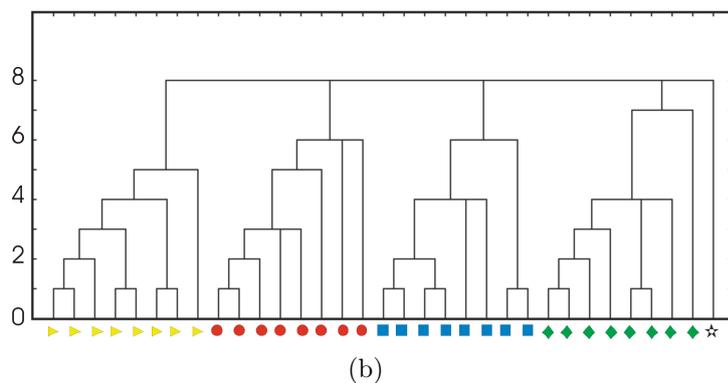
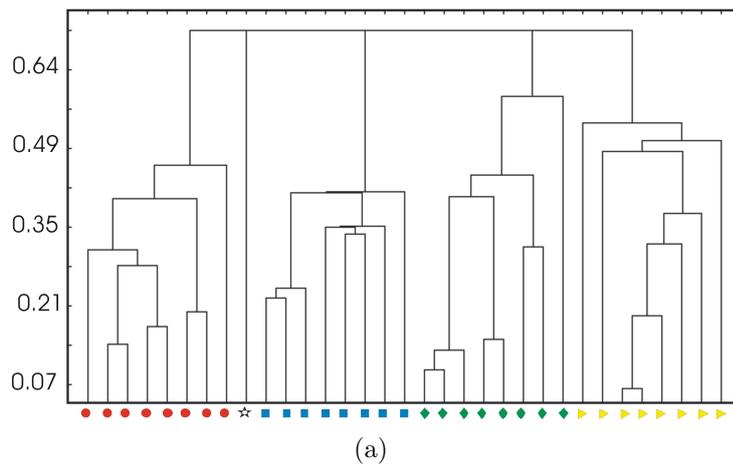


Figura 71: Resultados obtidos para o conjunto de dados 1. Figura 71(a): árvore hierárquica resultante da técnica de *single linkage*. Figura 71(b): árvore hierárquica resultante da técnica de caminhada do turista.

algoritmo da caminhada do turista não impõe tal estrutura arbitrária aos dados, privilegiando a estrutura natural dos dados. Ambas as árvores mostram a fusão dos quatro grupos distintos com o ponto mediano (rotulado com a estrela) no mesmo nível de hierarquia (em 0.7 unidades de distância para o *single linkage* e na janela de memória oito para a caminhada do turista). Ressaltamos que no método de *single linkage* (Figura 71(a)) esta fusão ocorre em cinco diferentes passos de iteração enquanto na caminhada do turista, isto acontece em um único passo. Uma característica interessante é que o valor de $\mu = 8$ atingido na fusão completa dos pontos da Figura 71(b) não ocorre por acaso. De fato, cada um dos quatro grupos colapsados neste passo possui exatamente oito elementos e, conforme enfatizado nas seções anteriores, os períodos dos atratores resultantes da caminhada são necessariamente maiores que μ . Assim, começando a caminhada em um dado ponto do conjunto de dados, a auto-repulsão impõe que o caminhante encontre um atrator com no mínimo nove pontos para $\mu = 8$. Como este procedimento é rodado para todos os N pontos simultaneamente, os atratores são sobrepostos, criando um único grupo de N pontos.

A Figura 72 ilustra a característica de “ação local” da caminhada do turista. O conjunto de dados 2 é construído por duas cópias de uma mesma estrutura em escalas diferentes. No dendrograma do *single linkage* (Figura 72(a)) pode-se ver dois agrupamentos distintos com densidades diferentes. O grupo menos denso da Figura 69 dá origem ao ramo da árvore com as maiores distâncias internodais. O mais denso ou, o menor grupo dá origem ao ramo mais denso da árvore. Ao contrário, a árvore da caminhada do turista é exatamente simétrica, representando os conjuntos equivalentes de maneira idêntica. Esta é uma característica importante do algoritmo proposto. Apesar de ter acesso a informações sobre todo o sistema (através da tabela de vizinhanças), a caminhada favorece informações locais no processo de fusão dos agrupamentos. Além disto, as distâncias entre os agrupamentos são desconsideradas na caminhada, que utiliza apenas as relações de ordenação entre os elementos. Estes resultados sugerem um possível avanço no uso de ferramentas de reconhecimento de padrões automáticas, na detecção de estruturas similares independentemente da escala em que aparecem.

Na Figura 70 podemos distinguir claramente dois grupos linearmente separáveis do conjunto de flores, o primeiro rotulado com os círculos e o segundo englobando os outros dois rótulos. Estes últimos não são linearmente separáveis e isto pode ser verificado nas Figuras 73(a) e 73(b) para o *single linkage* e a caminhada do turista respectivamente. Pode-se notar que ambos os métodos foram capazes de distinguir os dois grandes grupos referidos acima. Por sua vez, o grupo formado pelos rótulos quadrado e losango pode ser

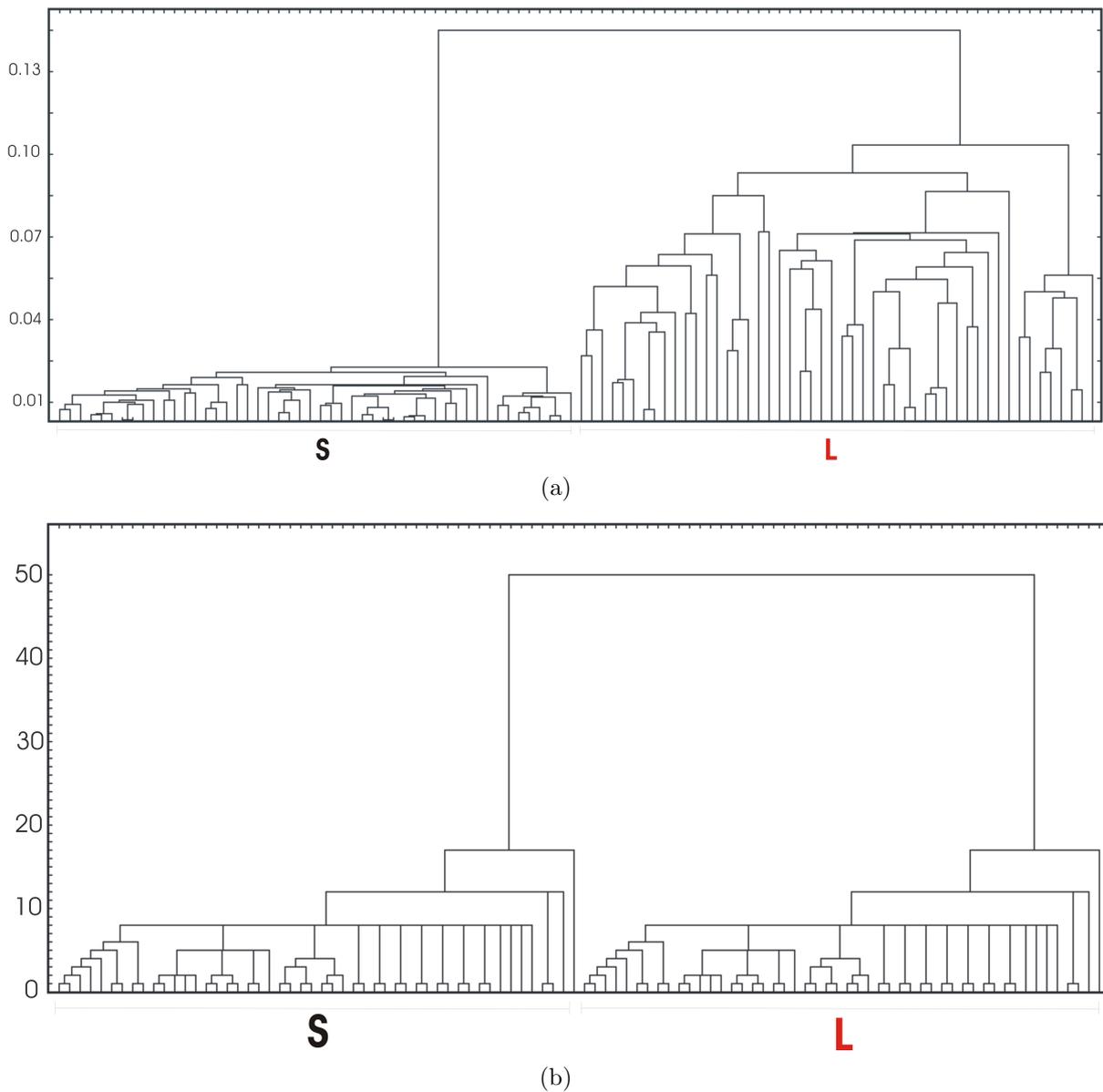


Figura 72: Resultados obtidos para o conjunto de dados 2. Figura 72(a): árvore hierárquica resultante da técnica de *single linkage*. Figura 72(b): árvore hierárquica resultante da técnica de caminhada do turista.

dividido em subgrupos de acordo com o corte escolhido e os resultados são marcadamente diferentes para ambos os métodos. O método de *single linkage*, ainda que com algumas classificações erradas e alguns pontos não agrupados, foi capaz de distinguir habilmente os subgrupos de quadrados e losangos. Isto não é observado na Figura 73(b) para a caminhada do turista que foi capaz de separar os rótulos, porém dentro de um único grande agrupamento. Um resultado similar pode ser encontrado em [49] com o uso de mapas de Kohonen. Os mapas de Kohonen e a caminhada do turista trabalham com uma vizinhança que não é essencialmente binária como no *single linkage*. De fato, se

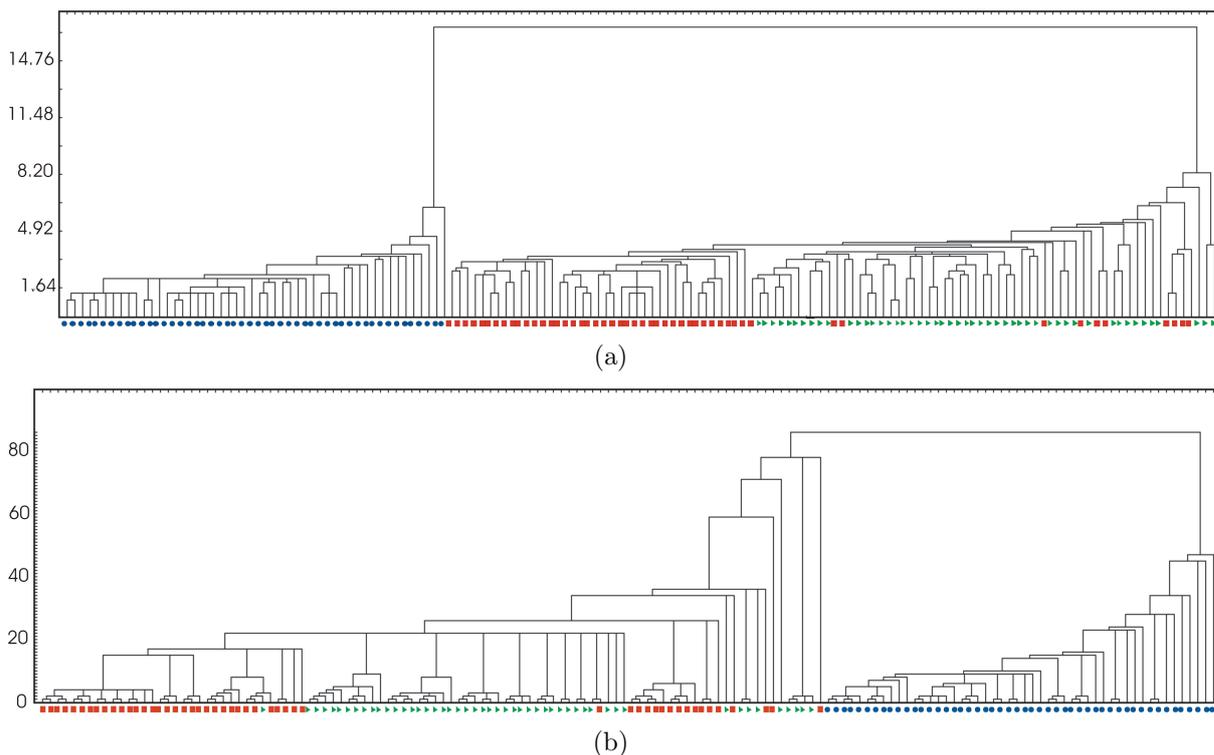


Figura 73: Resultados obtidos para o conjunto de dados 3. Figura 73(a): árvore hierárquica resultante da técnica de *single linkage*. Figura 73(b): árvore hierárquica resultante da técnica de caminhada do turista.

tentarmos reconhecer os agrupamentos visualmente na Figura 70 sem o conhecimento prévio dos rótulos e de posse apenas das distâncias Euclidianas entre os pontos, seríamos capazes de reconhecer apenas dois grupos. Este resultado é, portanto coerente com o método empregado.

6.4 Conclusões

O rápido desenvolvimento do poder computacional permitiu o acesso a imensas bases de dados e trouxe consigo a necessidade de métodos de análise elaborados e diversificados. Em muitas das aplicações emergentes, um único método de classificação não é suficiente para resolver o problema e, frequentemente múltiplos métodos devem ser aplicados para se atingir a resposta desejada [50].

Aqui nós apresentamos um algoritmo de reconhecimento de padrões baseado em uma caminhada determinista. O algoritmo permite uma representação hierárquica automática da estrutura implícita de um dado conjunto de dados de maneira eficiente. Algumas vantagens podem ser enfatizadas na utilização do algoritmo proposto:

1. A matriz de distâncias deixa de ser necessária e pode-se trabalhar apenas com a tabela de ordenamento de vizinhança para cada elemento. Isto poderia ser de grande utilidade em problemas onde as relações entre os objetos são qualitativas e a introdução de uma matriz de distâncias seria arbitrária. Note que não é possível adaptar o algoritmo de *single linkage* construindo-se uma matriz de distâncias a partir de uma tabela de vizinhanças, por exemplo, considerando-se o vizinho mais próximo com distância 1, o segundo mais próximo com distância 2 e assim por diante. Este método não seria implementável já que a tabela de vizinhanças é uma tabela de ordenamento e, portanto, não necessariamente simétrica. Dados dois pontos A e B desta tabela, se o ponto A é o n -ésimo vizinho de B , B não é necessariamente o n -ésimo vizinho de A . A “distância” baseada neste ranque não é simétrica, quebrando a definição formal de distância métrica imperativa ao algoritmo de *single linkage*.
2. A caminhada do turista favorece as informações locais de cada ponto dos dados e não faz restrições para a estrutura da árvore. Pontos são colapsados se eles pertencerem ao mesmo atrator, onde um atrator é um agrupamento de pontos mutuamente mais próximos. Por outro lado, dois atratores formam um nó na árvore se eles coalescerem em um dado nível de hierarquia.
3. Desconsiderar as distâncias entre os elementos do conjunto permite uma representação do conjunto de dados invariante por transformação de escala.

Além disto, a interpretação das relações entre os elementos de um conjunto apenas por sua distância de pares tem sido alvo de freqüentes críticas [51]. Como vimos no exemplo 3, classificações falsas, produzindo uma estrutura hierárquica que não existe no conjunto de dados são freqüentes neste tipo de abordagem. A utilização de um algoritmo que não imponha uma estrutura binária aos dados parece mais natural do ponto de vista da classificação não supervisionada.

Do ponto de vista da complexidade computacional, no entanto, o algoritmo da caminhada do turista recai em desvantagem com relação aos algoritmos hierárquicos tradicionais. Vimos no Capítulo 4 que as trajetórias crescem com o aumento de μ e que o valor de μ crítico, ou seja, o valor de memória necessário para que todo o sistema seja explorado parece crescer com o aumento da dimensionalidade do sistema. Assim, para problemas com muitas dimensões o algoritmo proposto não representa uma alternativa viável do ponto de vista computacional. No entanto, algumas modificações vêm sendo estudadas de modo a se manterem as características desejáveis do algoritmo e diminuir sua complexidade computacional.

7 Conclusão

Nosso trabalho visa apresentar a caminhada do turista como ferramenta de reconhecimento de padrões em conjunto de dados. Com este objetivo, exploramos as propriedades do modelo em diferentes meios e com diferentes parâmetros.

No Capítulo 4, apresentamos as características da caminhada do turista em meio contínuo em uma e duas dimensões. Uma questão pertinente em qualquer sistema dinâmico é sobre a existência de transição de fase e, no caso de haver transição, a determinação do parâmetro crítico. Em sistemas unidimensionais, estudamos dois sistemas que se diferenciam pela condição de contorno. Através das distribuições de tempos de transiente e das distribuições acumuladas complementares de ciclos, sugerimos os valores de memória crítica para a condição de contorno periódica como $\mu_c^{cpc} = \log_2 \frac{N}{4}$ e para condições abertas de contorno como $\mu_c^{cac} = \log_2 \frac{N}{2}$. Estes resultados devem ser comparados à memória crítica determinada analiticamente para um caminhante partindo do primeiro ponto do espaço e chegando ao último é $\mu_c^* = \log_2 N$. Esta transição corresponde a uma transição localização-delocalização no limite $N \rightarrow \infty$.

Em meios bidimensionais, pudemos apenas mostrar evidências de uma transição de fase em μ da ordem de N . A utilização da distribuição acumulada complementar permitiu-nos demonstrar que a distribuição de ciclos não segue lei de potência para μ finito. Este resultado é corroborado pelos gráficos de $\langle p \rangle$ em função de N que é constante dentro da faixa estudada.

No Capítulo 5 apresentamos uma abordagem inédita da caminhada do turista em redes regulares. Testamos três regras de movimentação, variando o parâmetro de memória e tamanho da rede. Os resultados dependem fortemente da regra de movimentação imposta. Nestes sistemas, assim como em sistemas contínuos, o tamanho médio da trajetória não depende de N e tanto $\langle p \rangle$ como $\langle t \rangle$ variam linearmente com μ .

No mesmo Capítulo, apresentamos as estatísticas das trajetórias para redes estruturadas em texturas. Mostramos que as estatísticas de tempos de transiente refletem de

maneira eficiente a estrutura embutida no conjunto de pixels. Motivados por esta propriedade, propusemos uma metodologia de análise de imagens baseada na caminhada do turista em que utilizamos os dados da matriz conjunta para caracterizar classes diferentes de texturas. Uma análise discriminante foi utilizada para classificar os vetores obtidos a partir das matrizes conjuntas obtendo uma eficiência de mais de 90% para 37 diferentes classes.

Finalmente, no Capítulo 6 apresentamos uma metodologia de reconhecimento de padrões em dados multidimensionais baseada na caminhada do turista. O algoritmo proposto constitui um método não supervisionado de agrupamento de dados que tem como resultado uma árvore hierárquica geral. Os métodos mais comumente utilizados baseiam-se na distância mínima entre pares de objetos dentro do conjunto de dados. Assim, o resultado obtido é uma árvore binária que impõe uma estrutura de pares ao conjunto analisado. Árvores gerais são difíceis de se obter e suas aplicações são raras. No entanto, parece-nos natural que uma metodologia de reconhecimento de agrupamentos naturais em conjuntos multidimensionais pudesse refletir a organização natural do conjunto, sem impor qualquer estrutura arbitrária. No algoritmo proposto, os agrupamentos são representados pelos atratores formados a cada passo de memória μ . Assim, as distâncias entre os elementos do conjunto são irrelevantes. Esta característica permite uma representação dos dados de forma independente de escala, ou seja, padrões semelhantes porém, em escalas diferentes são interpretados de forma idêntica pelo algoritmo proposto.

Mostramos que o método classifica eficientemente os conjuntos de dados testados e apresenta diferenças de atuação com os métodos tradicionais que podem ser vantajosas de acordo com a aplicação.

Uma das desvantagens do método é sua alta complexidade computacional. No Capítulo 4 demos a base para o cálculo desta complexidade e mostramos que ela ultrapassa a complexidade obtida com os métodos tradicionais no pior dos casos. Neste sentido, estudos estão sendo feitos para diminuir a redundância nas comparações e simplificar o método de determinação dos ciclos o que diminuiria consideravelmente a complexidade do método.

Nesta tese apresentamos resultados originais que contribuem para o avanço dos estudos sobre a caminhada do turista. Apresentamos também duas novas metodologias de análise de conjuntos de dados, uma com aplicação em imagens, outra com aplicação no agrupamento não supervisionado de dados multidimensionais e estudamos suas propriedades, vantagens e limitações. Ambas as metodologias foram publicadas em revistas internacionais indexadas de relevância na área em questão. Dentre as perspectivas de

trabalhos futuros, destacamos:

- o estudo da transição de fase em sistemas com $d = 2$ e em dimensionalidades mais altas;
- as diferentes regras de movimentação em redes regulares sugerem aplicações com diferentes objetivos. Um estudo das diferentes regras e suas implicações em imagens pode revelar novas e interessantes aplicações na área de diagnóstico médico por exemplo, com a detecção de bordas e de reconhecimento de texturas.
- o estudo da correlação entre a entropia de posição dos pixels e a distribuição de tempos de transiente correspondente para as caminhadas em texturas;
- a exploração de diferentes memórias na construção do vetor de características extraídos de texturas, concatenando informações acerca de diferentes escalas dentro da imagem, visando a melhoria da eficiência do método proposto de análise de imagens;
- melhorias na implementação do método de agrupamento baseado na caminhada do turista visando a diminuição de complexidade do método proposto.

Referências

- [1] J. Schöll and E. Schöll-Paschinger, “Classification by restricted random walks,” *Pattern recognition*, vol. 36, pp. 1279–1290, 2003.
- [2] S.-J. Yang, “Exploring complex networks by walking on them,” *Physical Review E*, vol. 71, p. 016107, 2005.
- [3] O. Kinouchi, A. S. Martinez, G. F. Lima, G. M. Lourenço, and S. Risau-Gusman, “Deterministic walks in random networks: an application to thesaurus graphs,” *Physica A*, vol. 315, pp. 665–676, December 2002.
- [4] F. Nascimento Silva and L. da Fontoura Costa, “Identifying complex networks by random walks.” physics/0612121v1, http://www.arxiv.org/PS_cache/physics/pdf/0612/0612121v1.pdf, 2006.
- [5] G. F. Lima, A. S. Martinez, and O. Kinouchi, “Deterministic walks in random media,” *Phys. Rev. Lett.*, vol. 87, p. 010603, Jun 2001.
- [6] H. E. Stanley and S. V. Buldyrev, “Statistical physics - the salesman and the tourist,” *Nature (London)*, vol. 413, pp. 373–374, 27 September 2001.
- [7] L. Lam, “Active walks: the first twelve years (part 1),” *International journal of bifurcation and chaos*, vol. 15, no. 8, pp. 2317–2348, 2005.
- [8] L. Lam, “Active walks: the first twelve years (part 2),” *International journal of bifurcation and chaos*, vol. 16, no. 2, pp. 239–268, 2006.
- [9] D. Boyer and H. Larralde, “Looking for the right thing at the right place: Phase transition in an agent model with heterogeneous spatial resources,” *Complexity*, vol. 10, pp. 52–55, January-February 2005.
- [10] J. P. Boon, “Random and deterministic walks.” physics/0501527v1, http://www.arxiv.org/PS_cache/cond-mat/pdf/0501/0501527v1.pdf, 2005.
- [11] L. A. Bunimovich and M. A. Khlabystova, “Walks in rigid environment,” *Journal of statistical physics*, vol. 108, no. 516, pp. 905–925, 2002.
- [12] D. Boyer, O. Miramontes, G. Ramos-Fernandez, J. L. Mateos, and G. Cocho, “Modeling the searching behavior of social monkeys,” *Physica A*, vol. 342, pp. 329–335, October 2004.
- [13] H. Freund and P. Grassberger, “The red queen’s walk,” *Physica A*, vol. 190, pp. 218–237, 1992.

- [14] U. Dieckmann, P. Marrow, and R. Law, “Evolutionary cycling in predator-prey interactions: population dynamics and the red queen,” *Journal of theoretical biology*, vol. 176, pp. 91–102, 1995.
- [15] C. G. Langton, “Studying artificial life with cellular automata,” *Physica D*, vol. 2, no. 1-3, pp. 120–149, 1986.
- [16] A. Moreira, A. Gajardo, and E. Goles, “Dynamical behavior and complexity of langton’s ant,” *Complexity*, vol. 6, no. 4, pp. 46–51, 2001.
- [17] A. Bhattacharjya and S. Liang, “Power-law distributions in some random boolean networks,” *Phys. Rev. Lett.*, vol. 77, pp. 1644–1647, August 1996.
- [18] S. Kauffman, “Metabolic stability and epigenesis in randomly constructed genetic nets,” *Journal of theoretical biology*, vol. 22, pp. 437–467, 1969.
- [19] B. Derrida and H. Flyvbjerg, “The random map model - a disordered model with deterministic dynamics,” *J. Phys. (Paris)*, vol. 48, pp. 971–978, June 1987.
- [20] B. Harris, “Probability-distributions related to random mappings,” *Ann. Math. Stat.*, vol. 31, no. 4, pp. 1045–1062, 1960.
- [21] A. G. Percus and O. C. Martin, “The stochastic traveling salesman problem: finite size scaling and the cavity prediction,” *J. Stat. Phys.*, vol. 94, pp. 739–758, March 1999.
- [22] D. Boyer, G. Ramos-Fernandez, O. M. J. L. Mateos, G. Cocho, H. Larralde, H. Ramos, and F. Rojas, “Scale-free foraging by primates emerges from their interaction with a complex environment,” *Proceedings of the Royal Society B-Biological Sciences*, vol. 273, pp. 1743–1750, JUL 22 2006.
- [23] A. S. Martinez, Osame, and S. Risau-Gusman, “Exploratory behavior, trap models, and glass transitions,” *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 69, no. 1, p. 017101, 2004.
- [24] H. Schildt, *Borland C++ completo e total versão 5*. São Paulo, 1997.
- [25] C. A. S. Terçariol and A. S. Martinez, “Analytical results for the statistical distribution related to a memoryless deterministic walk: dimensionality effect and mean-field models,” *Physical review E*, vol. 72, p. 021103, 2005.
- [26] G. F. de Lima, *Caminhadas determinísticas em meios desordenados: problema da caminhada do turista*. PhD thesis, Universidade de São Paulo, Ribeirão Preto - SP, Junho 2002. (in Portuguese).
- [27] C. A. S. Terçariol, R. S. González, and A. S. Martinez, “Exact analytical calculation for the percolation crossover in deterministic partially self-avoiding walks in one-dimensional random media.” physics/0702030v1, http://www.arxiv.org/PS_cache/cond-mat/pdf/0702/0702030v1.pdf, 2007.
- [28] M. Newman, “Power laws, pareto distributions and zipf’s law.” cond-mat/0412004v3, http://www.arxiv.org/PS_cache/cond-mat/pdf/0412/0412004v3.pdf, 2007.
- [29] W. T. R. Oliveira comunicação privada, 2007.

- [30] S. Coppersmith and Z. Z. Leo P. Kadanoff, “Reversible boolean networks ii: Phase transitions, oscillations and local structures.” cond-mat/0009019v2, http://www.arxiv.org/PS_cache/cond-mat/pdf/0009/0009019v2.pdf.
- [31] S. Coppersmith and Z. Z. Leo P. Kadanoff, “Reversible boolean networks i: Distribution of cycle lengths.” cond-mat/0004422v1, http://www.arxiv.org/PS_cache/cond-mat/pdf/0004/0004422v1.pdf.
- [32] C. A. S. Terçariol, R. S. González, and A. S. Martinez, “Exploring random media with partially self-avoiding deterministic walks,” in *31st Conference of the Middle European Cooperation in Statistical Physics (MECO 31)*, vol. 31, (Primošten - Croácia), p. 84, Abril 2006.
- [33] O. Kinouchi and C. P. C. Prado, “Robustness of scale invariance in models with self-organized criticality,” *Physical Review E*, vol. 59, pp. 4964–4969, 1999.
- [34] P. Brodatz, *A Photographic Album for Arts and Design*, vol. 1. Toronto: Dover Publishing Co., 1966.
- [35] M. G. Campitelli, P. D. Batista, O. Kinouchi, and A. S. Martinez, “Deterministic walks as an algorithm of pattern recognition,” *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 74, no. 2, p. 026703, 2006.
- [36] S. Li, “On discontinuity-adaptative smoothness priors in computer vision,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 6, pp. 576–586, 1995.
- [37] P. Howarth and S. Ruger, “Evaluation of texture features for content-based image retrieval,” *Lect. Note Comput. Sci.*, vol. 3115, pp. 326–334, 2004.
- [38] A. K. Jain and F. Farrokhnia, “Unsupervised texture segmentation using gabor filters,” *Pattern Recogn.*, vol. 24, no. 12, pp. 1167–1186, 1991.
- [39] A. Webb, *Statistical Pattern Recognition*, vol. 1. Newnes, 1999.
- [40] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005.
- [41] R. O. Duda and P. E. Hart, *Pattern Classification*. John Wiley and Sons, 2000.
- [42] J. Vesanto and E. Alhoieni, “Clustering of the self-organizing map,” *IEEE Transactions on neural networks*, vol. 11, no. 3, pp. 586 – 600, 2000.
- [43] M. G. Campitelli, A. S. Martinez, and O. M. Bruno, “An image analysis methodology based on deterministic tourist walks,” *Lect. Note Comput. Sci.*, vol. 4140, pp. 159–167, 2006.
- [44] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási, “Hierarchical organization of modularity in metabolic networks,” *Science*, vol. 297, pp. 1551 – 1555, 2002.
- [45] M. S. Su and Y. C. Liu, “A new approach to clustering data with arbitrary shapes,” *Pattern Recogn.*, vol. 38, pp. 1887 – 1901, 2005.

-
- [46] R. Xu, “Survey of clustering algorithms,” *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645 – 678, 2005.
- [47] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” *ACM Computing Surveys*, vol. 31, no. 3, pp. 264 – 323, 1999.
- [48] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, pp. 178–188, 1936.
- [49] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, “Self-organizing map in matlab: the som toolbox,” in *Proceedings of the Matlab DSP Conference*, pp. 35–40, 1999.
- [50] A. K. Jain, R. P. W. Duin, and J. Mao, “Statistical pattern recognition: a review.,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 4–37, 2000.
- [51] N. A. Alves, “Unveiling community structures in weighted networks.” physics/0703087, http://www.arxiv.org/PS_cache/physics/pdf/0703/0703087v1.pdf.