UNIVERSIDADE FEDERAL DE ALAGOAS
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE – FEAC
CURSO DE MESTRADO EM ECONOMIA APLICADA – CMEA

ADHEMAR RANCIARO NETO

TEORIA DA INFORMAÇÃO ALGORÍTMICA, EFICIÊNCIA RELATIVA DE MERCADO E PERDA DE MEMÓRIA EM SÉRIES DE RETORNOS DE ALTA FREQUÊNCIA EM ATIVOS NEGOCIADOS NA BM&F BOVESPA

Maceió 2010

ADHEMAR RANCIARO NETO

TEORIA DA INFORMAÇÃO ALGORÍTMICA, EFICIÊNCIA RELATIVA DE MERCADO E PERDA DE MEMÓRIA EM SÉRIES DE RETORNOS DE ALTA FREQUÊNCIA EM ATIVOS NEGOCIADOS NA BM&F BOVESPA

Dissertação apresentada como requisito parcial à obtenção do título de Mestre em Ciências Econômicas do Programa de Pós-Graduação em Economia Aplicada da Universidade Federal de Alagoas (UFAL)

Orientador: Prof. Dr. Iram Marcelo Gléria

Maceió 2010

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária Responsável: Helena Cristina Pimentel do Vale

R185t Ranciaro Neto, Adhemar.
Teoria da informação algorítmica, eficiência relativa de mercado e perda de memória em séries de retornos de alta frequência em ativos negociados na BM&F BOVESPA / Adhemar Ranciaro Neto, 2010.
78 f. grafs., tabs.

Orientador: Iram Marcelo Gléria.
Dissertação (mestrado em Economia Aplicada) – Universidade Federal de Alagoas. Faculdade de Economia, Administração e Contabilidade, Maceió, 2010.

Bibliografia: f. 87-95.
Anexos: f. 96-107.

1. Econofisica. 2. Eficiência de mercado. 3. Informação algorítmica. 4. Complexidade algorítmica. 5. Bolsa de valores – Brasil. I. Título.

CDU: 336.018(81)

Universidade Federal de Alagoas
Faculdade de Economia, Administração e Contabilidade
Programa de Pós-Graduação em Economia Aplicada

Teoria da informação algorítmica, eficiência relativa de mercado e perda de memória em séries de retornos de alta freqüência em ativos negociados na BM&F Bovespa

ADHEMAR RANCIARO NETO

Dissertação submetida ao corpo docente do Programa de Pós-Graduação em Economia Aplicada da Universidade Federal de Alagoas e aprovada em 05 de julho de 2010.

Banca Examinadora:

Prof. Dr. Iram Marcelo Gléria (CMEA-UFAL)
(Orientador)

Prof. Dr. Francisco José Peixoto Rosário (CMEA-UFAL)
(Examinador Interno)

Prof. Dr. Eraldo Sergio Barbosa da Silva (UFSC)
(Examinador Externo)

Dedico este trabalho à minha
esposa com amor e carinho

AGRADECIMENTOS

À Faculdade de Economia, Administração e Contabilidade pela estrutura física e administrativa que contribuiu para a realização do projeto, em especial, à figura do Prof. Dr. Francisco José Peixoto Rosário;

À FAPEAL pelo apoio financeiro por 12 meses;

Ao professor Iram Marcelo Gléria pela orientação do trabalho e, também, pelo incentivo ao desenvolvimento de minha carreira;

Aos professores doutores Eraldo Sérgio Barbosa da Silva e Francisco José Peixoto Rosário pela composição das bancas de qualificação e de defesa, que contribuíram muito para este trabalho;

À família Bandini pelo apoio emocional;

À BM&F BOVESPA pelo fornecimento da base de dados que possibilitou a execução do trabalho;

“Toda a nossa ciência, comparada com a realidade, é primitiva e infantil, E, no entanto, é a coisa mais preciosa que temos."

(Albert Einstein)

RESUMO

O presente trabalho tem por objetivos: 1) aplicar a teoria da complexidade de Kolmogorov utilizando a medida proposta por Lempel e Ziv (1976) para analisar o comportamento desta diante de alterações em parâmetros como tamanho de janela, salto e de região de estabilidade em séries financeiras de retornos de alta freqüência de ativos negociados na BM&F BOVESPA; 2) avaliar a evolução da medida ao se ampliarem os intervalos entre as negociações; e finalmente, 3) verificar a possibilidade de existir algum indício de relação entre o valor daquela medida e o comportamento das curvas de autocorrelação apresentadas para cada intervalo de negociação especificado. Foi também discutido o critério utilizado para a medida de eficiência relativa de mercado proposto por Giglio (2008).

Palavras-chave: Eficiência de Mercado, Informação Algorítmica, Complexidade Algorítmica

ABSTRACT

This paper aims to apply the Kolmogorov algorithmic complexity theory using the measure proposed by Lempel and Ziv (1976) to analyze its behavior due to changes in parameters such as window size, jumps and the region of stability of high frequency financial series returns of assets traded on the BM&F BOVESPA, as well as to assess the evolution of such a measure when the intervals between the negotiations are extended and to verify the possible evidence of a relationship between the value of the complexity measure and the behavior of autocorrelation curves presented for each trading interval specified. We also discuss the criterion used to measure the relative efficiency of the market proposed by Giglio (2008).

Keywords: Market Efficiency, Algorithmic Information, Algorithmic Complexity.

LISTA DE FIGURAS

Figura 1: Esquema de uma máquina de Turing. 14
Figura 2: Representação de possíveis operações de uma máquina de Turing. 15
Figura 3: Evolução da entropia da fonte *versus* probabilidade para dois eventos. 26
Figura 4: Diagrama para o algoritmo do cálculo da complexidade de uma sequência de comprimento n 29
Figura 5: LZ médio *versus* RE BNCA3 JANELA 5000. 37
Figura 6:LZ médio *versus* RE BNCA 3 JANELA 10000. 37
Figura 7: LZ médio *versus* RE BNCA3 JANELA 30000. 38
Figura 8: LZ médio *versus* RE BRTP3 JANELA 5000. 38
Figura 9: LZ médio *versus* RE BRTP3 JANELA 10000. 39
Figura 10: LZ médio *versus* RE BRTP3 JANELA 30000. 39
Figura 11:LZ médio *versus* RE CGAS5 JANELA 5000. 40
Figura 12:LZ médio *versus* RE CGAS5 JANELA 10000. 40
Figura 13: LZ médio *versus* RE CGAS5 JANELA 30000. 41
Figura 14: LZ médio *versus* RE CLSC6 JANELA5000. 41
Figura 15: LZ médio *versus* RE CLSC6 JANELA 10000. 42
Figura 16: LZ médio *versus* RE CLSC6 JANELA 30000. 42
Figura 17: LZ médio *versus* RE LIGT3 JANELA 5000. 43
Figura 18: LZ médio *versus* RE LIGT3 JANELA 10000. 43
Figura 19: LZ médio *versus* RE LIGT3 JANELA 30000. 44
Figura 20: LZ médio *versus* RE SBSP3 JANELA 5000. 44
Figura 21: LZ médio *versus* RE SBSP3 JANELA 10000. 45
Figura 22: LZ médio *versus* RE SBSP3 JANELA 30000. 45
Figura 23: LZ médio *versus* RE SDIA4 JANELA 5000. 46
Figura 24: LZ médio *versus* RE SDIA4 JANELA 10000. 46
Figura 25: LZ médio *versus* RE SDIA4 JANELA 30000. 47
Figura 26: LZ médio *versus* RE TCSL3 JANELA 5000. 47
Figura 27: LZ médio *versus* RE TCSL3 JANELA 10000. 48
Figura 28: LZ médio *versus* RE TCSL3 JANELA 30000. 48
Figura 29: LZ médio *versus* RE TLPP4 JANELA 5000. 49
Figura 30: LZ médio *versus* RE TLPP4 JANELA 10000. 49
Figura 31: LZ médio *versus* RE TLPP4 JANELA 30000. 50
Figura 32: LZ médio *versus* RE TMAR5 JANELA 5000. 50
Figura 33: LZ médio *versus* RE TMAR5 JANELA 10000. 51
Figura 34: LZ médio *versus* RE TMAR5 JANELA 30000. 51
Figura 35: LZ médio *versus* RE TNLP3 JANELA 5000. 52
Figura 36: LZ médio *versus* RE TNLP3 JANELA 10000. 52
Figura 37: LZ médio *versus* RE TNLP3 JANELA 30000. 53
Figura 38: LZ médio *versus* RE TRPL4 JANELA 5000. 53
Figura 39: LZ médio *versus* RE TRPL4 JANELA 10000. 54
Figura 40: LZ médio *versus* RE TRPL4 JANELA 30000. 54
Figura 41: LZ médio *versus* RE UGPA4 JANELA 5000. 55

Figura 42: LZ médio *versus* RE UGPA4 JANELA 10000. 55
Figura 43: LZ médio *versus* RE UGPA4 JANELA 30000. 56
Figura 44: LZ médio *versus* RE USIM3 JANELA 5000. 56
Figura 45: LZ médio *versus* RE USIM3 JANELA 10000. 57
Figura 46: LZ médio *versus* RE USIM3 JANELA 30000. 57
Figura 47: Variável estocástica *versus* tempo. (a) Região de estabilidade 0,19. (b) Região de estabilidade 0,22. 59
Figura 48: Eficiência Relativa *versus* RE BNCA3 JANELA 5000. 61
Figura 49: Eficiência Relativa *versus* RE BNCA3 JANELA 10000. 61
Figura 50: Eficiência Relativa *versus* RE LIGT3 JANELA 5000. 62
Figura 51: Eficiência Relativa *versus* RE SBSP3 JANELA 5000. 62
Figura 52: Eficiência Relativa *versus* RE SDIA4 JANELA 5000. 63
Figura 53: Eficiência Relativa *versus* RE SDIA4 JANELA 10000. 63
Figura 54: LZ médio *versus* intervalo entre negócios para a ação CGAS5. 64
Figura 55: Eficiência relativa *versus* intervalo entre negócios para a ação CGAS5. 65
Figura 56: Desvio padrão do LZ médio *versus* intervalo entre negócios para a ação CGAS5. 65
Figura 57: Número de valores absolutos de autocorrelação superiores a 0,05 *versus* intervalo entre negócios para a ação CGAS5. 67
Figura 58: LZ médio *versus* eficiência relativa de mercado (GIGLIO 2008). 68
Figura 59: LZ médio *versus* intervalo entre negócios para a ação ELPL6. 68
Figura 60: Eficiência relativa *versus* intervalo entre negócios para a ação ELPL6. 69
Figura 61: Desvio padrão do LZ médio *versus* intervalo entre negócios para a ação ELPL6. 69
Figura 62: Número de valores absolutos de autocorrelação superiores a 0,05 *versus* intervalo entre negócios para a ação ELPL6. 71
Figura 63: LZ médio *versus* eficiência relativa de mercado (GIGLIO 2008). 71

LISTA DE TABELAS

Tabela 1: Posições e tamanho dos dados na linha de registro 32
Tabela 2: Ações negociadas na BOVESPA utilizadas no estudo da evolução do LZ. 35
Tabela 3: Janelas, Saltos e regiões de estabilidades para análise de dados negócio a negócio. 36
Tabela 4: Maiores diferenças de LZ médio observadas entre saltos de tamanhos distintos 59

LISTA DE ABREVIATURAS E DE SIGLAS

BRASIL T PAR – Brasil Telecom Participacões S/A.

CELESC - Centrais Elétricas de Santa Catarina S/A.

COMGÁS – Companhia de Gás de São Paulo S/A.

ELETROPAULO – Eletropaulo Metropolitana Eletricidade De São Paulo S/A.

LIGHT – LIGHT S/A.

NOSSA CAIXA – Banco Nossa Caixa S/A.

ON – Ação ordinária

PN – Ação preferencial

PNA - Ação preferencial classe A

PNB - Ação preferencial classe B

SABESP - Companhia De Saneamento Básico Do Estado De São Paulo S/A.

SADIA - SADIA S/A.

TELEMAR e TELEMAR N L - Telemar Norte Leste S/A.

TELESP – Telecomunicações de São Paulo S/A.

TIM PART – TIM Participações S/A.

CTEEP – Companhia de Transmissão de Energia Elétrica Paulista S/A.

ULTRAPAR - Ultrapar Participações S/A.

USIMINAS - Usinas Siderúrgicas de Minas Gerais S/A.

SUMÁRIO

1 **INTRODUÇÃO** 1

1.1 **Sistemas Complexos e Econofísica** 1

1.2 **Eficiência de Mercado** 3

1.3 **Autocorrelação e Tempo em Séries Financeiras** 6

2 **MÉTODO** 8

2.1 **Apresentação** 8

2.2 **Aleatoriedade, complexidade e informação algorítmica** 8

2.2.1 Sequências Aleatórias 9

2.2.2 Complexidade de Kolmogorov, Teoria da Informação Algorítmica e compressividade. ... 13

2.2.3 Alguns procedimentos para a verificação da aleatoriedade de uma sequência de caracteres. 18

2.3 **O modelo de Lempel e Ziv e o posterior desenvolvimento computacional de Kaspar e Schuster**. 21

2.3.1 O modelo de Lempel e Ziv 21

2.3.2 Desenvolvimento computacional proposto por Kaspar e Schuster (1987). 28

2.4 Aplicação do método aos dados 29

3 **RESULTADOS E DISCUSSÃO** 35

3.1 **Considerações gerais** 35

3.2 **Análise dos resultados no caso de intervalos de tempo igual a uma negociação** 36

3.2.1 Evolução do LZ médio em relação à região de estabilidade 36

3.2.2 Evolução da Eficiência relativa (GIGLIO 2008) em relação à região de estabilidade. 60

3.3 **Análise dos resultados no caso de intervalos de tempo maior que uma negociação**. 64

4 **CONCLUSÃO** 72

REFERÊNCIAS 74

1 **INTRODUÇÃO**

1.1 **Sistemas Complexos e Econofísica**

Ao final do século XX, parte dos físicos dedicou-se a estudar a dinâmica dos sistemas complexos. Um sistema complexo é de acordo com Heylighen (1988) definido quando existem dois ou mais componentes interagindo de modo a formar uma estrutura estável. Tais interações são não lineares (GLÉRIA, MATSUSHITA e DA SILVA 2004). A partir daquela premissa, não se pode estudar tal sistema a partir de seus componentes separadamente (reducionismo), pois as interações seriam descartadas e tampouco avaliá-lo como um todo, baseando-se quase que somente nas interações entre os componentes (holismo). A análise destes sistemas deve conjugar as duas correntes de pensamento não se podendo negar a existência naquele dos elementos distinção (separação do todo) e conexão (indissociabilidade do todo sem perder parte do significado original). Exemplos de modelos que satisfazem tal análise são os baseados em rede. Uma rede é composta por nós (representando os componentes) e as conexões entre os nós representando as interações.

Em um sistema complexo há uma relação circular entre a estrutura global do sistema e as interações locais entre os componentes. Cada componente interage com a estrutura global e esta, por sua vez, é formada pelas interações entre os componentes e seus vizinhos. No estudo da evolução dos sistemas complexos, propriedades interessantes são observadas sob certas circunstâncias tais como: leis de escala, auto similaridade, fractais e criticalidade auto organizada.

De acordo com Gléria et. al. (2004), um fenômeno crítico ocorre geralmente em um sistema que está longe do equilíbrio, em processos nos quais a história é importante. Quando o sistema entra em um estado crítico, pequenas perturbações podem levá-lo a eventos sem causa determinada e de qualquer magnitude. Conforme observado por Bak e Paczluski (1995) tais eventos não podem ser previstos, mas sua distribuição estatística pode ser determinada.

Bak e Paczluski (1995) apontaram que sistemas dinâmicos de tamanho grande e que apresentam muitos graus de liberdade naturalmente se organizam para atingir o estado crítico (criticalidade auto organizada). O sistema passa a apresentar equilíbrios

pontuais onde períodos de regularidade são interrompidos por tormentas intermitentes. Então, sistemas que apresentam eventos comuns podem ser levados a um estado crítico produzindo eventos catastróficos, no sentido de inesperados, incomuns e que podem assumir qualquer magnitude.

Empiricamente, Mandelbrot percebeu que, para muitos sistemas, a distribuição de probabilidade de grandes eventos é dada pela mesma função de distribuição de eventos pequenos, levando a um forte indício de uma origem dinâmica comum entre aqueles sistemas. Esta função pertence à família de distribuições de Pareto-Lévy. Estas apresentam propriedades de ausência de escala característica e caudas que se distribuem em leis de potência. Para descrever estes tipos de comportamento em um sistema, Mandelbrot utilizou o termo "fractal". (BAK e PACZLUSKI 1995 p. 2)

As distribuições de Pareto-Lévy (exceto a Gaussiana) possuem caudas longas. Esta peculiaridade faz com que quase todos os momentos da distribuição sejam infinitos, exceto o primeiro momento.

Sistemas que apresentam leis de potência são acompanhados do ruído (espectro de potência) $\frac{1}{f}$[1]. Bak, Tang e Wiesenfeld (1987) após analisarem diversos tipos de sistemas dinâmicos, perceberam que os fractais e os ruídos $\frac{1}{f}$ podem surgir na natureza sem perturbação externa. Os autores chegaram até a propor, a partir de estudos numéricos, que qualquer sistema dinâmico que apresente as características apropriadas pode se auto organizar para a criticalidade.

Em Economia, uma série de estudos empíricos foi realizada e foram percebidas as características de um sistema complexo dinâmico que se dirige para um estado crítico. O sistema econômico é composto de unidades que interagem entre si e estas com o todo. As crises poderiam ser explicadas pela chegada do sistema à criticidade, sendo então, vistas como intrínsecas à natureza do sistema.

Algumas das pesquisas mais famosas são as de Pareto, que verificou que a renda em uma localidade era distribuída sob a forma de lei de potência; de Mandelbrot (1963) que, ao analisar os preços dos contratos de algodão, percebeu que sua distribuição não era Gaussiana (obedecia a uma lei de potência) e ainda observou que,

[1] Em Eletrônica este tipo de espectro de potência é chamado de ruído rosa. Ele é formado a partir do inverso da freqüência de um sinal.

ao ampliar a escala de tempo, era vislumbrado o mesmo comportamento dos preços que em relação à escala reduzida (ausência de escala), levando o autor a desenvolver a teoria dos fractais (entes de dimensão fracionada); e em Mantegna e Stanley (2000), que observaram leis de escala no comportamento do índice *Standard & Poors* 500 (S&P 500) e que retomaram os estudos de Mandelbrot no sentido de encontrar alguma forma de determinar o segundo momento finito da distribuição de preços em séries financeiras, dado que o significado econômico deste é importante[2].

Devido aos grandes avanços no estudo da dinâmica dos sistemas complexos e nas diversas verificações de suas propriedades na Economia, Mantegna e Stanley (2000) propuseram o termo Econofísica para esta junção dos estudos econômicos com a teoria da dinâmica dos sistemas complexos.

1.2 **Eficiência de Mercado**

Em Economia, um mercado é dito eficiente quando os preços de mercado refletem completamente toda a informação disponível. Tal hipótese foi proposta por Fama (1970) na tentativa de compreender as bases para a verificação da utilidade das análises técnica e fundamentalista[3] utilizadas em finanças.

Uma definição semelhante à de Fama (1970) é a encontrada em Malkiel (2003), Mantegna e Stanley (2000) e em Schimidt (2005), que diz que um mercado eficiente é aquele cuja informação nova é incorporada instantaneamente ao preço do ativo. Com isso, tanto a análise técnica quanto a fundamentalista seriam inúteis na tentativa de prever os preços dos ativos.

Em trabalho independente ao de Fama, Samuelson (1965) mostrou que, se os preços incorporam totalmente e instantaneamente as informações, então eles deverão variar de forma aleatória. O preço do instante x incorpora a informação do instante x. O preço no instante $x + t$ dependerá exclusivamente da informação nova, que é fornecida e incorporada em $x + t$. Tal fenômeno apresentado por Samuelson

[2] A raiz quadrada do segundo momento (variância) é a medida de volatilidade das variáveis em um processo estocástico. Para os Economistas, o desvio padrão é uma medida de risco.

[3] De acordo com Lo (2007, p.2), análise técnica refere-se ao uso de padrões geométricos de preços e de planilhas de volume para prever os futuros movimentos de preço de um ativo e a análise fundamentalista refere-se ao uso de dados econômicos e contábeis para determinar o valor correto de um ativo.

(1965) nos preços dos mercados que atendem a condição do parágrafo acima, é chamado, na literatura especializada, de passeio aleatório.

A partir das idéias propostas pelos dois autores, a hipótese dos mercados eficientes (HME) é associada à idéia de passeio aleatório.

Em um mercado eficiente não é possível obter lucro econômico persistente[4] por meio de qualquer negócio baseado em informação[5]. Para Malkiel (2003), ser um negociador experiente em um mercado eficiente equivale a ser um chipanzé vendado atirando dardos sobre os ativos como forma de escolhê-los para realizar uma operação.

Em relação à questão do lucro nos mercados eficientes, Jensen (1978) definiu que os preços refletem apenas as informações cujos custos de obtenção não excedam os benefícios de sua utilização.

Com relação à informação, Fama (1991) apresentou condições para que um mercado pudesse ser enquadrado como eficiente:

a) Não existem custos de transação na negociação de um ativo
b) Todas as informações estão disponíveis a todos os agentes do mercado
c) Todos os agentes têm a mesma opinião sobre os impactos dessa nova informação sobre os preços futuros e atuais do ativo negociado (expectativas homogêneas)

Em Jensen (1978) encontram-se as formas da hipótese de eficiência de mercado propostas em Fama (1970). Este autor fez tal separação para poder verificar que tipo de teste empírico pode ser utilizado dependendo da categoria do conjunto de informação disponível para a análise. São elas:

a) Forma fraca da HME: é a forma na qual o conjunto de informação possui somente a informação contida na história de preços até o instante t. Nesta classificação da hipótese, os testes consistem em avaliar a somente a previsibilidade do preço de um ativo com base neste conjunto de informações.

[4] Nos mercados eficientes observam-se as características de um jogo justo (onde o resultado esperado da próxima rodada é igual a zero, não importando o conteúdo histórico até a referida rodada) podendo, portanto, serem modelados sob a forma de *martingales*.

[5] Para Jensen (1978) um mercado é eficiente com relação ao conjunto de informação X se for impossível obter lucros econômicos negociando com base nas informações de X.

b) Forma semi-forte da HME: é a forma na qual o conjunto de informação possui toda a informação pública disponível no instante t (inclusive a informação histórica). Nesta versão da hipótese é possível realizar testes para verificar o ajustamento de preços devido à chegada de uma nova informação para todos.
c) Forma forte da HME: é a forma na qual o conjunto de informação contém toda a informação disponível no instante t seja ela pública ou privada, de conhecimento geral ou de apenas um participante. Nesta categoria de hipótese, o teste necessitaria ser alimentado com informações privadas, o que constitui em uma limitação devido à restrição em sua disponibilidade.

Diversos testes empíricos foram realizados mostrando que a HME é consistente na maioria dos casos. (JENSEN 1978)

Um teste famoso de rejeição da forma fraca da HME foi o proposto por Lo e MacKinlay (1988).

Ao avaliar a eficiência de mercado dos índices de retornos ações dos Estados Unidos por meio da hipótese de passeio aleatório, Lo e MacKinlay (1988) perceberam que retornos semanais dos índices eram correlacionados devido ao aumento de variância entre retornos analisados com defasagem alta ser superior ao aumento determinado pelo modelo de passeio aleatório, desta forma verificando que alguns resultados futuros conseguem ser previstos pelo conjunto de informações anteriores. O teste de hipótese realizado consistia em analisar quocientes entre variâncias.

Todavia, testar uma rejeição da HME para um determinado mercado de forma absoluta é complicado, pois, de acordo com Lo (2007) a HME não é bem definida[6] e empiricamente refutável. Por isso os testes têm que ter hipóteses auxiliares conjuntas, assim, não sendo possível saber qual elemento é inconsistente com os dados, caso o teste falhe.

Devido à dificuldade de se testar a HME em sua versão original, Lo (2007) propôs que a eficiência de um mercado seja testada de forma relativa a outro mercado em analogia aos sistemas físicos que necessitam de um referencial para que suas

[6] De acordo com Lo (2007, p.6), a HME faz uma afirmação sobre dois aspectos distintos dos preços: o conteúdo informacional e o mecanismo de formação de preços. Logo, qualquer teste de eficiência deve levar em conta o tipo de informação que é refletida nos preços e como essa informação vem a ser refletida nos preços.

grandezas sejam medidas. Tal proposta é corroborada em Grossman e Stiglitz (1980) que teorizam sobre a impossibilidade de se chegar à HME, mas que pode ser utilizada para uma análise comparada (relativa). Estes autores argumentam que os custos das transações são responsáveis pelas ineficiências observadas nos mercados reais. Ainda afirmam que se não houvesse lucro em adquirir informação, então haveria pouca razão para negociar provocando o colapso dos mercados. Com certo grau de ineficiência de mercado, os investidores sentem-se motivados a negociar com base em informação e, portanto, um equilíbrio de mercado estável surge quando há oportunidades de lucro.

Mantegna e Stanley (2000) sugerem um modelo para verificar a versão fraca da HME em caráter relativo. A teoria da complexidade algorítmica proposta por Kolmogorov (1965) e Chaitin (1966) serve para verificar se uma sequência de dados pode apresentar ou não características aleatórias (o que verificaria a imprevisibilidade em uma série de preços de um ativo). As definições de aleatoriedade, bem como os mecanismos de sua verificação são mostrados no capítulo 2 deste trabalho.

1.3 **Autocorrelação e Tempo em Séries Financeiras**

De acordo com Mantegna e Stanley (2000) a análise de autocorrelação é importante para a verificação da independência entre as variações de preços nas séries financeiras. Em dados de alta freqüência, geralmente as autocorrelações entre os valores caem para um valor insignificante após uma defasagem entre preços em intervalo de tempo menor que 1 dia. Este tempo somente pode ser mensurado de forma adequada para tais dados.

Processos cuja autocorrelação possui memória curta possuem espectro de freqüência do tipo $s(f) = \frac{1}{f^2}$, o que evidencia independência entre as variáveis.

Para processos cuja autocorrelação possui memória longa, podem ser acompanhados de ruídos do tipo $s(f) = \frac{1}{f}$, evidenciando um processo com ausência de escala.

Com relação às escalas de tempo a serem escolhidas, Mantegna e Stanley (2000) definem que pode haver 3 candidatos;

a) Tempo físico
b) Tempo de mercado
c) Número de transações

Todas as três opções possuem argumentos favoráveis e contrários. Por exemplo: o tempo físico apresenta a regularidade, mas não consegue captar os efeitos do fechamento do mercado, fins de semana e feriados. O tempo de mercado consegue evitar o problema dos efeitos do fechamento do mercado, como, por exemplo, o efeito fim de semana (variância dos preços entre os finais de semana é maior que entre os preços de fechamento e de abertura do mercado em dia da semana). Porém surgem desvantagens, tais como: a atividade de mercado tem que ser considerada uniforme nesta escala e mudança de preços durante o fechamento do mercado são captadas como mudanças de curto prazo.

Ao escolher trabalhar com o número de transações (utilização de dados de alta freqüência), os autores explicaram que esta escala de tempo consegue captar todos os efeitos do mercado ao longo de um dia de negociação. Mas a irregularidade do tempo entre as negociações é uma das fontes de aleatoriedade do modelo. Por isso, ao utilizar dados negócio a negócio é possível eliminar uma das fontes. Entretanto, o volume de transações, outra fonte geradora de aleatoriedade, ainda permanece no modelo.

Para fins deste trabalho, as medidas de aleatoriedade foram feitas em base de dados negócio a negócio.

2 **MÉTODO**

2.1 **Apresentação**

O método utilizado para análise da eficiência relativa de mercado, bem como para verificação da perda de memória das séries financeiras, foi baseado no modelo de determinação da medida de complexidade de uma sequência de dados proposto por Lempel e Ziv (1976) e desenvolvido computacionalmente por Kaspar e Schuster (1987).

Serão apresentadas, nas seções seguintes, a relação entre a teoria da informação algorítmica, complexidade e aleatoriedade, a descrição da medida proposta por Lempel e Ziv (1976) com o desenvolvimento de Kaspar e Schuster (1987) e o estudo das correlações entre retornos de séries financeiras apresentado por Mantegna e Stanley (2000).

Ao final deste capítulo será apresentado o procedimento para a análise da evolução da complexidade de Lempel e Ziv em séries de preços de alta frequência de ações que foram componentes do índice da BOVESPA ao longo dos anos de 2007 e de 2008, além da análise da eficiência relativa de mercado para intervalos maiores que uma negociação e da análise da perda de memória proposta por Mantegna e Stanley (2000) para vários comprimentos de intervalo.

2.2 **Aleatoriedade, complexidade e informação algorítmica**

O estudo das sequências aleatórias levaram Kolmogorov, Solomonoff e Chaitin a desenvolver, de forma independente, a teoria da informação algorítmica (VOLCHAN 2002). A partir dos fundamentos desta, foi possível Lempel e Ziv (1976) criarem uma medida numérica para avaliar o quão aleatória é uma sequência de dados. Sendo assim, este trabalho preocupou-se em, primeiramente, enunciar alguns dos conceitos de aleatoriedade de uma série de dados e formas de mensuração para, finalmente, no bojo da teoria da informação algorítmica, enunciar a ideia de complexidade de uma sequência proposta por Kolmogorov.

Na última seção foram apresentados alguns testes de aleatoriedade reunidos em Ruhkin (2000).

2.2.1 Sequências Aleatórias

Nesta seção são mostrados os principais conceitos de aleatoriedade de uma sequência que, de acordo com Campani e Menezes (2004) e Calude (1999) levaram Kolmogorov a desenvolver a sua teoria da complexidade. Os autores também evidenciaram que, ao começar a estudar as sequências aleatórias deve-se ter em mente que a definição de aleatoriedade é dependente da distribuição de probabilidade envolvida. Por exemplo, em uma moeda simétrica (cara e coroa equiprováveis), uma sequência em que ocorressesm mais caras que coroas seria logo identificada como não aleatória. Mas, uma moeda em que houvesse um desequilíbrio entre suas faces, poderia resultar em uma sequência deste tipo, a qual seria tratada como aleatória em relação a esta segunda distribuição.

Definir sequência aleatória como sequência imprevisível no sentido de não haver um conjunto de estratégias que possam levar um jogador a obter melhores resultados que poderia obter se apostasse ao acaso, levaria a um problema de dependência de duas variáveis: o tamanho mínimo da sequência para que haja convergência da freqüência relativa e o tamanho do raio de convergência.

Além disso, definir uma sequência como aleatória a partir de suas probabilidades levaria a outro problema, pois as sequências binárias 1111111111 e 1001011000 possuem a mesma probabilidade de ocorrência, apesar de a primeira intuitivamente ser menos aleatória que a segunda.

Em vista dos problemas apresentados, Volchan (2002) mostrou as três principais noções de aleatoriedade de sequências que foram propostas ao longo da história:

1) Estocasticidade ou estabilidade na freqüência. Autores: Von Mises, Wald e Church;
2) Incompressividade ou comportamento caótico. Autores: Solomonoff, Kolmogorov e Chaitin; e
3) Tipicalidade. Autor: Martin-Löf.

O presente trabalho tem por preocupação, enunciar de forma sucinta, os estudos desenvolvidos pelos autores mencionados que seguem logo abaixo.

A noção utilizada por Richard Von Mises para definir sequências aleatórias foi baseada na existência das sequências que possuíam limite de freqüência relativa. Estas foram chamadas de *Kollektivs*. A definição de Von Mises segue abaixo:

Definição 2.1: Uma sequência binária infinita $x = x_0x_1x_2$ ..., é uma sequência aleatória (*Kollektiv*) se:

1. $\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} x_i = p$
2. Toda sequência $x_{k^0}x_{k^1}x_{k^2}$ ..., obtida de x pela aplicação de alguma regra de seleção de posição admissível ϕ, deve satisfazer o item 1 desta Definição.

O problema apresentado na definição de Von Mises está no uso generalizado de funções parciais admissíveis de seleção de posição. Por exemplo: ao utilizarmos a função $\Phi_o(x_i) = V$ se $x_i = 1$ e indefinido caso contrário, obtém-se uma sequência não aleatória.

Em vista deste problema, Abraham Wald propôs tomar apenas um conjunto contável de funções para ser usado como regra de seleção de posição e, além disso, Alonzo Church sugeriu que somente as funções parciais recursivas poderiam ser tomadas como regras de seleção ensejando o uso de estratégias computáveis. A Definição 2.1 alterada pela implementação de funções parciais recursivas caracteriza as sequências Wald-Church aleatórias.

Porém, existem sequências Wald-Church aleatórias que, com limite de freqüência $\frac{1}{2}$ satisfazem $\frac{1}{n}\sum_{i=0}^{n-1} x_i \geq \frac{1}{2}$. Com isso, é possível obter ganhos infinitos ao apostar em um único número invalidando, desta forma, a aleatoriedade.

Campani e Menezes (2004) e Volchan (2002) observam que a tipicidade proposta por Martin-Löf (1966) refere-se a uma sequência não possuir características que a distingam das demais, ser comum. Se a sequência é típica, seu conjunto é bem maior do que o das sequências não típicas. Utilizando o conceito de medida de conjuntos (ISNARD 2007), verificou-se que, como o conjunto das sequências não típicas é pequeno demais, sua medida seria nula. No entanto é necessário ter cuidado, pois não é correto afirmar que uma sequência realmente aleatória satisfaz a todas as leis da aleatoriedade[7]. Isso quer dizer, em termos de medida, que é falsa a idéia de que uma

[7] Von Mises, Wald e Church verificaram que somente a Lei dos Grandes Números bastava para garantir aleatoriedade, o que também não estava correto (CAMPANI E MENEZES, 2004)

sequência típica satisfaz a todas as propriedades da aleatoriedade que ocorram com probabilidade igual a 1 no conjunto de todas as possíveis sequências ($\Sigma^{\mathbb{N}}$), ou seja, a sequência deveria pertencer à intersecção de conjuntos onde a propriedade fosse verificada com probabilidade unitária ($x \in \bigcap_{\alpha} \Sigma_{\alpha}$). Por exemplo, seja uma distribuição de Bernoulli (λ) para um alfabeto binário (uma propriedade da aleatoriedade). Logo, cada letra tem probabilidade ½ de ocorrência em uma sequência. Seja x uma sequência binária. Então $\lambda(\{x\}) = 0$ para todo x e $\lambda(\Sigma^{\mathbb{N}} - \{x\}) = 1$. Mas, $\bigcap_{\alpha} \Sigma_{\alpha} = \phi$, assim não havendo sequências típicas, o que é falso. Então, a formalização matemática de tipicidade precisaria sofrer alterações.

Martin-Löf (1966) propôs que as propriedades deveriam ser testadas utilizando-se um teste de aleatoriedade (uma função parcial recursiva) que exclui as sequências que apresentam alguma regularidade (não típicas). Desta forma, a definição de Martin-Löf propõe a existência de um conjunto de sequências ditas aleatórias, que possui complemento recursivamente enumerável, isto é, solucionável por uma máquina de Turing. Para detalhes sobre a máquina de Turing ver a seção 2.2.2.

Um conjunto de medida nula $X \subset \Sigma^{\mathbb{N}}$ é um conjunto que pode ser coberto por conjuntos elementares de tal forma que o conjunto de cobertura (união destes conjuntos elementares) possua medida tão pequena quanto se queira. Assim, uma propriedade vale quase toda parte em $X \subset \Sigma^{\mathbb{N}}$ quando existe conjunto de medida nula $A \subseteq X$ tal que a propriedade vale para todo $x \in X - A$ (ISNARD 2007).

Os conjuntos elementares definidos foram os cilindros (VOLCHAN 2002). São conjuntos da forma $\Gamma_w = \{x \in \Sigma^{\mathbb{N}} : x = wy\}$, onde $w \in \{0,1\}^{*}$[8]. Um cilindro nada mais é do que o conjunto de todas as sequências que possuam prefixo w[9].O intervalo real de um cilindro em [0,1] é dado por $(0, w; 0, w + 2^{-|w|}]$, onde $|w|$ é o comprimento do prefixo w (definido oportunamente na seção 0.).

Definição 2.2: Seja μ a medida de probabilidade ($0 \leq \mu \leq 1$) em $\Sigma^{\mathbb{N}}$. $X \subset \Sigma^{\mathbb{N}}$ é um conjunto de medida μ nula se, e somente se, para um dado número racional $\varepsilon > 0$, existe uma sequência de prefixos $w_0, w_1, w_2, \ldots \in \{0,1\}^{*}$ tal que:

$$X \subset \bigcup_{k \geq 1} \Gamma_{w_k} \qquad (2.1)$$

[8] $\{0,1\}^{*}$ é o conjunto de todas as sequências binárias finitas
[9] wy é uma concatenação de sequências que forma x

$$\sum_{k\geq 1} \mu(\Gamma_{w_k}) < \varepsilon$$

$\cup_{k\geq 1} \Gamma_{w_k}$ é a cobertura de X. Caso μ seja uma medida de Bernoulli (½, ½), tem-se $\mu(\Gamma_{w_k}) = 2^{-|w_k|}$.

Um conjunto $X \subset \Sigma^{\mathbb{N}}$ de medida μ efetivamente nula, conforme proposto por Martin-Löf, é aquele para o qual existe um algoritmo (uma máquina de Turing) que recebe um número racional $\varepsilon > 0$ como entrada e que enumera um conjunto de prefixos $w_0, w_1, w_2, \ldots$ tais que as condições da Definição 2.2 sejam satisfeitas. Um conjunto de medida efetiva unitária é obtido por complementação. Isso implica que um conjunto de medida μ efetivamente nula é um conjunto que possui medida nula e que pode ser algoritmicamente gerado, sendo, portanto, recursivamente enumerável.

Qualquer subconjunto de um conjunto de medida μ efetivamente nula é também um conjunto de medida μ efetivamente nula. Um conjunto unitário é um conjunto de medida μ efetivamente nula se seu elemento é computável (gerado por algoritmo, portanto, não aleatório).

Uma medida de probabilidade μ é computável quando, para cada racional positivo ε e para cada prefixo $w \in \{0,1\}^*$ existe uma função computável (por uma máquina de Turing) F que toma como entrada o vetor (ε, w) e gera a saída $F(\varepsilon, w)$ de tal modo que: $|F(\varepsilon, w) - (\varepsilon, w)| \leq \varepsilon$. Um exemplo de medida computável é a de Bernoulli.

Martin-Löf obteve, então, o seguinte resultado:

Teorema 2.1: Seja μ uma medida de probabilidade computável. A intersecção dos conjuntos de medida μ efetiva unitária é não vazia e possui medida μ efetiva unitária.

O teorema acima mostra a consistência entre o conjunto de sequências aleatórias e o conjunto de sequências típicas.

Os testes sequenciais de aleatoriedade de Martin-Löf (1966) apud Volchan (2002) consistem em uma sequência enumerável recursiva de intervalos $\{I_m^n\}$ tais que, para cada m, temos $\mu(I_m^n) < 2^{-m} = \varepsilon$. Aplicar este teste em uma sequência x significa escolher um nível de confiança m e verificar se x pertence ou não a $\{I_m^n\}$ para $n \geq 1$. x falha no teste quando $x \in \{I_m^n\}$, sendo considerada não aleatória no nível m. Caso passe no teste, x passa para a próxima avaliação. Então, cada teste verifica certa propriedade de não aleatoriedade (regularidade) de x. Portanto, a sequência é tida como aleatória no sentido de Martin-Löf se ela passar por todos os testes de

aleatoriedade. O teorema 2.1 implica a existência de um teste sequencial universal que, caso seja aprovada neste, a sequência x é tida como aleatória.

Corolário 2.1: Uma sequência computável $x \in \Sigma^{\mathbb{N}}$ é μ-típica se, e somente se, $\mu\{x\} > 0$. Ou seja, para a sequência ser típica, ela deve ocorrer com probabilidade diferente de zero.

Para Volchan (2002) e Campani e Menezes (2004), a definição de aleatoriedade de Martin-Löf é a melhor existente, apesar do requisito de computabilidade inerente aos testes e à sequência

Assim, os pontos principais da definição de sequência aleatória de Martin-Löf foram mostrados e, na próxima seção será mencionada a definição de aleatoriedade de Kolmogorov sendo, de acordo com Calude (1999) equivalente à de seu discípulo Martin-Löf.

2.2.2 Complexidade de Kolmogorov, Teoria da Informação Algorítmica e compressividade.

A definição de aleatoriedade de uma sequência segundo Kolmogorov encontra-se inserida na teoria da informação algoritmica.

A teoria da informação algorítmica relaciona-se tanto com a teoria da informação como com os fundamentos da Ciência da Computação. Esta disciplina relaciona-se com o conceito de aleatoriedade por meio da definição de complexidade proposta por Kolmogorov. Gregory Chaitin define informalmente a Teoria da Informação Algoritmica como sendo a junção da teoria de informação proposta por Shannon com a teoria da computabilidade proposta por Turing (GIGLIO 2008).

Shannon (1948) mostrou que o grau de incerteza inerente ao conteúdo informacional transmitido sob a forma de uma sequência de caracteres pode ser medida a partir da entropia da fonte emissora. Quando esta medida atinge seu grau máximo, a sequência seria tida como perfeitamente aleatória (para uma definição matemática da entropia de Shannon ver seção 0)

De acordo com Teixeira (1998), Alan Turing desenvolveu uma máquina ideal (autômato), que constitui a melhor formalização da noção de um algoritmo que se tem noção na história da matemática, para buscar a solução do Problema de Decisão

(*Eitschendungsproblem*) formulado por Hilbert que consistia em determinar se todos os enunciados matemáticos verdadeiros poderiam ou não ser provados, ou seja, serem deduzidos a partir de um dado conjunto de premissas.

Para Turing um algoritmo é um processo ordenado por regras, que enuncia o procedimento para que um problema seja resolvido (TEIXEIRA, 1998, p.20).

Uma máquina de Turing possui dois componentes:

1) Uma fita infinitamente longa, dividida em pequenos quadrados contendo em cada um destes um conjunto finito de símbolos (entrada); e
2) Um dispositivo mecânico (*scanner*) que lê, imprime e apaga os símbolos que estão no quadrado.

A representação gráfica habitual de uma máquina de Turing é mostrada na Figura 1

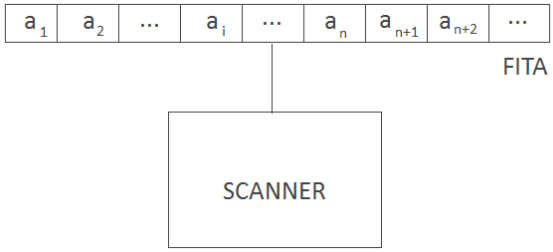

Figura 1: Esquema de uma máquina de Turing[10].

O comportamento desta máquina é determinado por um programa, definido por um número finito de instruções que informam à máquina o tipo de computação que ela deve efetuar. Tais instruções são inseridas em uma lista de estados (número finito) característica de cada máquina e que fica armazenada no *scanner*. A partir de um estado inicial e de um símbolo de entrada, a máquina inicia a execução do programa. A Figura 2 ilustra um exemplo de esquema de operações e, para cada par estado-entrada, uma instrução é executada.

[10] Teixeira (1998) com alterações

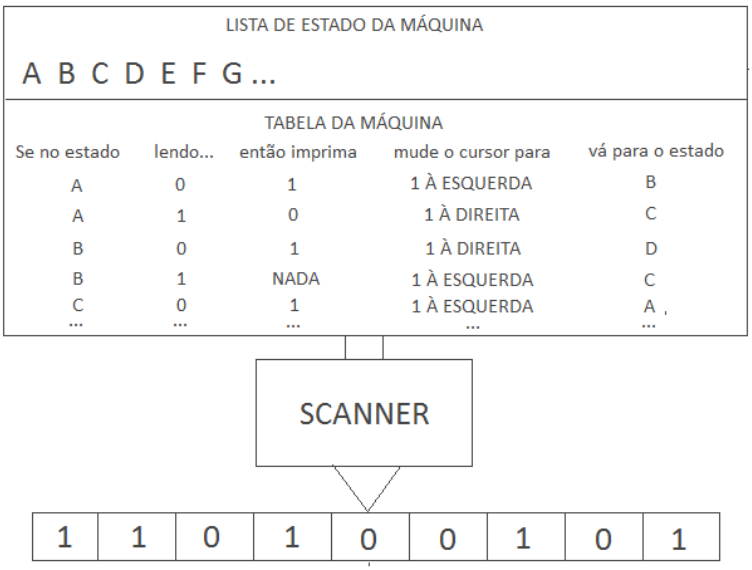

Figura 2: Representação de possíveis operações de uma máquina de Turing[11].

Com exemplo, pode-se tomar a máquina de Turing da Figura 2 para explicar um pouco seu funcionamento. Seja A o estado inicial da máquina e a posição inicial do *scanner* sobre a fita idêntica à da figura. Logo, para uma entrada igual a "0", o conjunto de instruções diz que deve imprimir o símbolo "1" na fita, deslocá-la uma casa para a esquerda e mudar para o estado B. Ao deslocar a fita para a esquerda, o *scanner* aponta para o símbolo "1". O par (B,1) indica que deve-se seguir a instrução da linha 4: não imprimir nada, deslocar a fita uma casa para a esquerda e mudar para o estado C. Então, a partir do novo par (C,0) parte-se para a instrução da linha 5 e inicia-se uma nova computação. O procedimento e repetido até o surgimento da instrução "PARE", no qual a máquina encerra as operações.

A definição formal de uma máquina de Turing (M) é dada por uma sétupla, de acordo com Moreira (2003):

$$\mathrm{M} = (\mathrm{Q}, \Sigma, \Gamma, \mathrm{s}, \mathrm{b}, \mathrm{F}, \delta) \tag{2.2}$$

onde:

a) Q é um conjunto finito de estados

b) Σ é um alfabeto finito de símbolos

c) Γé o alfabeto da fita (conjunto finito de símbolos)

d) $s \in Q$ é o estado inicial

[11] Teixeira (1998) com alterações.

e) $b \in \Gamma$ é o símbolo branco (o único símbolo que se permite ocorrer na fita infinitamente em qualquer passo durante a computação)
f) $F \subseteq Q$ é o conjunto dos estados finais
g) $\delta: Q \times \Gamma \to Q \times \Gamma \times \{E, D\}$ é uma função parcial chamada função de transição, onde E é o movimento para a esquerda e D é o movimento para a direita da fita.

Existem outros tipos de máquina de Turing, como máquina com várias fitas, com fitas n-dimensionais, com o programa sendo executado a partir de instruções contidas em uma fita.

Uma máquina de Turing universal (MTU) é um dispositivo que executa um programa codificado em uma fita. Se esta contém o programa de uma máquina T, então a máquina universal irá produzir os mesmos resultados que T produziria para uma mesma fita com dados de entrada. Uma das formas de se pensar em uma MTU, é um computador programável. Na máquina de Turing usual, o programa já está inserido nela e não é possível alterá-lo. Já na MTU, para mudar o programa (conjunto de instruções), basta mudar a codificação da fita.

A partir das idéias mencionadas e com base na tese de Church-Turing, a qual enuncia que, dado um problema cuja resolução é obtida por meio de algoritmos (problemas efetivamente computáveis), certamente existe uma máquina de Turing que o resolva (MOREIRA, 2003), além de outras descobertas, propiciaram que Kolmogorov, Solomonoff e Chaitin desenvolvessem, de forma independente, a Teoria da informação algorítmica.

Isso significa que, se uma máquina de Turing consegue gerar uma determinada sequência, esta sequência é efetivamente computável. Chaitin (1966) utilizou uma máquina de Turing para reproduzir uma sequência finita e de alfabeto binário e, também, buscou determinar o número de instruções (tamanho do programa) que a produziria.

De acordo com Li e Vitányi (1997) é possível pensar, intuitivamente, que algumas sequências são consideradas complexas, dado o elevado número de bits que um programa teria que ter para reproduzi-la e outras são simples, como 00000000000 (o procedimento de descrição utilizado para esta sequência possui tamanho pequeno).

Para que se faça uma descrição apropriada do conteúdo informacional da sequência, é necessário que se tenha um método universal e que os algoritmos utilizados sejam os melhores possíveis, no sentido de que têm o menor tamanho possível para cada tipo de sequência que se pretende descrever.

Além disso, a linguagem de programação (tipo de computador) pode influenciar a produção de informação sobre uma dada sequência. Portanto, para que a informação sobre uma sequência seja adequada, é necessário que ela seja independente dos meios de descrição.

A complexidade de uma sequência proposta por Kolmogorov consiste em encontrar o menor tamanho de programa (critério informacional) que, ao ser executado em um computador determinado a priori, consiga reproduzi-la (LI E VITÁNYI 1997).

Formalizando matematicamente a idéia do parágrafo anterior:

$$\begin{cases} C_f(x) = \min\{l(p): f(p) = n(x)\} \\ C_f(x) = \infty \; quando \; p \; não \; existir \end{cases} \tag{2.3}$$

onde C_f é o número que representa a complexidade, p é o programa, $l(p)$ é o tamanho do programa, $f(p)$ é uma função parcial recursiva (computador) e $n(x)$ é uma enumeração de uma sequência $x \in S$ (S é o conjunto de sequências).

Para que o cálculo da complexidade fique independente do tipo de computador, adota-se como linguagem padrão a máquina de Turing Universal.

Para Kolmogorov, uma sequência é tida como aleatória (complexa) quando $C_f(x) \geq |x|$.

Isso quer dizer que o tamanho do menor programa que gera uma sequência aleatória é, pelo menos, o tamanho em bits desta. A sequência aleatória é, portanto incompressível no que se refere a não ser possível reduzi-la a um objeto de tamanho inferior a ela.

Calude (1999) demonstrou que são equivalentes as proposições de aleatoriedade formuladas por Martin-Löf e pela incompressividade de uma sequência, desta forma garantindo a consistência das idéias de Kolmogorov, dado que o conceito de sequência aleatória de Martin-Löf, de acordo com Campani e Menezes (2004), é tido como robusto por muitos cientistas.

A complexidade de Kolmogorov possui duas características principais: incomputabilidade e invariância.

Por incomputabilidade, entende-se que, dada uma sequência x não existe máquina de Turing que produza $C_f(x)$. Isso implica que qualquer procedimento computacional para encontrar o valor exato de $C_f(x)$ não funcionará. Somente será possível obter aproximação para tal medida.

É também possível perceber que $C_f(x)$ independe da escolha do computador (linguagem). É possível realizar medidas aproximadas de complexidade com máquinas diferentes. A esta propriedade é dada o nome de invariância. Formalizando a propriedade, tem-se:

$$C_f(x) \leq C_g(x) + c_g \tag{2.4}$$

onde f e g são linguagens distintas e c_g é uma constante que depende somente das linguagens utilizadas e não da sequência a ser analisada. Logo, máquinas diferentes utilizadas para medir complexidade necessitam de programas cuja diferença de tamanho é, no máximo, uma constante. Desta forma, se a sequência é considerada complexa por uma linguagem, também o será por outra.

2.2.3 Alguns procedimentos para a verificação da aleatoriedade de uma sequência de caracteres.

Rukhin (2000) reuniu em seu trabalho alguns testes para verificar o caráter aleatório de uma sequência de dados. Nesta seção foram desenvolvidos alguns procedimentos dentre os apontados pelo autor, enquanto outros foram apenas listados.

A primeira das verificações apresentadas é o teste de hipótese de Kolmogorov – Smirnov em que a hipótese nula é a aleatoriedade e a hipótese alternativa é o comportamento não aleatório.

Seja S uma sequência que é partida em N subsequências de comprimento M. Suponha que existam $K + 1$ classes de sequências cujas freqüências observadas sobre a partição de S são dadas por $v_0, v_1, v_2, \dots, v_k$ e que $v_0 + v_1 + \cdots + v_k = N$. Suponha que $\pi_0, \pi_1, \pi_2, \dots, \pi_k$ sejam as probabilidades teóricas das classes mencionadas. Sabe-se que a estatística

$$\sum_{i=0}^{K} \frac{(v_i - N\pi_i)^2}{N\pi_i} \tag{2.5}$$

tem distribuição qui-quadrado com K graus de liberdade. Quanto menor o valor p do teste há mais razões para se rejeitar a hipótese nula.

Outros procedimentos são baseados nas propriedades do passeio aleatório. Tais avaliações servem para sequências binárias cujo alfabeto é {0,1}. Há dois testes enunciados para esta categoria.

No passeio aleatório, cada termo da sequência é denotado por ε_k, $k = 1,2,\dots,n$, podendo assumir os valores 0 ou 1. Por questões de conveniência é feita a seguinte transformação $X_k = 2\varepsilon_k - 1, \quad k = 1,2,\dots,n$. Sabe-se que, no limite, $\frac{S_n}{\sqrt{n}}$ tem distribuição normal com média zero e variância 1, onde $S_n = X_1 + X_2 + \cdots + X_n$. A hipótese nula para os testes utilizando passeio aleatório implica aleatoriedade da sequência.

O teste baseado no maior dos valores absolutos das somas parciais do passeio aleatório é realizado com base na seguinte estatística: $\max_{1\le k\le n}|S_k|$. A estatística

$$z = \max_{1\le k\le n} \frac{|S_k|(observado)}{\sqrt{n}} \tag{2.6}$$

é utilizada para o teste, sendo o valor P deste dado por $1 - H(z)$, onde

$$H(z) = \frac{4}{\pi}\sum_{j=0}^{\infty} \frac{(-1)^j}{2j+1} e^{\left\{-\frac{(2j+1)^2\pi^2}{8z^2}\right\}} \tag{2.7}$$

quando z for pequeno. Para valores grandes de z,

$$H(z) \approx 1 - \frac{4}{3z\sqrt{2\pi}} e^{\left\{-\frac{9z^2}{2}\right\}}. \tag{2.8}$$

Quanto maior o valor de z, há mais evidências para se rejeitar a hipótese nula, ou seja, a hipótese de aleatoriedade.

Ainda nos testes envolvendo passeio aleatório existem aqueles baseados no número de visitas realizadas sobre uma sequência binária. Seja J o número de zeros dentre as somas S_k, $k = 1,2,\dots,n$, onde $S_0 = 0$. A hipótese de aleatoriedade pode ser rejeitada se J observado for muito pequeno. O valor p do teste é:

$$P[J < J(obs)] \approx \sqrt{\frac{2}{\pi}} \int_0^{J(obs)/\sqrt{n}} e^{-\frac{u^2}{2}}\, du \tag{2.9}$$

Existem ainda os ***run tests*** que podem ser baseados na concepção clássica de *run* ou na concepção de Feller de *run*. Para tais avaliações, a hipótese nula é de aleatoriedade.

Na concepção clássica, *run* é a quantidade de sucessões de um ou mais zeros que são seguidos ou precedidos tanto por um quanto por nenhum símbolo. O comprimento (r) de um *run* é dado pelo número de zeros consecutivos que ele possui.

Para Feller (1968) apud Rukhin (2000) os *runs* são eventos recorrentes e não se sobrepõem a outros *runs*. A partir desta definição, é possível admitir uma função geradora para os momentos de ocorrência de um *run* de comprimento r. A média é dada por $\mu = 2^{r+1} - 2$ e a variância $\sigma^2 = 2^{2(r+1)} - (2r+1)2^{r+1} - 2$.

Com isso, a estatística:

$$z(obs) = \frac{\sqrt{\mu}(\mu N_r(obs) - n)}{\sigma\sqrt{n}} \qquad (2.10)$$

possui distribuição aproximadamente normal padronizada com valor p igual a $2(1 - \Phi(|z(obs)|))$, onde n é o tamanho da sequência e $N_r(obs)$ é o número de *runs* de comprimento r observados na sequência.

Existe uma distribuição similar, no limite, para a definição clássica de um *run*. Se M_r é o número total de *runs* com comprimento r, então a estatística

$$\frac{(M_r - E[M_r])}{\sqrt{Var[M_r]}}, \qquad (2.11)$$

onde $E[M_r]$ é o valor esperado para M_r e $Var[M_r]$, é a variância de M_r, aproxima-se de uma normal padronizada.

Há também um grupo de testes baseados em padrões, que utilizam a ocorrência de palavras repetidas (conjunto de símbolos concatenados) de um dado tamanho ao longo da sequência para verificar a aleatoriedade. Um número grande de ocorrência de palavras repetidas pode ser um forte indício para descartar o caráter aleatório daquela. Nesta categoria, foram elencados os procedimentos abaixo:

a) Testes baseados em freqüência de padrões;

b) Testes baseados no número de palavras faltantes;

c) Teste de entropia aproximada; e

d) Testes seriais.

O último grupo de testes considerado significativo para Rukhin (2000) são baseados em compressão de dados. A idéia é que as sequências aleatórias não podem

ser comprimidas (ver seção 2.2.2 deste trabalho). Os procedimentos que fazem parte deste grupo são:

a) Teste de Lempel e Ziv (ver seção 2.3 deste trabalho);
b) Teste de Maurer;
c) Teste do posto de matrizes aleatórias; e
d) Teste de complexidade linear.

Este trabalho utilizou somente o procedimento de Lempel e Ziv (1976).

2.3 **O modelo de Lempel e Ziv e o posterior desenvolvimento computacional de Kaspar e Schuster.**

Esta seção foi reservada para a apresentação do modelo apresentado por Lempel e Ziv (1976) em seu artigo "On the Complexity of Finite Sequences" e para mostrar a extensão feita por Kaspar e Schuster (1987) em seu artigo "Easily calculable measure for the complexity of spatiotemporal patterns".

2.3.1 O modelo de Lempel e Ziv

O objetivo proposto por Lempel e Ziv (1976) foi o de definir um número que verificasse a existência de aleatoriedade de uma sequência finita. Porém, segundo os autores, a medida proposta não é absoluta: esta é produzida a partir do uso de uma ***self delimiting learning machine*** que, ao ler uma sequência da esquerda para a direita, adiciona uma palavra nova à sua memória quando descobre uma subsequência de dígitos consecutivos que não foi lida anteriormente, ou seja, somente são permitidas as operações de leitura e de impressão.

O processo de leitura e aprendizado dos elementos da sequência pela máquina referida no parágrafo anterior foi definido pelos autores, bem como a notação necessária à sua compreensão. Uma descrição detalhada das definições e dos termos segue abaixo. Nesta seção foram utilizadas as mesmas notações propostas por Lempel e Ziv (1976).

Seja A^* o conjunto de todas as sequências definidas em um alfabeto finito A. Este representa o conjunto de caracteres distintos que formarão as sequências em A^*. Seja

também $l(S)$ o comprimento de uma sequência $S \in A^*$. O conjunto A^n é o conjunto de todas as possíveis sequências de *A** cujo comprimento é igual a *n*. A sequência nula Λ é aquela que $l(\Lambda) = 0$.

Uma sequência $S \in A^n$ é completamente especificada quando se escreve $S = s_1 s_2 \dots s_n$. Quando S é formada a partir de um único elemento de A, é possível escrevê-la como $S = a^n; a \in A$.

Uma subsequência $S(i,j)$ de S que inicia na posição i e termina na posição j, fica definida como $S(i,j) = s_i s_{i+1} s_{i+2} \dots s_j$, se $i \leq j$ ou $S(i,j) = \Lambda$, se $i > j$.

As sequências $Q \in A^m$ e $R \in A^n$ formam a sequência $S = QR = q_1 q_2 \dots q_m r_{m+1} r_{m+2} \dots r_{m+n} \in A^{m+n}$ por meio do processo de concatenação. $Q = S(1;m)$ e $R = S(m+1;m+n)$ são subsequências (ou palavras) de S. Ao concatenar S com ela mesma, utiliza-se a notação $S^2 = SS$. As extensões deste tipo de notação são: $S^0 = \Lambda$ e $S^i = S^{i-1}S \;\; , i \geq 1$.

Q é chamado de prefixo de $S \in A^*$, e S é chamado de extensão de Q se existir um inteiro i tal que $Q = S(1,i)$, $i \geq 1$. O prefixo Q e a extensão S são tidos como próprios quando $l(Q) < l(S)$.

Define-se um operador π para identificar prefixos de S. O operador aplicado a S gera prefixos sob a forma $S\pi^i = S(1, l(S) - i)$, $i \in \mathbb{N}$. As extensões da definição de π são $S\pi^0 = S$ e $S\pi^i = \Lambda, i \geq l(S)$.

Exemplo: $S = 12345$. Então $S\pi^2 = 123$

Define-se vocabulário ($v(S)$) de uma sequência S, $v(S) \subseteq A^*$ o conjunto formado por todas as palavras de S.

Exemplo: $v(0010) = \{\Lambda, 0,1,00,01,10,001,010,0010\}$

Uma palavra $Q \in v(S)$ é chamada autopalavra (*eigenword*) de S quando ela não pertencer ao vocabulário de nenhum prefixo próprio de S. O autovocabulário (*eigenvocabulary*) $e(S)$ é o conjunto das autopalavras de S.

Exemplo: $S = 0010$. Prefixos próprios de S: $Q = 0$, $R = 00$, $T = 001$. Então, $v(Q) = \{\Lambda, 0\}$, $v(S) = \{\Lambda, 0,00\}$, $v(T) = \{\Lambda, 0,00,01,001\}$. Logo, $W = 10$, $K = 010$, $U = 0010$ são autopalavras de S. $e(S) = \{10,010,0010\}$.

A partir das definições de autopalavra e de autovocabulário, são desenvolvidos os seguintes lemas (a numeração dos lemas foi mantida a mesma do artigo original de Lempel e Ziv (1976)):

Lema 2.1: $v(S\pi) \subset v(S)$

Lema 2.2: $e(S) = v(S) - v(S\pi)$ (2.12)

Seja uma sequência S e $W = S(i,j)$ uma de suas palavras. A concatenação $R = SW$ pode ser vista como sendo obtida a partir de S por um procedimento de cópia onde w_m é copiado de s_{i+m-1}, $m = 1,2,\dots,j-i+1$ assumindo a posição em R no termo $r_{m+l(S)}$. Baseado no mesmo procedimento de cópia é possível gerar uma extensão $R = SQ$ de S, que pode ser maior do que qualquer palavra em $v(S)$. A única condição é que $Q \in v(SQ\pi)$. Tal condição implica a existência de um inteiro positivo $p \leq l(S)$ de modo que o termo q_i assuma a posição $r_{p+i-1}, i = 1,2,\dots,l(Q)$ e que R possa ser gerada a partir de S por meio da cópia dos termos que se iniciam na posição p na sequência R.

Uma extensão $R = SQ$ é reprodutível a partir de S ($S \to R$) se $Q \in v(SQ\pi)$. Uma posição p de S tal que $Q = R(p, l(Q) + p - 1)$ é chamada ponteiro para a reprodução ($S \to R$).

Uma sequência S é produtível a partir de seu prefixo $S(1,j)$, se $S(1,j) \to S\pi$ e $j < l(S)$. A notação utilizada é $S(1,j) \Rightarrow S$ e $S(1,j)$ é uma base de S. Toda sequência não nula possui uma base e Λ é uma base para todo símbolo $a \in A$, mas não para outra sequência $S \in A^*$.

A produtibilidade é diferente da reprodutibilidade no que se refere ao processo de cópia recursiva que é próprio da última denominação. A produção permite que se inove um caractere ao final do processo de cópia.

Qualquer ponteiro para $S \to R\pi$ também será ponteiro para $S \Rightarrow R$ (Por exemplo: $01 \Rightarrow 0100$ possui ponteiro p = 1, mas $01 \nrightarrow 0100$ (0100 não é reprodutível a partir de 01)).

O processo de produção de S é dado pela repetição dos passos de produção $S(1,h_i) \Rightarrow S(1,h_{i+1})$ consecutivamente iniciando com $i = 0$ e acabando em $i = l(S) - 1$. O resultado da produção $S(1,h_k)$ do k-ésimo passo é chamado k-ésimo estado do processo.

A história de produção de uma sequência S, $H(S)$ é dada por:

$H(S) = S(1,h_1)S(h_1+1,h_2)S(h_2+1,h_3)\dots S(h_{m-1}+1,h_m)$, com i=1,2...,m

Os componentes de $H(S)$ são as palavras $H_i(S) = S(h_{m-1}+1,h_m)$ e $h_0 = 0$

O componente $H_i(S)$ e sua produção correspondente $S(1, h_{i-1}) \Rightarrow S(1, h_i)$ são chamados de exaustivos se $S(1, h_{i-1}) \nrightarrow S(1, h_i)$.

A história $H(S)$ é chamada de exaustiva se todos os seus componentes forem exaustivos com exceção do último que pode ou não ser exaustivo. Denota-se história exaustiva por $E(S)$. Por exemplo: a sequência S = 0001101001000101 pode ser decomposta da seguinte forma: $0 \cdot 001 \cdot 10 \cdot 100 \cdot 1000 \cdot 101$

Os termos à esquerda de cada ponto são exaustivos. O último termo não é exaustivo (não possui ponto à direita).

A medida de complexidade de uma sequência finita proposta por Lempel e Ziv (1976) $c(S)$ é calculada da seguinte forma: $c(S) = \min\{c_H(S)\}$, onde $c_H(S)$ é o número de componentes de uma história de S. A minimização é dada sobre todas as histórias de S.

Lempel e Ziv (1976) enunciaram alguns teoremas e lemas decorrentes de sua medida de complexidade que não serão demonstrados neste trabalho. As demonstrações encontram-se no próprio artigo de Lempel e Ziv, bem como o trecho que se refere à história primitiva de uma sequência.

Teorema 2.2: $c(S) = c_E(S)$, onde $c_E(S)$ é o número de componentes em $E(S)$, a história exaustiva de S.

Para enunciar os outros teoremas é necessário observar as seguintes convenções: o tamanho do alfabeto (número de elementos) A é dado por $\alpha = |A|$ e $\log(x)$ significa logaritmo de x na base α salvo menção em contrário.

Teorema 2.3: Para todo $S \in A^n$, tem-se:

$$c(S) < \frac{n}{(1-\varepsilon_n)\log(n)}, \text{ onde } \varepsilon_n = 2\frac{1+loglog(\alpha n)}{\log(n)} \quad (2.13)$$

O teorema 2.4 abaixo mostra que, a partir da definição de complexidade de Lempel e Ziv (1976) quase todas as sequências de comprimento suficientemente grande são complexas. Tal propriedade mostra-se útil quando complexidade está relacionada com aleatoriedade.

Teorema 2.4: Para todo número positivo ε,

$$\lim_{n\to\infty} P\left(c(S) < \frac{n(1-\varepsilon)}{\log(n)} \middle| l(s) = n\right) = 0 \quad (2.14)$$

Como o teorema 2.4 não é muito restritivo, surgiu a necessidade de se demonstrar que o critério adotado por Lempel e Ziv (1976) não permitiria que sequências que não

apresentassem comportamento aleatório fossem classificadas como complexas. A solução encontrada para o problema acima foi considerar a sequência como se tivesse sido emanada de uma fonte discreta e ergódica e relacionar a complexidade da sequência com a entropia da fonte. Com isso obteve-se o seguinte limite superior:

$$c(s) = \frac{hn}{\log(n)} \tag{2.15}$$

onde h é a entropia normalizada da fonte ergódica.

De acordo com Shannon (1948), as fontes discretas de informação podem ser representadas por meio de processos de Markov (caso geral: Um sistema deve possuir um número finito de estados possíveis $S_k \;\; k = 0,1,2,\dots,n$ e, também, deve possuir as probabilidades de transição p_{ij} de um estado i para um estado j). Dentre tais processos existe um subconjunto que possui propriedades importantes para o estudo da comunicação que são os processos ergódicos. Ainda, segundo o autor, toda sequência produzida por um mesmo processo ergódico é a mesma em propriedades estatísticas. Isso significa que, a freqüência que um determinado caractere aparece na sequência tende a atingir um limite bem definido à medida que o tamanho daquela aumenta. Isto não é verdade para qualquer sequência, mas o conjunto de sequências para o qual a afirmação é falsa tem probabilidade zero.

Um processo ergódico, de acordo com Bendat e Piersol (1986) pode se dividir em fracamente ergódico e em fortemente ergódico.

O processo é caracterizado por fracamente ergódico quando a média espacial (média de um processo aleatório $\{x_k(t)\}$, onde k representa cada categoria de função amostral e t representa o tempo, para valores fixos deste) assume valor igual ao da média temporal (média aritmética das imagens de uma função amostral $x_k(t)$ ficando k, neste caso, fixo) e quando as funções de auto-covariância e de auto-correlação possam ser calculadas utilizando a média temporal em suas expressões em vez da média espacial.

Já no processo fortemente ergódico, qualquer estatística que utilize a média espacial em sua expressão poderá ser substituída pela média temporal.

Ainda em Shannon (1948), a entropia da fonte (h) é uma medida da taxa de informação produzida por uma fonte ergódica. Dado que exista um conjunto de eventos com probabilidade de ocorrência p_i, $i = 1,\dots,n$ torna-se necessário medir o

grau de incerteza sobre a escolha de um evento dentre os existentes. O autor mostrou que

$$h = -k \sum_{i=1}^{n} p_i \log(p_i) \tag{2.16}$$

onde k é apenas um fator de escala e o logaritmo é na base n.[12]

O valor de h é máximo quando os eventos são equiprováveis (maior incerteza possível para a "escolha" de um evento) e zero quando um dos eventos tiver probabilidade igual a 1 de ocorrer. A Figura 3 ilustra o comportamento de h em função de p para 2 (dois) eventos.

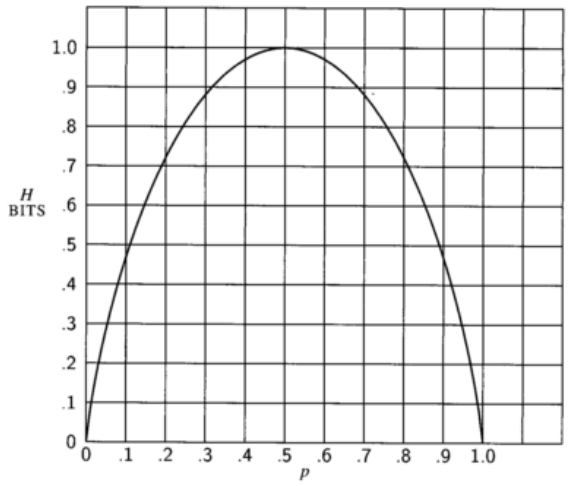

Figura 3: Evolução da entropia da fonte *versus* probabilidade para dois eventos[13].

A consequência do Teorema 2.4 implica que, para uma sequência S, tem-se

$$\lim_{n\to\infty} \frac{c(S)\log(n)}{hn} = 1. \tag{2.17}$$

O valor $\frac{c(S)\log(n)}{hn}$ é chamado de complexidade de Lempel e Ziv normalizada pela entropia (neste trabalho tal valor é denominado *LZ*).

Teorema 2.5: Subaditividade: Sejam Q e S duas sequências, então:

$$c(QS) \leq c(Q) + c(S) \tag{2.18}$$

Uma sequência de De Bruijn, conforme apresentado em Hurlbert e Isaak (1993) é uma sequência cíclica S onde, dado o tamanho do alfabeto m (m inteiro) e dado o comprimento $l(S) = m^n$, onde n é um número inteiro e S possui a propriedade de que toda subsequência de S de comprimento n aparece exatamente uma vez no ciclo. Em De Bruijn (1975), ao apresentar os ciclos P_n de sequências cíclicas binárias de comprimento 2^n, a propriedade destes é a de que todos os conjuntos ordenados de n

[12] Neste trabalho, k = 1 como Kaspar e Schuster (1987)
[13] Shannon (1948)

dígitos ocorrem exatamente uma vez no ciclo. Por exemplo: P_3 apresenta dois ciclos que podem ser dispostos sob as seguintes sequências (binárias): 00010111 e 11101000. Esta sequência é importante do ponto de vista do estudo da complexidade dada sua aplicação em estudos de números pseudo aleatórios (HURLBERT E ISAAK 1993) e devido ao fato de as sequências de De Bruijn serem consideradas boas aproximações finitas de sequências complexas (LEMPEL E ZIV 1976)

Teorema 2.6: Se S é uma sequência de De Bruijn de comprimento $n = \alpha^k + k - 1$, então

$$c(S) \geq \frac{n}{\log(n)} \tag{2.19}$$

O Teorema 2.6 implica que, caso S seja uma sequência de De Bruijn tem-se que $\frac{c(S)\log(n)}{n}$ tem um limite inferior de 1.

Analisando a história exaustiva de uma sequência além da produção de história relacionada à taxa de crescimento de seu vocabulário, Lempel e Ziv (1976) propuseram os lemas e teoremas abaixo.

Seja $e(S)$ o autovocabulário de S. A cardinalidade de $e(S)$ é dada por $k(S)$ e é referida como autovalor de S.

Lema 2.3: $k(\Lambda) = 0$ e para $S \neq \Lambda$, $1 \leq k(S) \leq l(S)$

$k(S) = l(S)$ se e somente se o último símbolo de S difere de todos os seus predecessores para $S \neq \Lambda$

Teorema 2.7: $e(s) = \{S(i, l(S)) | 1 \leq i \leq k(S)\}$

Lema 2.4: $k(S\pi) \leq k(S)$

Então, se $S\pi$ é estendida até S, as autopalavras de $S\pi$ se tornam autopalavras de S e S pode ter novas autopalavras devido ao último símbolo de S.

O conjunto $e(S)$ representa o crescimento por símbolo do vocabulário de S na ultima posição de S.

O lema 2.4 mostra que a taxa de crescimento do vocabulário em uma dada posição de S possui pelo menos o mesmo valor do que possuía na posição anterior de S.

Lema 2.5: $S(i, j) \rightarrow S$ se, e somente se $k(S) \leq j \leq l(S)$

Lema 2.6: $S(i, j) \Rightarrow S$ se, e somente se $k(S\pi) \leq j \leq l(S) - 1$.

Então, o comprimento do menor prefixo de S para que esta seja reprodutível é igual a $k(S)$. Analogamente, o menor comprimento do prefixo de S para que esta seja produtível é $k(S\pi)$.

2.3.2 Desenvolvimento computacional proposto por Kaspar e Schuster (1987).

O trabalho dos autores Kaspar e Schuster (1987) foi aplicar a medida de complexidade proposta por Lempel e Ziv (1976) para analisar a complexidade de padrões espaço-temporais. Estes autores apresentaram um algoritmo para o cálculo da complexidade com base na classe de programas que permite somente duas operações: inserção e cópia de elementos de uma sequência.

Seja uma sequência $s_1 s_2 s_3 \dots s_n$ que foi reconstruída a partir de um programa até o dígito s_r e que este dígito foi inserido (ou seja, não foi copiado a partir de $s_1 s_2 s_3 \dots s_{r-1}$). Seja $S = s_1 s_2 s_3 \dots s_r \cdot$, onde o $\cdot$ indica que s_r foi inserido. Para determinar se o resto da sequência foi inserido ou copiado, procede-se da seguinte forma: toma-se $Q = s_{r+1}$ e verifica se $Q \in v(SQ\pi)$. Em caso afirmativo, Q é obtida por meio de cópia de uma palavra de S. Então, concatena-se o próximo elemento a Q, obtendo $Q = s_{r+1}s_{r+2}$ e verifica se $Q \in v(SQ\pi)$ até o momento em que Q não possa mais ser obtida por meio de cópia de um elemento de $v(SQ\pi)$ sendo, então, inserida em S e $Q = \Lambda$ voltando ao início do procedimento O número c de inserções em S (passos de produção) somado com 1 (um), caso a última cópia não seja seguida por uma inserção, é a complexidade da sequência $s_1 s_2 s_3 \dots s_n$.

Segue o exemplo abaixo para a compreensão do procedimento para o cálculo de c a partir da sequência 0010.

Passo 1: O primeiro dígito deve sempre ser inserido $\rightarrow 0 \cdot$

Passo 2: $S = 0, Q = 0, SQ = 00,\ SQ\pi = 0, Q \in v(SQ\pi) \rightarrow 0 \cdot 0$

Passo 3: $S = 0, Q = 01, SQ = 001,\ SQ\pi = 00, Q \notin v(SQ\pi) \rightarrow 0 \cdot 01 \cdot$

Passo 4: $S = 001, Q = 0, SQ = 0010,\ SQ\pi = 001, Q \in v(SQ\pi) \rightarrow 0 \cdot 01 \cdot 0$ (última cópia não seguida de inserção)

Então, $c = 3$ para 0010.

O diagrama para criar o algoritmo é mostrado na Figura 4

O procedimento foi adotado neste trabalho para desenvolver o programa de computador em linguagem *Visual Basic for Applications* (VBA) para medir a complexidade de uma sequência.

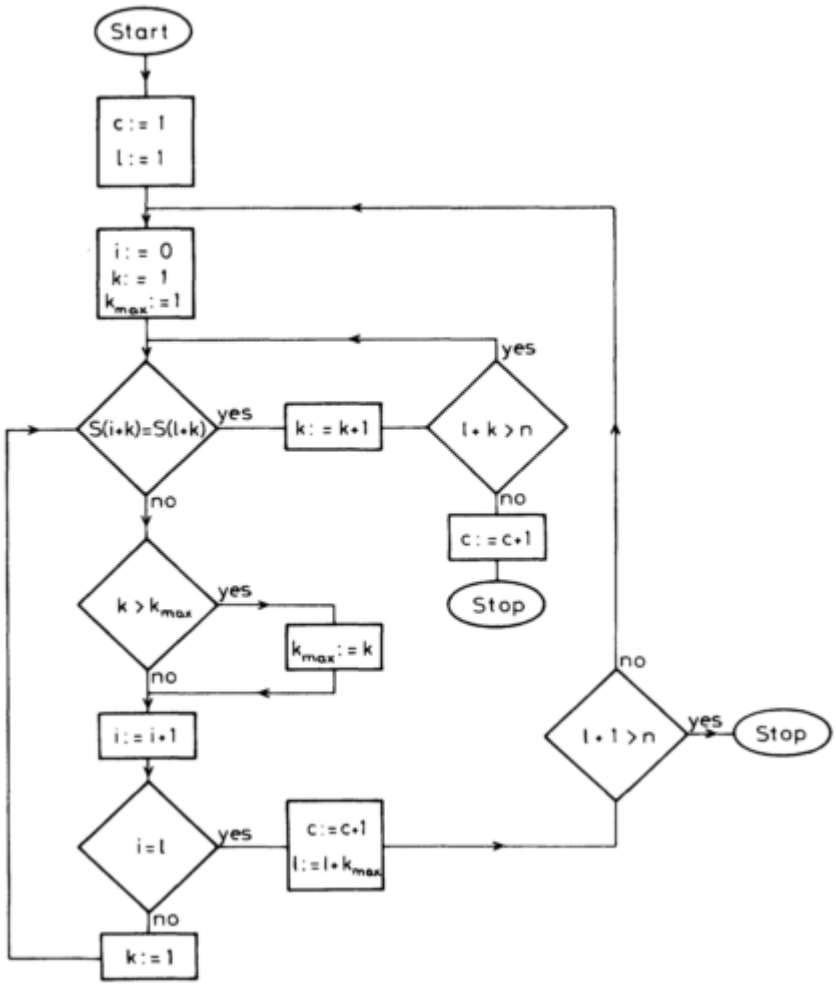

Figura 4: Diagrama para o algoritmo do cálculo da complexidade de uma sequência de comprimento n[14].

2.4 **Aplicação do método aos dados**

Nesta seção foi apresentado o procedimento geral de análise das ações negociadas na BM&F BOVESPA utilizando o grau de complexidade de uma sequência de informações dado por Lempel e Ziv (1976). Para cada análise haverá descrição mais detalhada sobre procedimento no Capítulo 3: Resultados.

[14] Kaspar e Schuster (1987)

A base de dados utilizada para análise foi fornecida pela BM&F BOVESPA contendo todas as negociações do mercado acionário realizadas entre o dia 1 de janeiro de 2007 e o dia 31 de dezembro de 2008, incluindo os extremos do intervalo. A Tabela 2 mostra o número de negociações de cada ação ao longo deste período.

Os dados foram armazenados em arquivos do editor de textos Bloco de Notas® (empresa Microsoft). A disposição de cada negociação foi descrita na linha abaixo:

2007-01-02;BNCA3 ;0000010;000000000049.100000;00000000000000100;11:00:01.412301;2007-01-02;000009;000000000049.180000;00000000000000100;10:46:30.702024;2007-01-02;000026;000000000049.100000;00000000000000100;10:57:19.181241

A Tabela 1 mostra as posições em relação à linha para a extração dos dados, o tamanho da sequência que representa cada informação, além da descrição de cada uma.

Os campos que representam hora são colocados no formato HH:MM:SS.NNNNNN.

O delimitador das colunas de detalhes é representado por ponto e vírgula (;).

Na base de dados foram feitas as análises com os retornos dos preços negociados (Coluna: Preço do Negócio) de cada ação dispostos em sucessão, partindo da negociação mais antiga para a mais nova. O retorno de preços foi dado pela relação:

$$r = \frac{P(t)-P(t-k)}{P(t-k)}, \tag{2.20}$$

onde $P(t)$ é o preço na negociação t e k é o intervalo arbitrado .

Cada retorno formado na série temporal foi convertido a um caractere que pertencesse ao alfabeto {0,1,2} com base no critério de codificação proposto por Shmilovici, Alon-Brimer e Hauser (2003) que utiliza regiões de estabilidade.

Seja x o valor a ser convertido, ρ a região de estabilidade pré determinada e s o valor codificado a partir de x. O critério de codificação é:

$$\begin{aligned} &x > \rho \Rightarrow s = 1 \\ &-\rho < x \leq \rho \Rightarrow s = 2 \\ &x \leq \rho - \Rightarrow s = 0 \end{aligned} \tag{2.21}$$

Após a codificação da série de retornos, todos os caracteres foram concatenados formando, assim, uma sequência.

A partir da sequência formada, partiu-se para o cálculo da complexidade normalizada pela entropia de Shannon (notação: LZ). Para informações sobre a determinação entropia de Shannon, ver item 2.3.1.

O LZ assume o valor de, pelo menos 1, para uma sequência de de Bruijn (este é o parâmetro comparativo de aleatoriedade).

Tabela 1: Posições e tamanho dos dados na linha de registro

Coluna	Posição Inicial	Tamanho	Descrição
Data Sessão	1	10	Data de sessão
Papel	12	12	Código do papel
Nr.Negócio	25	7	Número do negócio
Preço	33	19	Preço do negócio
Quantidade	53	18	Quantidade negociada
Hora	72	15	Horário da negociação
Data Oferta Compra	88	10	Data da oferta de compra
Seq.Oferta Compra	99	6	Número sequencial da oferta de compra
Preço Of.Compra	106	19	Preço da Oferta
Qtd.Negociada Of.Compra	126	18	Quantidade Negociada
Hora Prior. Of.Compra	145	15	Hora de registro da oferta no sistema
Data Oferta Venda	161	10	Data da oferta de venda
Seq.Oferta Venda	172	6	Número sequencial da oferta de venda
Preço Of.Venda	179	19	Preço da Oferta
Qtd.Negociada Of.Venda	199	18	Quantidade Negociada
Hora Prior. Of.Venda	218	15	Hora de registro da oferta no sistema

Fonte: BM&F BOVESPA

Baseado no trabalho de Giglio (2008) foi possível perceber que não era necessário utilizar todos os elementos da sequência de caracteres para calcular o LZ, pois há probabilidade não nula de se obter valores altos de LZ para uma sequência não complexa devido à flutuação aleatória da medida (LEMPEL e ZIV, 1976). Aquele autor fez menção ao uso do método das janelas deslizantes.

O método consiste em utilizar subsequências de comprimento fixo chamadas de janelas, que fazem parte da sequência original, para o cálculo do LZ.

Para a obtenção desta medida, deve-se seguir o procedimento abaixo:

a) Define-se o comprimento da janela (j) e o comprimento do salto (s), que consiste determinar em que posição iniciará a nova janela. Exemplo: Para uma janela de tamanho igual a 4 e salto de tamanho igual a 3, a sequência 0123456789 terá as seguintes subsequências: 0123, 3456, 6789;
b) Extrai-se o LZ de cada janela; e
c) Calcula-se a média dos LZ de todas as janelas. Esta estatística é representativa do LZ da sequência toda.

Para o cálculo da entropia de Shannon (h), foi feita a suposição, neste trabalho, de que cada janela teria sido gerada a partir de uma fonte ergódica. Com isso, as probabilidades de ocorrência dos caracteres "0", "1" e "2" têm como estimativas as freqüências relativas em que aparecem em cada janela. Existem outras formas de calcular tais probabilidades, como o método proposto por Bandt e Pompe (2002).

Tal método consiste em, a partir de uma série temporal finita de tamanho T, fracamente estacionária, e com um determinado número n de dimensões incorporadas, determinar o número de permutações do padrão π_i (total de permutações: $n!$) que ocorrem ao longo de sub-séries de elementos consecutivos de tamanho n da série original. A freqüência relativa do padrão de permutação π_i é dada por:

$$p(\pi_i) = \frac{\#\{t | t \leq T - n, (x_{t+1}, x_{t+2}, \dots, x_{t+n}) \text{é do tipo } \pi_i\}}{T - n + 1} \quad (2.22)$$

Com isso, a entropia de Shannon assume a forma $H(n) - \sum_{i=1}^{n!} p(\pi_i) \log p(\pi_i)$.

Giglio (2008) propôs uma medida de eficiência relativa de mercado verificando a razão entre o número de janelas para o qual $LZ > 1$ e o número total de janelas. Esta medida, segundo o autor, seria útil no sentido de verificar a hipótese de mercados eficientes na sua classificação fraca, de acordo com Jensen (1978) e de comparar níveis de eficiência entre ativos distintos.

As medidas de correlação entre retornos defasados foram feitas utilizando como estimativa a correlação de Pearson, que é dada por:

$$\rho = \frac{\sum_{i=1}^{T-\tau}\left(X_i - \frac{\sum_{i=1}^{T-\tau} X_i}{T-\tau}\right)\left(X_{i+\tau} - \frac{\sum_{i=1}^{T-\tau} X_{i+\tau}}{T-\tau}\right)}{\sqrt{\sum_{i=1}^{T-\tau}\left(X_i - \frac{\sum_{i=1}^{T-\tau} X_i}{T-\tau}\right)^2 \sum_{i=1}^{T-\tau}\left(X_{i+\tau} - \frac{\sum_{i=1}^{T-\tau} X_{i+\tau}}{T-\tau}\right)^2}} \quad (2.23)$$

onde T é o número de termos da série temporal, τ é a defasagem entre os termos que são correlacionados e X é o valor que assume a variável aleatória estocástica.

Tal medida foi utilizada na análise da perda de memória nos processos estocásticos envolvendo as séries de retorno mencionadas.

Os detalhes sobre a aplicação das correlações e do LZ estão no capítulo 3: Resultados.

3 **RESULTADOS E DISCUSSÃO**

Neste capítulo foram apresentados os experimentos realizados com os retornos de ações negociadas na BM&F BOVESPA aplicando as ferramentas enunciadas no capítulo anterior.

3.1 **Considerações gerais**

Para que fosse feita uma observação mais rigorosa a respeito da evolução do LZ e também sobre a análise de correlações foram utilizados dados de alta freqüência (negócio a negócio) relativos às 15 ações apresentadas na Tabela 2, todas negociadas na BM&F BOVESPA. Tais ações foram escolhidas por serem componentes do índice Bovespa, dessa forma estando entre os ativos com maior volume de negociação em bolsa.

Tabela 2: Ações negociadas na BOVESPA utilizadas no estudo da evolução do LZ.

Código da ação	Empresa e característica da ação	Número de Negociações
CGAS5	COMGÁS – PNA	107.341
CLSC6	CELESC – PNB	121.634
TLPP4	TELESP –PN	164.103
TMAR5	TELEMAR N L – PNA	182.223
BRTP3	BRASIL T PAR – ON	198.345
UGPA4	ULTRAPAR – PN	200.378
LIGT3	LIGHT – ON	209.541
TCSL3	TIM PART – ON	251.784
TRPL4	CTEEP – ON	254.948
BNCA3	NOSSA CAIXA – ON	274.492
TNLP3	TELEMAR – ON	280.735
USIM3	USIMINAS – ON	285.773
SBSP3	SABESP – ON	348.698
SDIA4	SADIA – PN	950.501
ELPL6	ELETROPAULO – PNB	451.312

Fonte: BM&F BOVESPA

3.2 **Análise dos resultados no caso de intervalos de tempo igual a uma negociação**

3.2.1 Evolução do LZ médio em relação à região de estabilidade

Para os estudos relativos a esta seção foram utilizadas todas as ações presentes na Tabela 2 exceto a ação ELPL6 (Eletropaulo). A Tabela 3 evidencia os tamanhos de janela, de salto e valores de região de estabilidade aplicados na análise da eficiência relativa de mercado daqueles ativos.

Tabela 3: Janelas, Saltos e regiões de estabilidades para análise de dados negócio a negócio.

Janela	Salto	Região de estabilidade (%)
30000	10000	0,001
10000	7500	0,010
5000	5000	0,020
	2500	0,030
	1000	0,040
		0,050
		0,100
		0,150
		0,200
		0,250
		0,300
		0,350
		0,400
		0,500

As figuras 5 a 46 mostram o comportamento do LZ médio para as ações da Tabela 2 em relação à região de estabilidade fixando, para cada diagrama, o tamanho das janelas. Nestas figuras “J” significa janela e “S” significa salto.

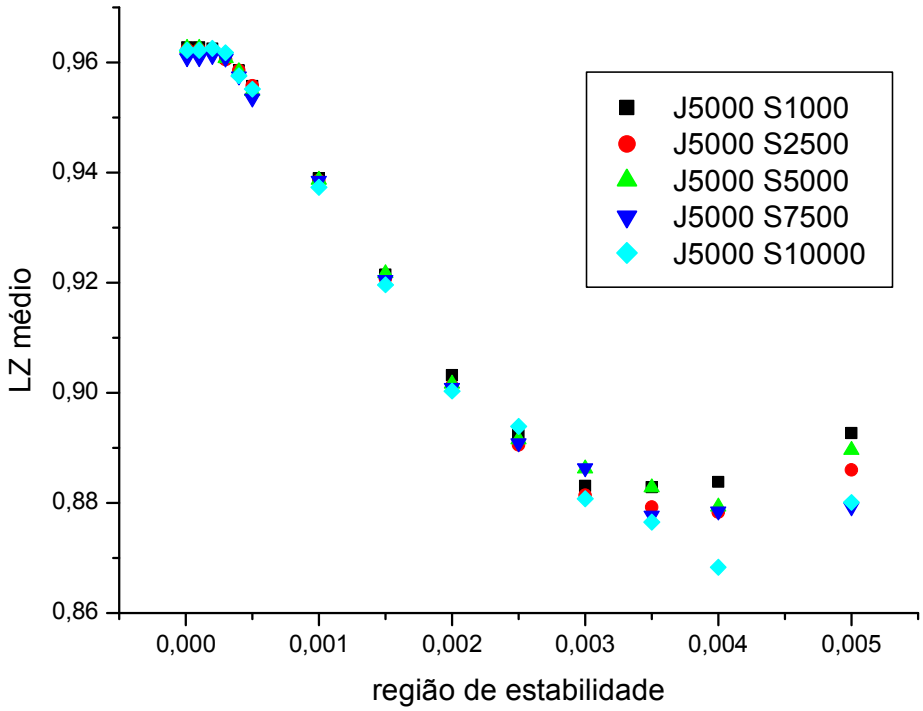

Figura 5: LZ médio *versus* RE BNCA3 JANELA 5000.

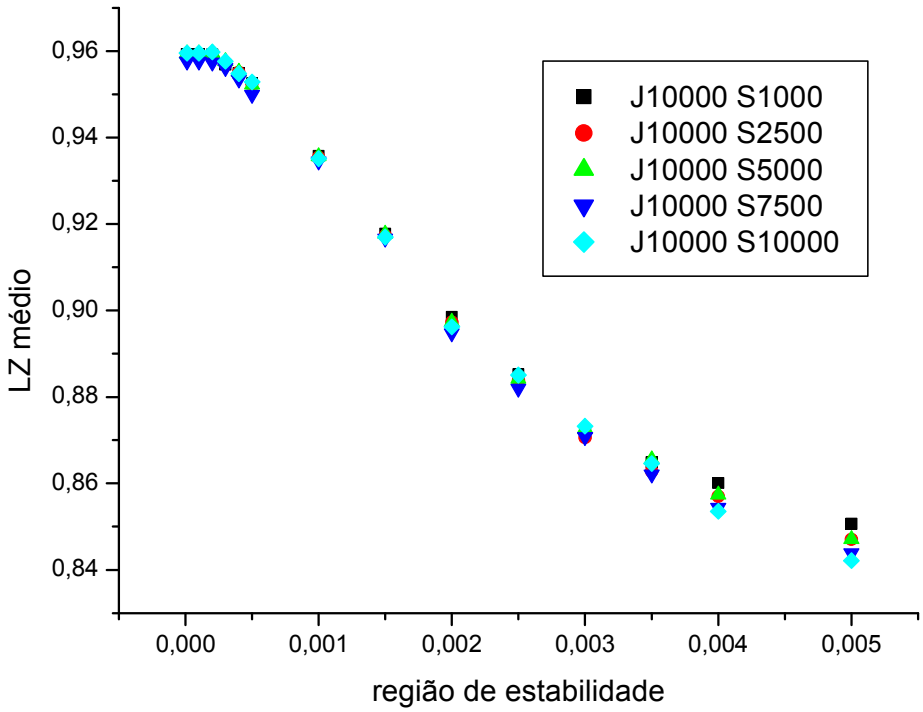

Figura 6:LZ médio *versus* RE BNCA 3 JANELA 10000.

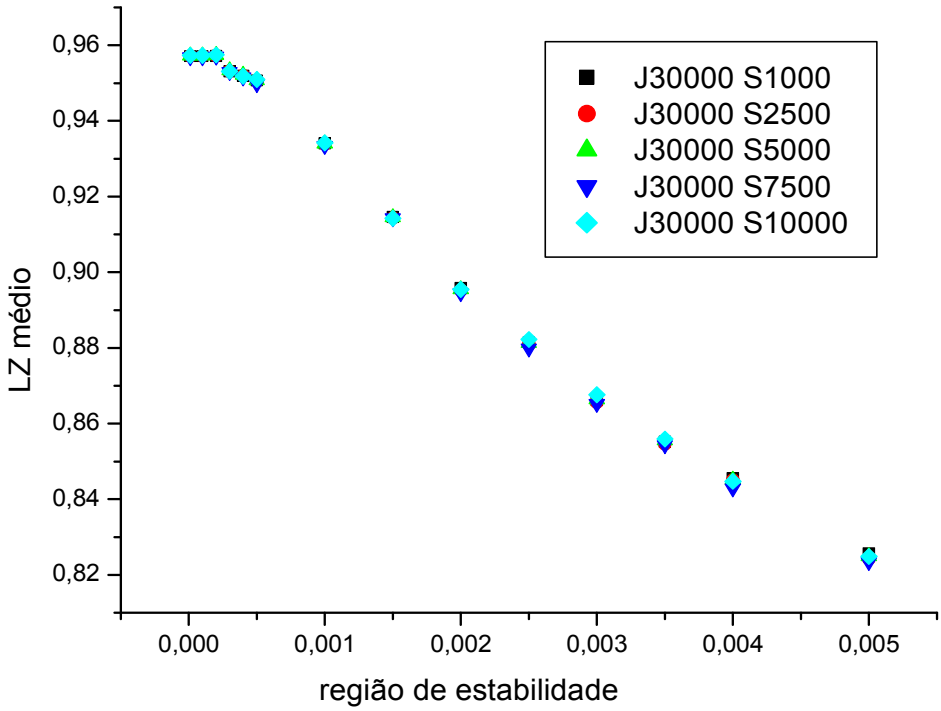

Figura 7: LZ médio *versus* RE BNCA3 JANELA 30000.

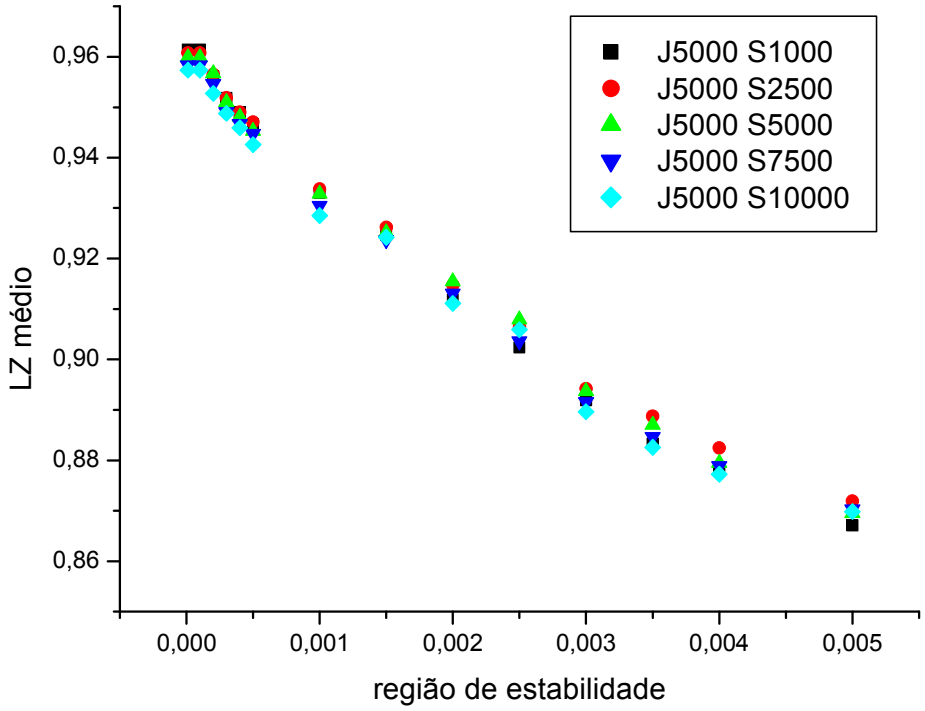

Figura 8: LZ médio *versus* RE BRTP3 JANELA 5000.

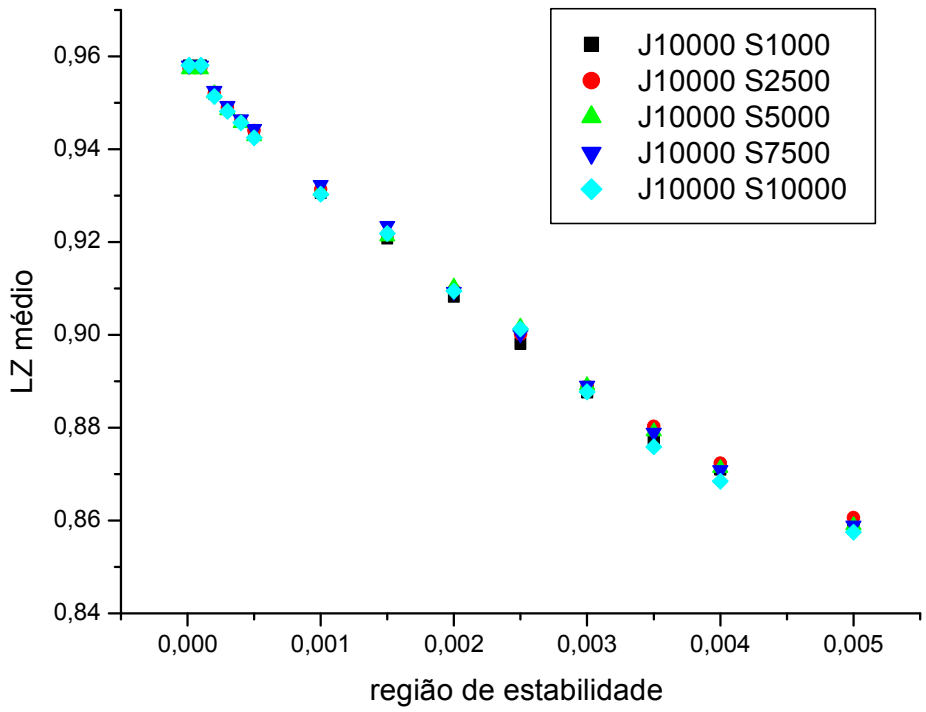

Figura 9: LZ médio *versus* RE BRTP3 JANELA 10000.

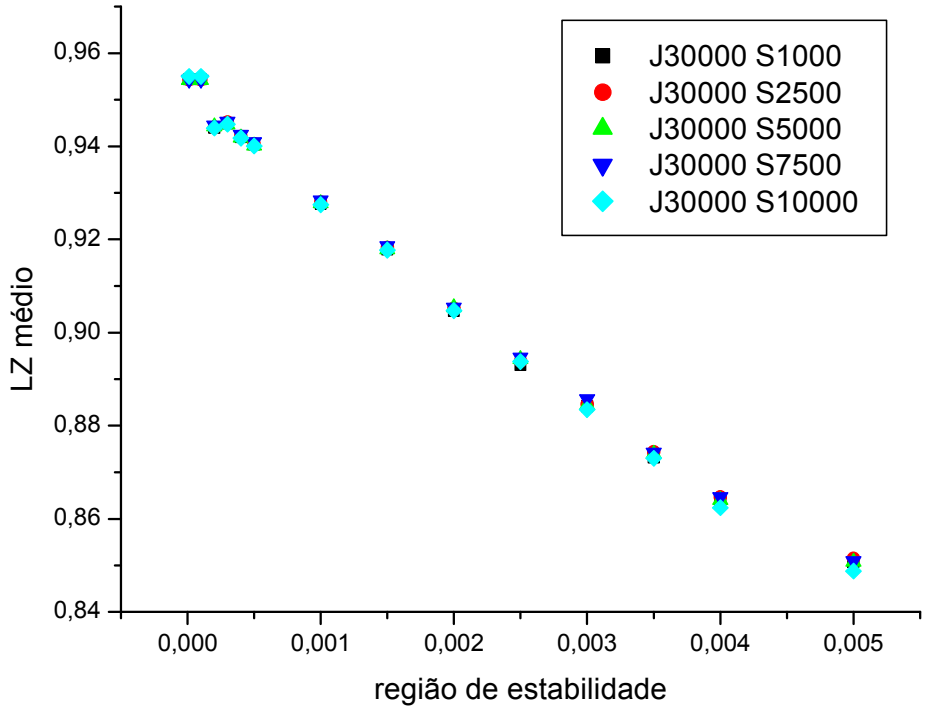

Figura 10: LZ médio *versus* RE BRTP3 JANELA 30000.

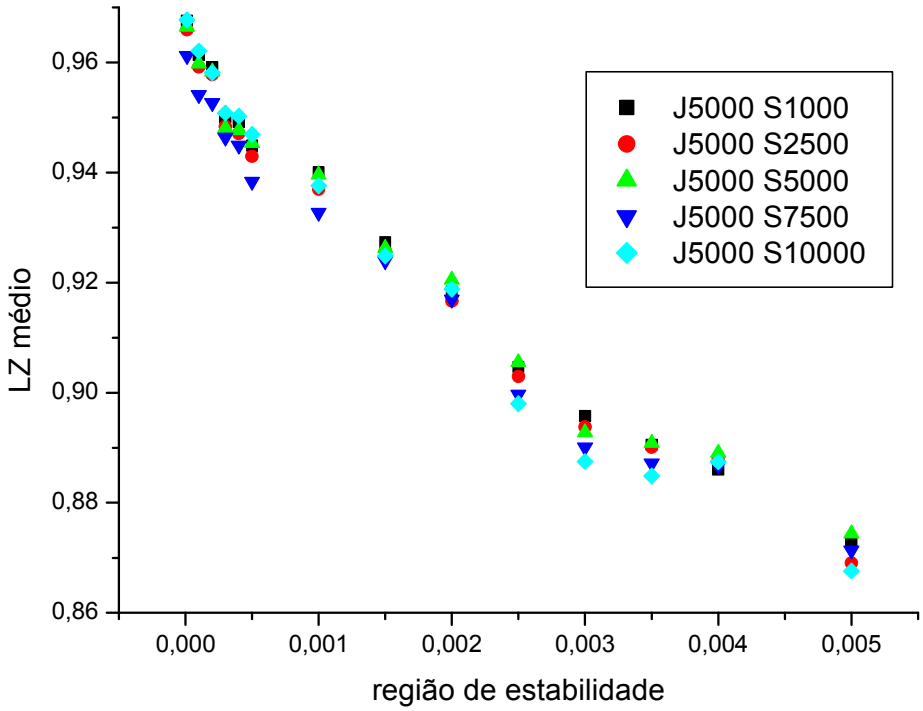

Figura 11:LZ médio *versus* RE CGAS5 JANELA 5000.

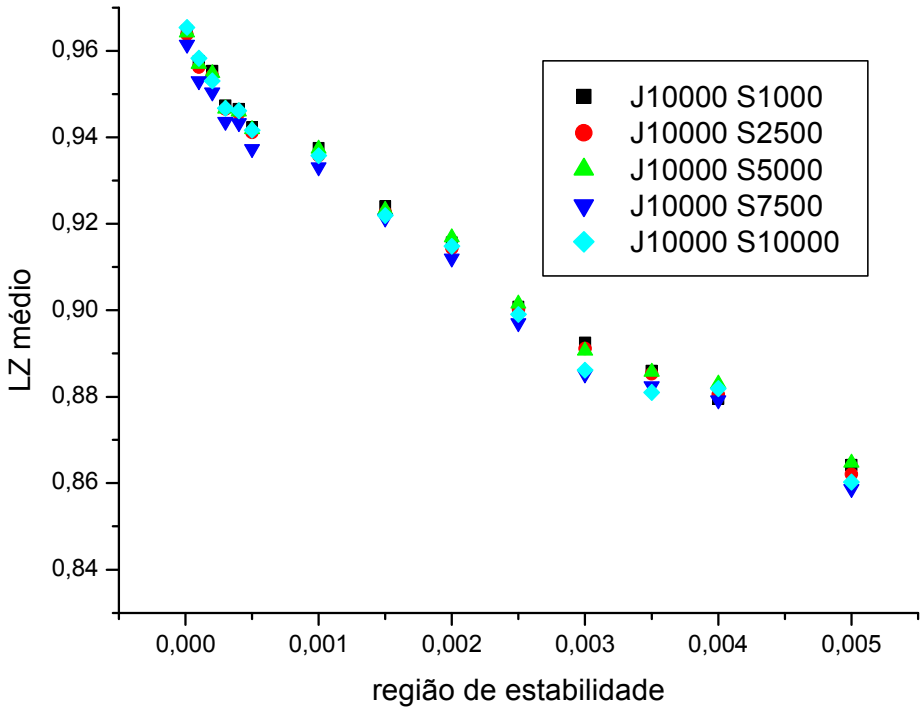

Figura 12:LZ médio *versus* RE CGAS5 JANELA 10000.

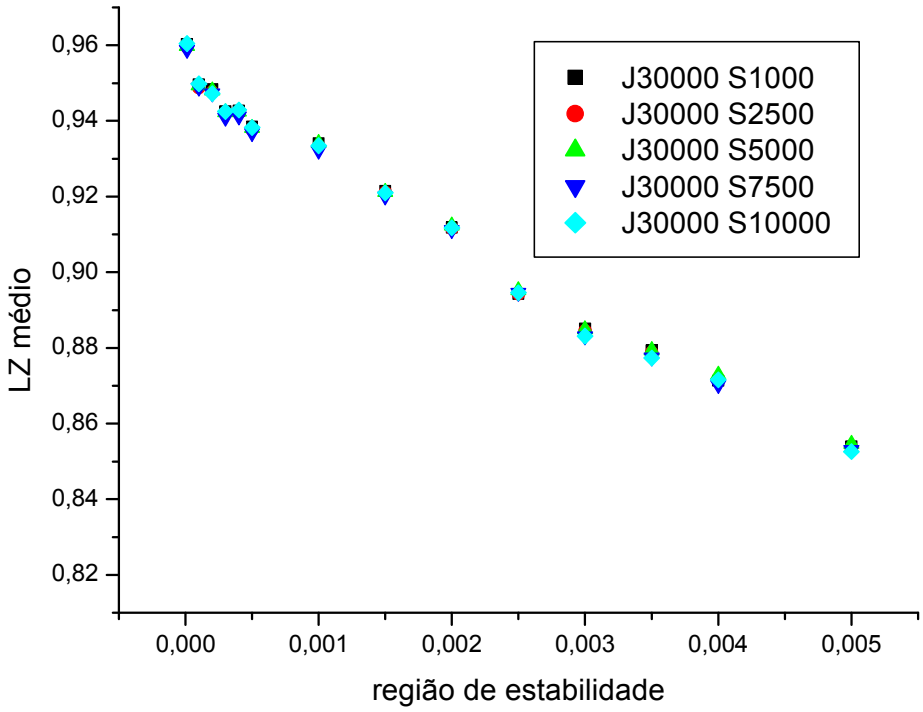

Figura 13: LZ médio *versus* RE CGAS5 JANELA 30000.

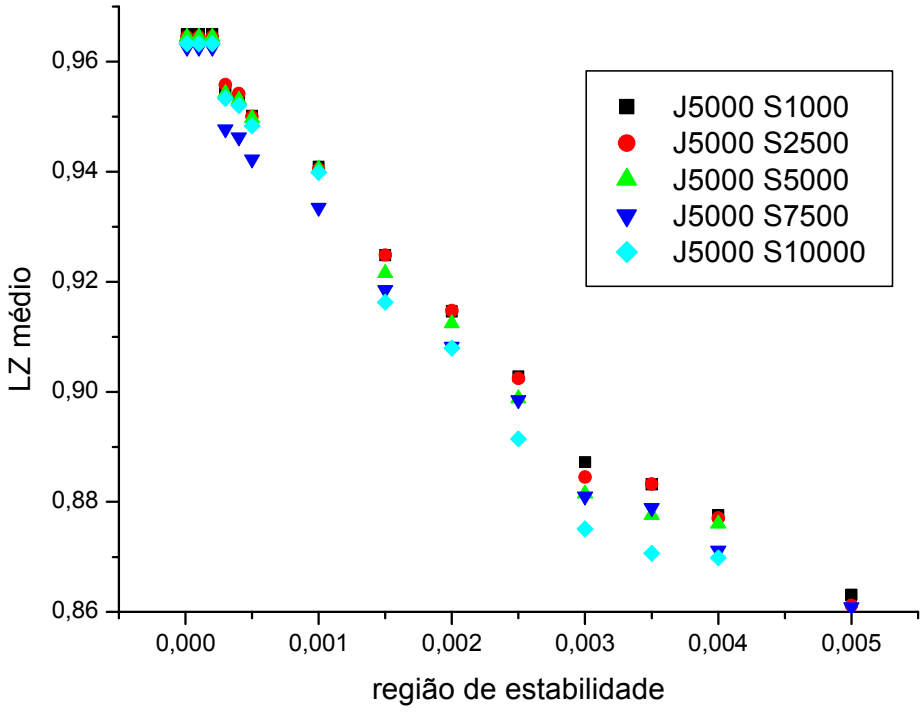

Figura 14: LZ médio *versus* RE CLSC6 JANELA5000.

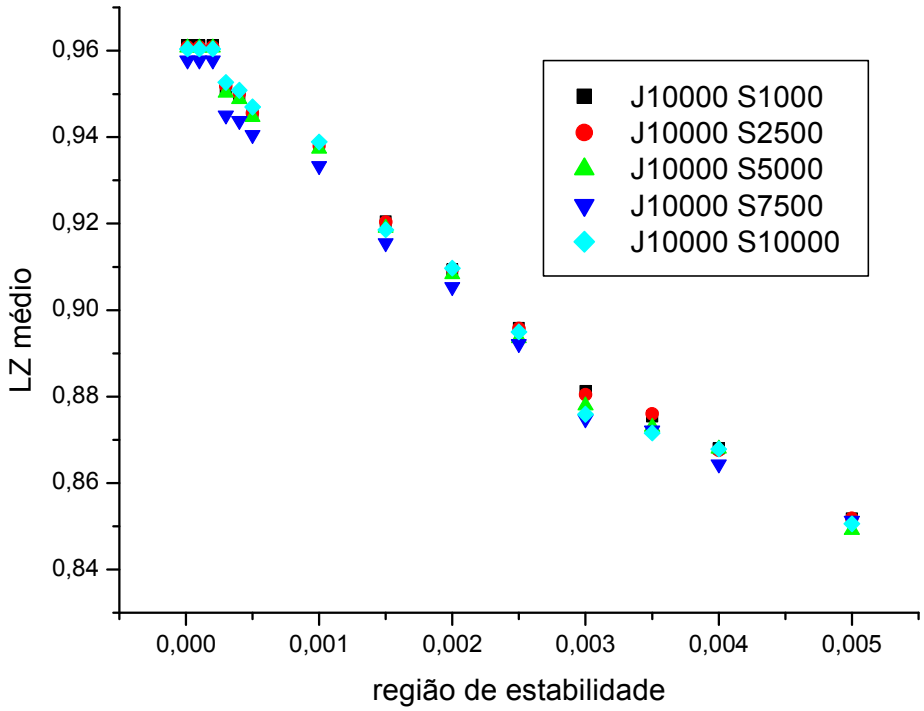

Figura 15: LZ médio *versus* RE CLSC6 JANELA 10000.

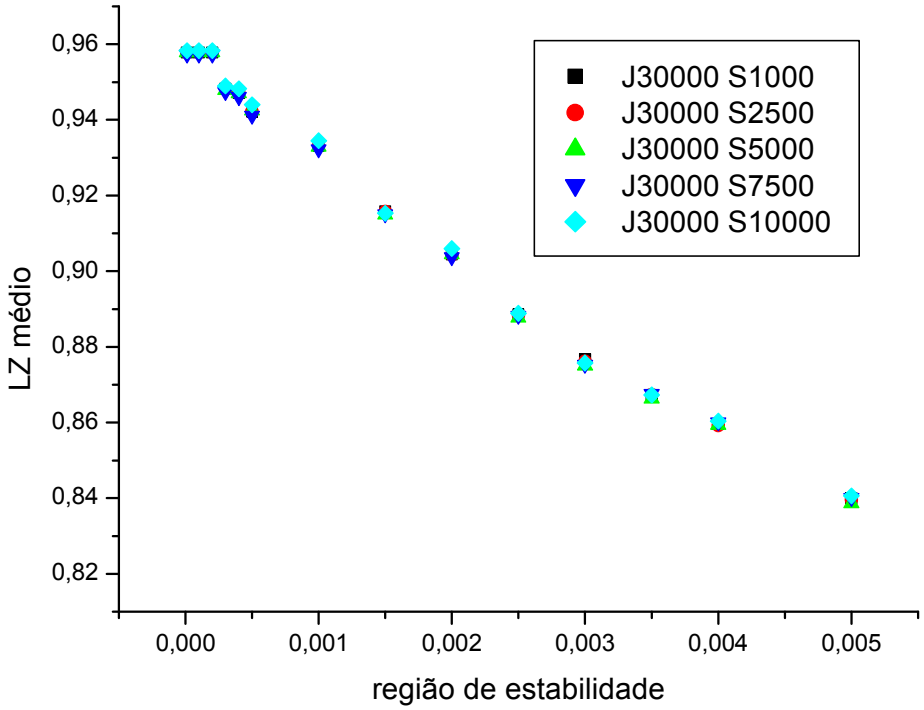

Figura 16: LZ médio *versus* RE CLSC6 JANELA 30000.

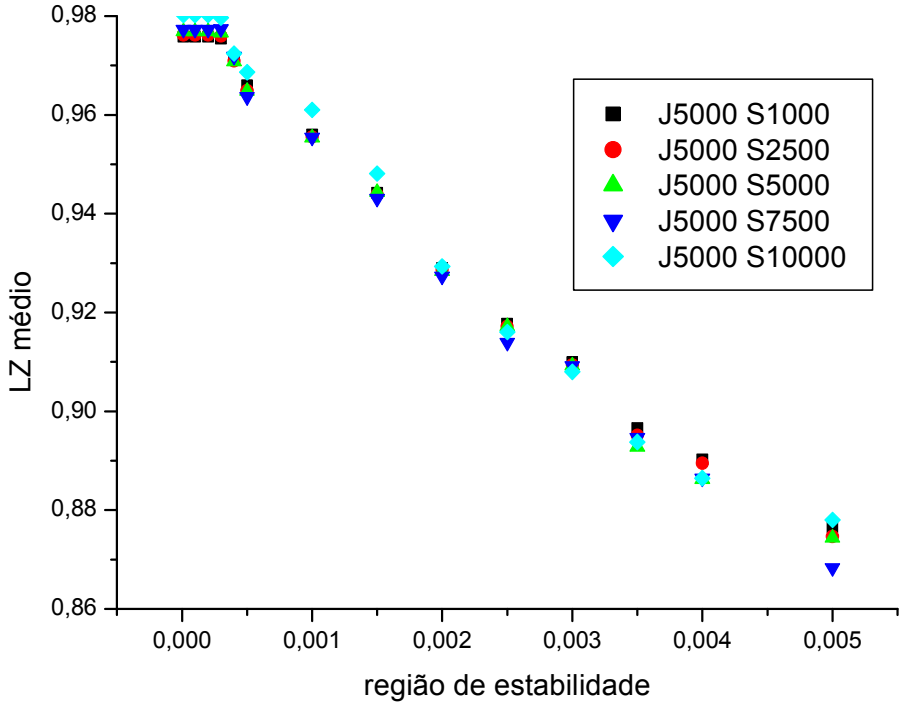

Figura 17: LZ médio *versus* RE LIGT3 JANELA 5000.

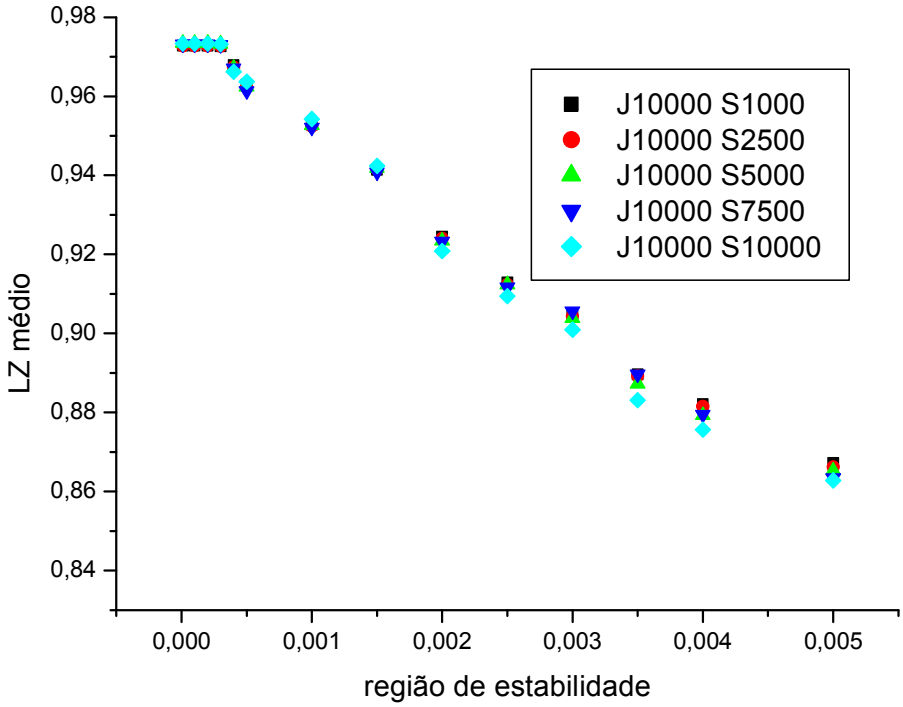

Figura 18: LZ médio *versus* RE LIGT3 JANELA 10000.

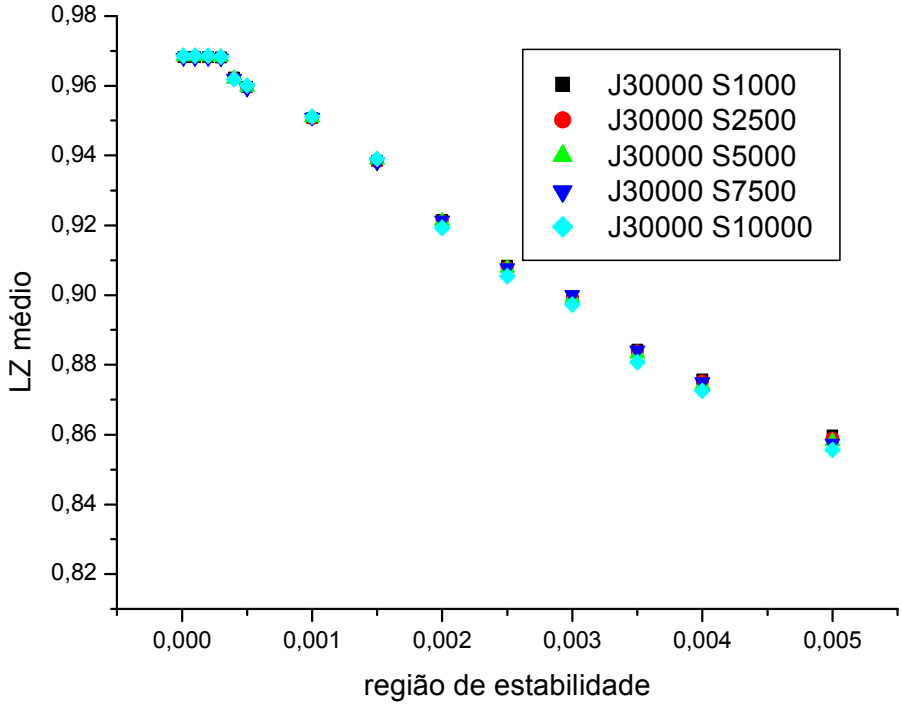

Figura 19: LZ médio *versus* RE LIGT3 JANELA 30000.

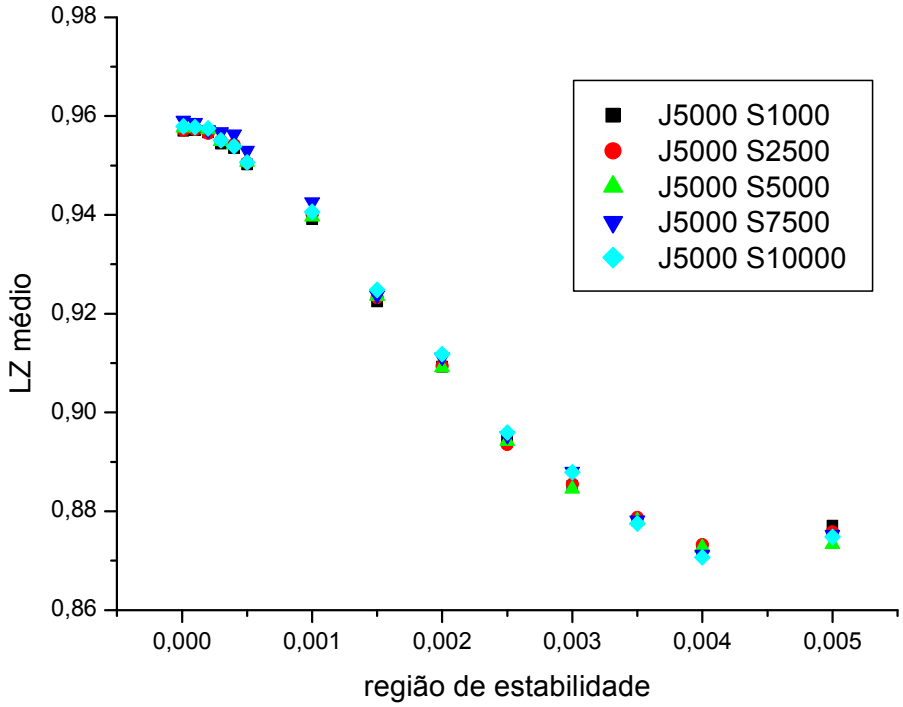

Figura 20: LZ médio *versus* RE SBSP3 JANELA 5000.

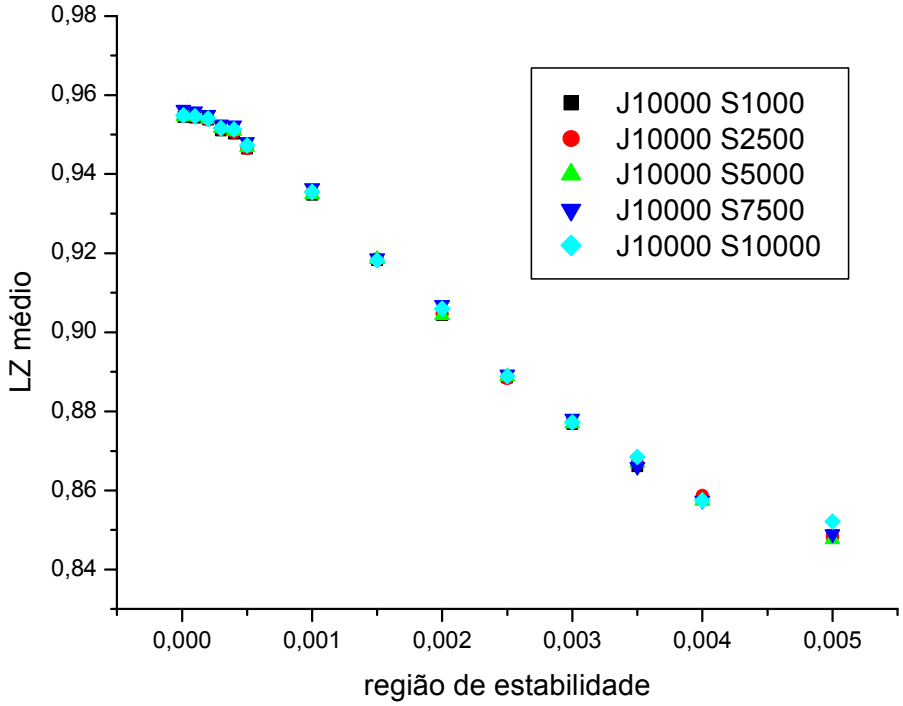

Figura 21: LZ médio *versus* RE SBSP3 JANELA 10000.

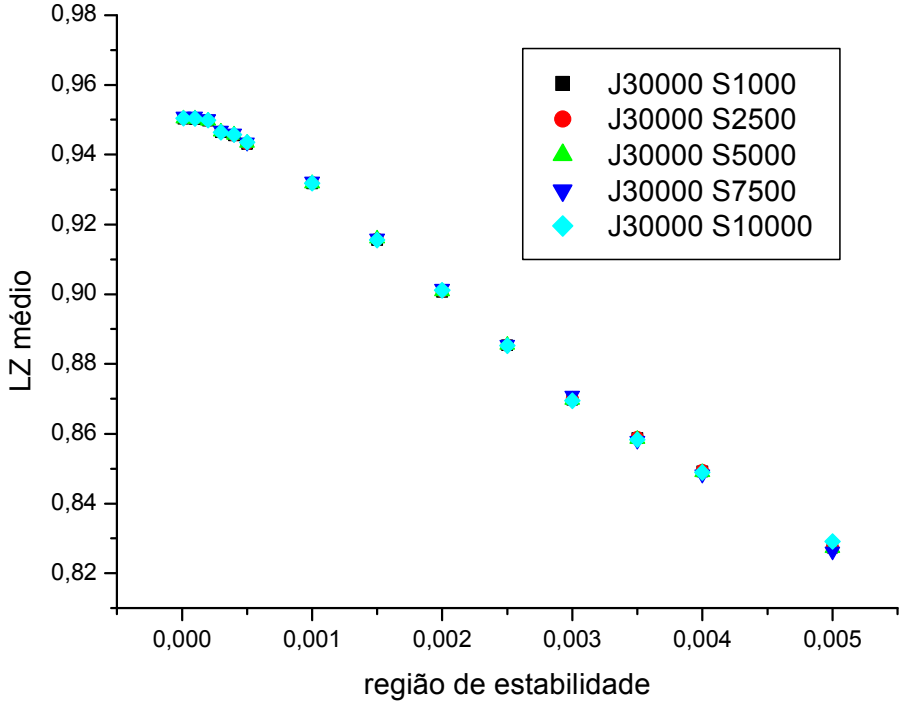

Figura 22: LZ médio *versus* RE SBSP3 JANELA 30000.

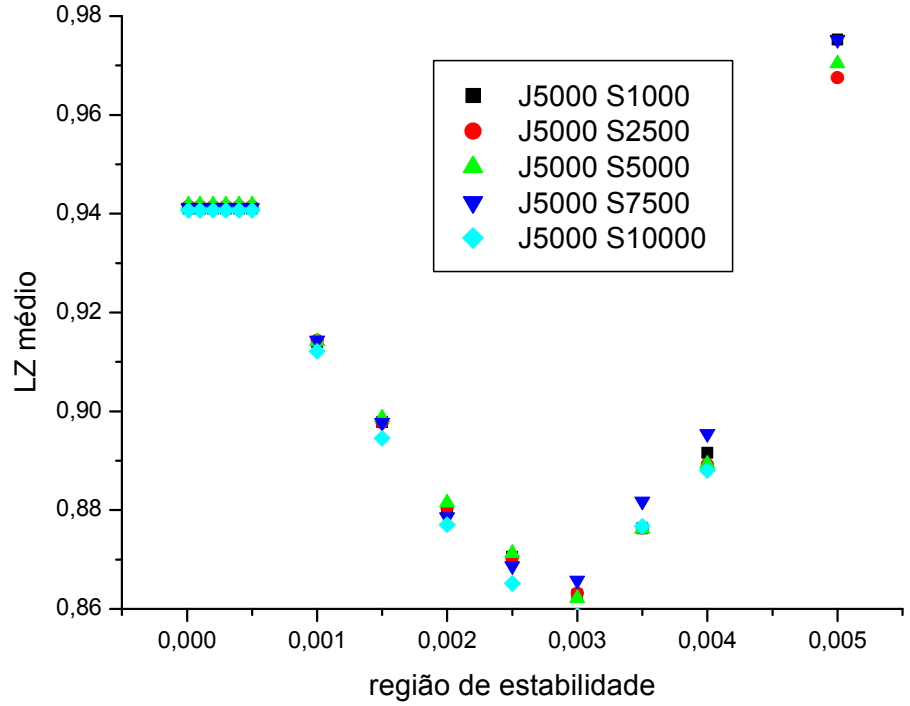

Figura 23: LZ médio *versus* RE SDIA4 JANELA 5000.

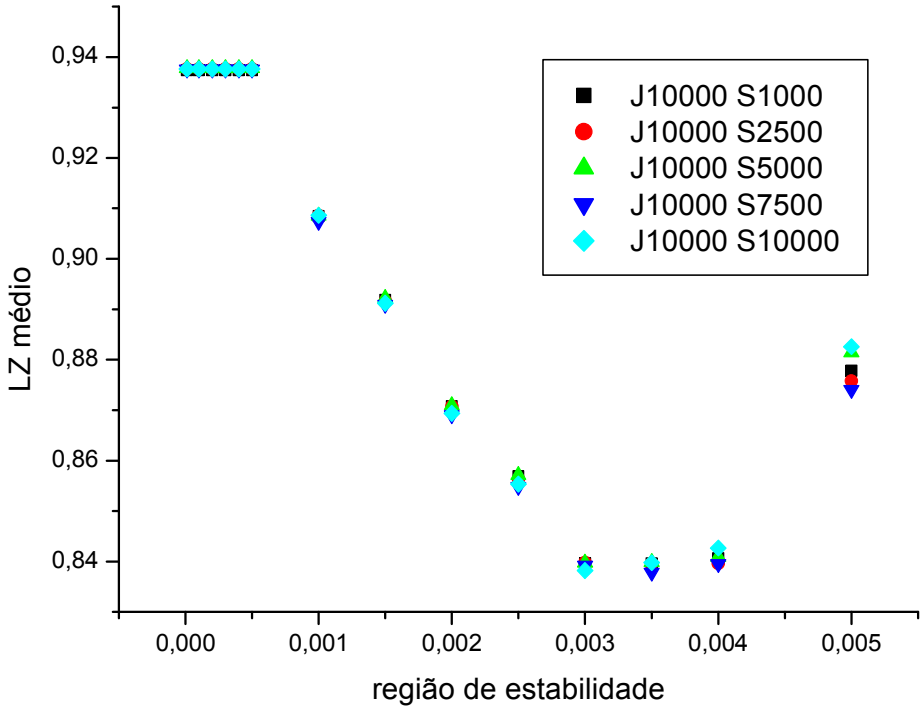

Figura 24: LZ médio *versus* RE SDIA4 JANELA 10000.

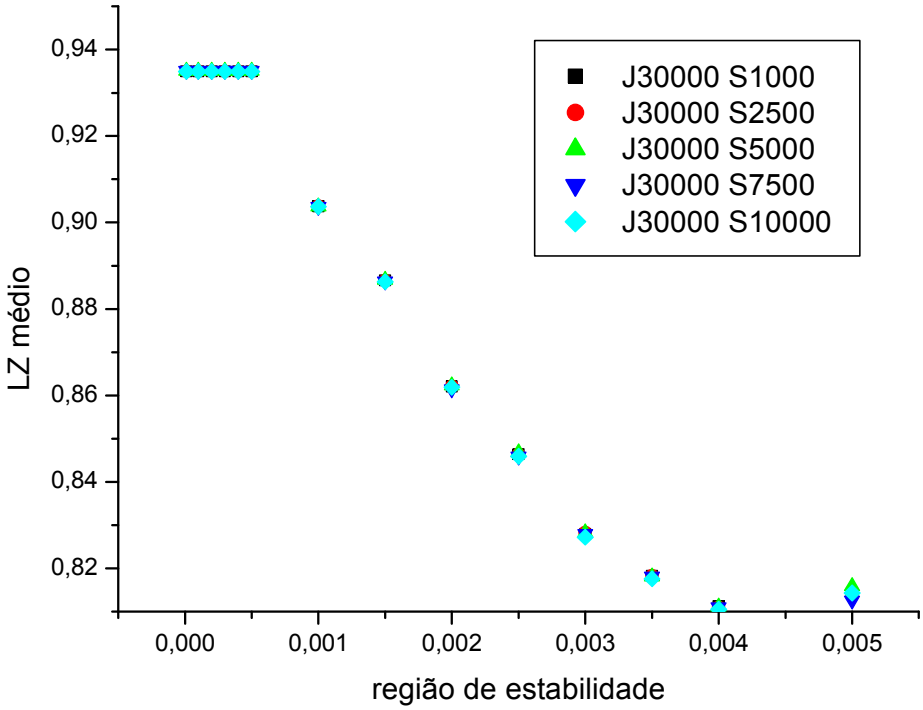

Figura 25: LZ médio *versus* RE SDIA4 JANELA 30000.

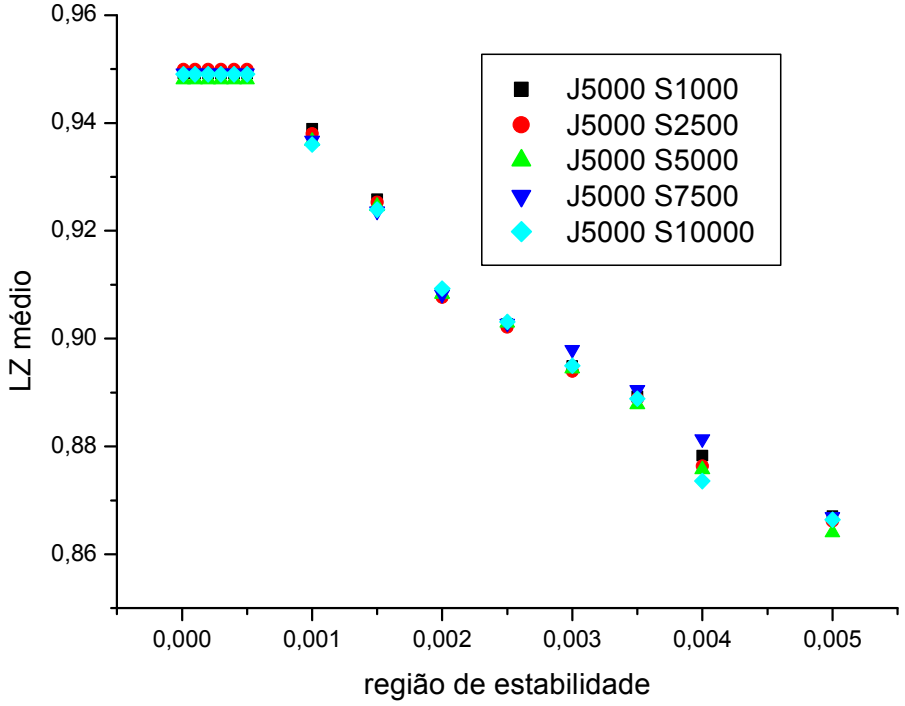

Figura 26: LZ médio *versus* RE TCSL3 JANELA 5000.

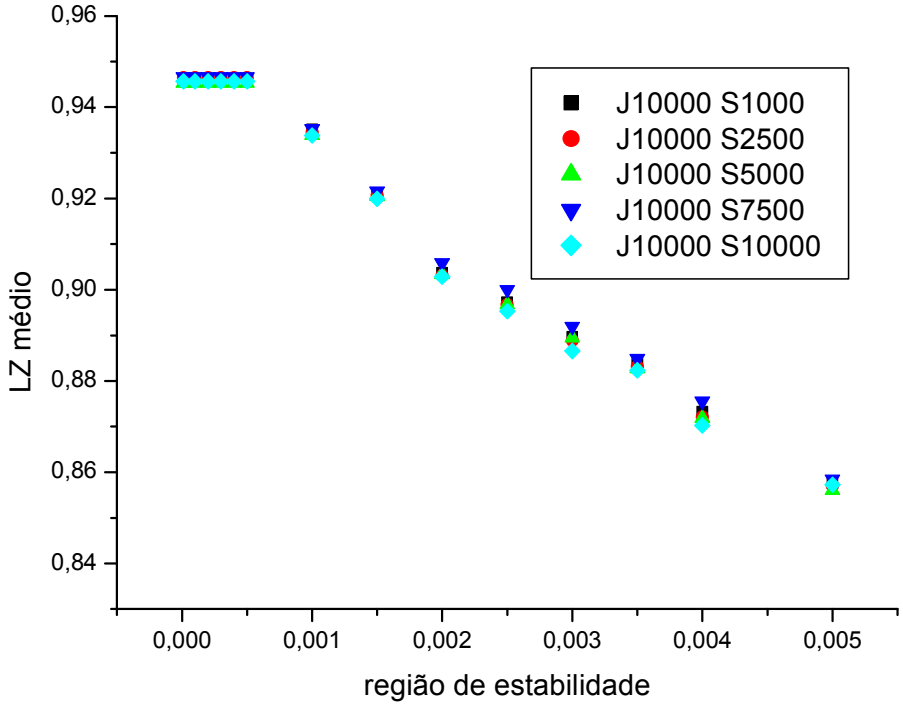

Figura 27: LZ médio *versus* RE TCSL3 JANELA 10000.

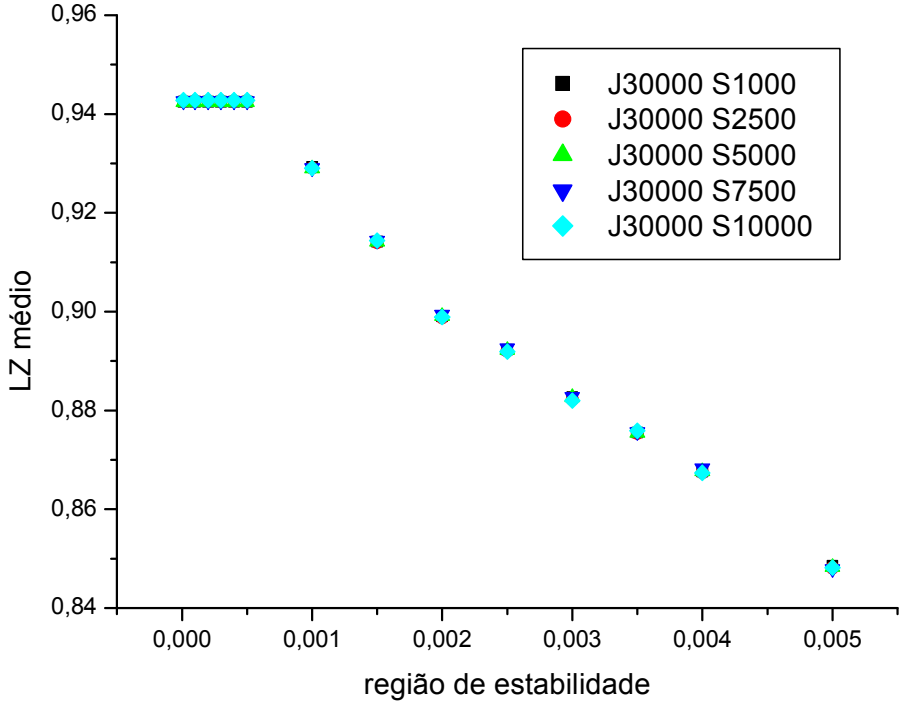

Figura 28: LZ médio *versus* RE TCSL3 JANELA 30000.

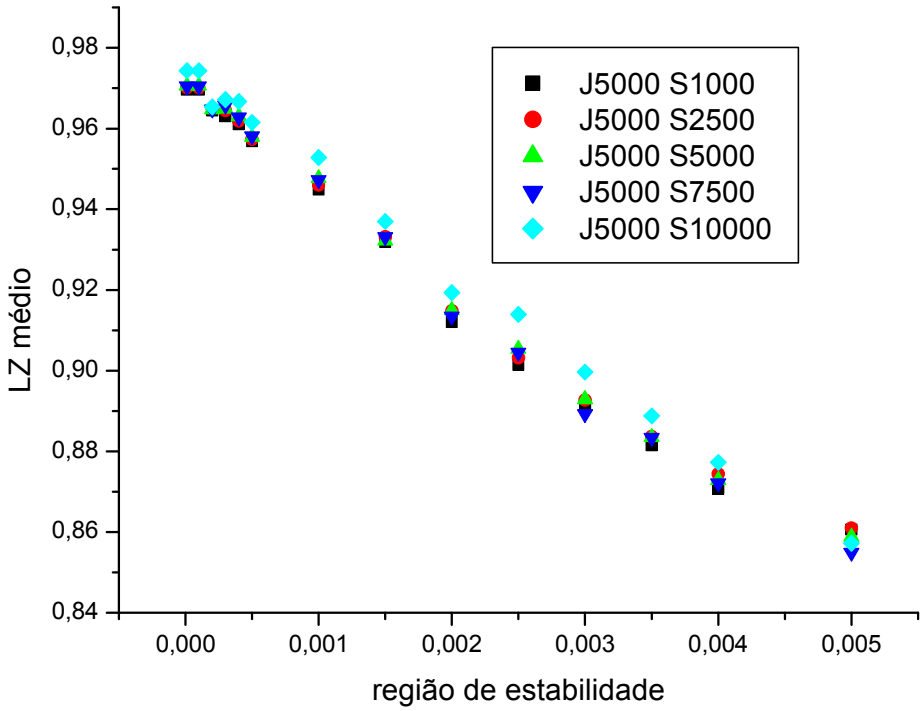

Figura 29: LZ médio *versus* RE TLPP4 JANELA 5000.

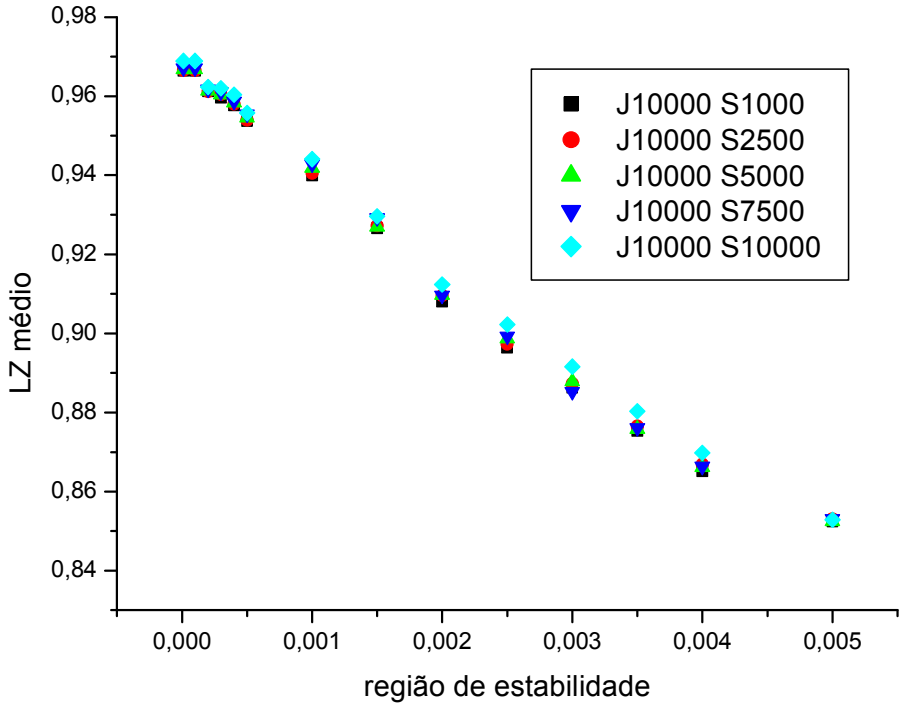

Figura 30: LZ médio *versus* RE TLPP4 JANELA 10000.

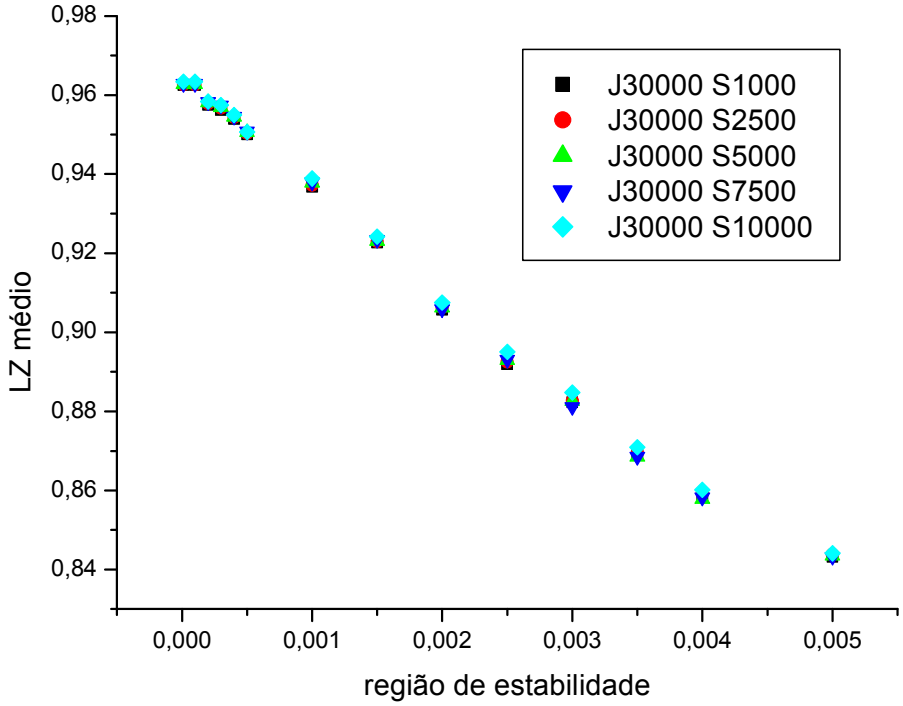

Figura 31: LZ médio *versus* RE TLPP4 JANELA 30000.

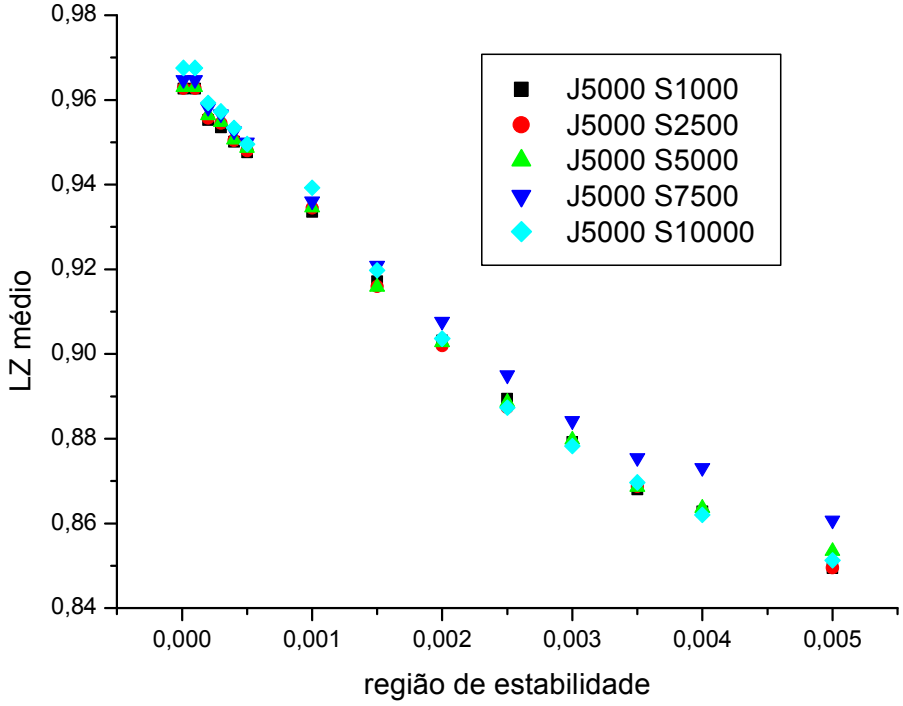

Figura 32: LZ médio *versus* RE TMAR5 JANELA 5000.

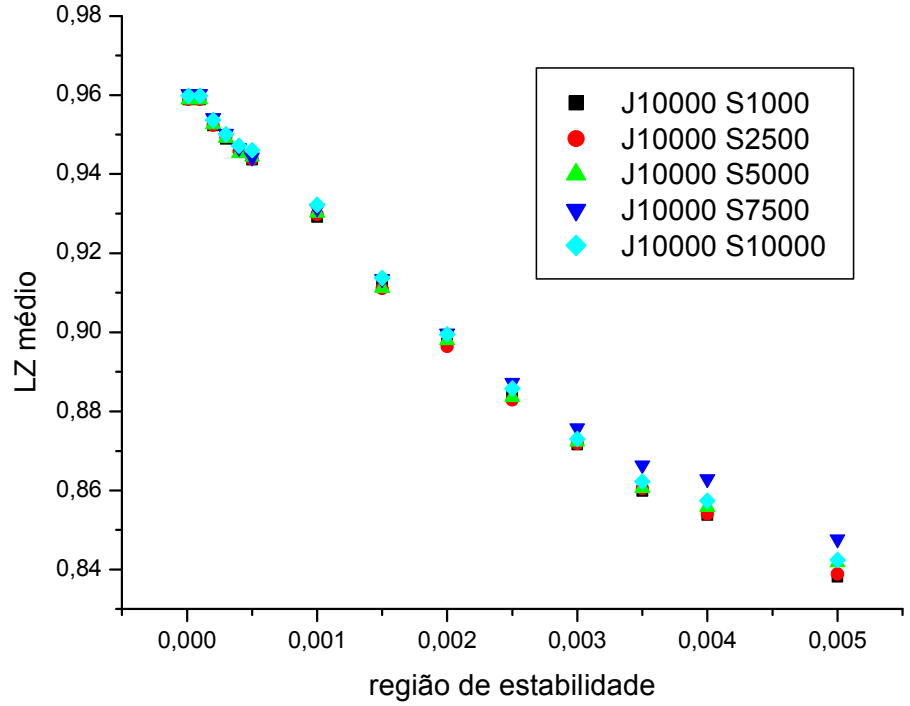

Figura 33: LZ médio *versus* RE TMAR5 JANELA 10000.

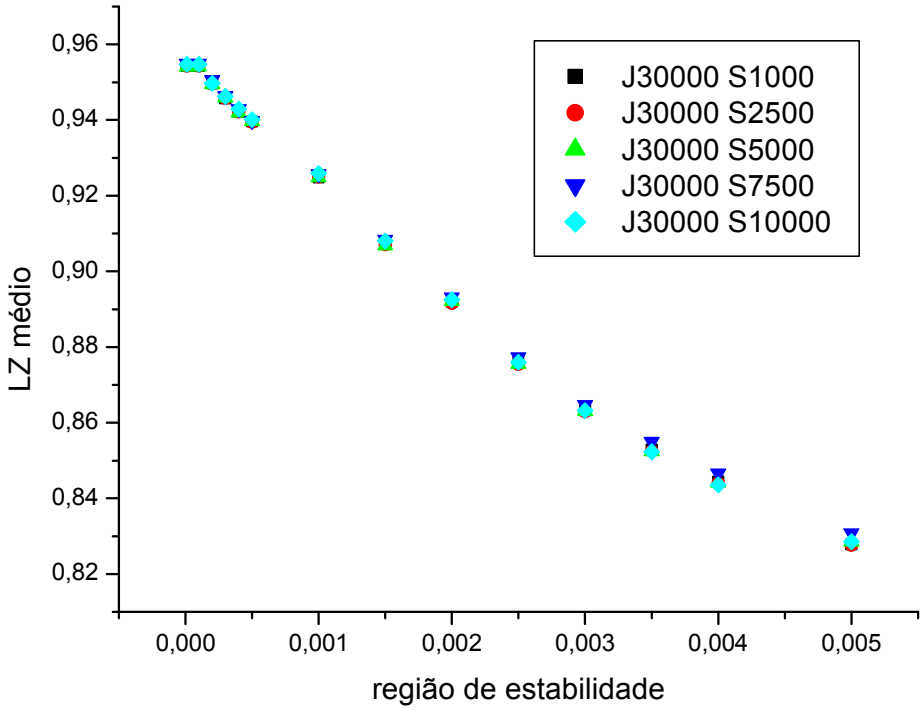

Figura 34: LZ médio *versus* RE TMAR5 JANELA 30000.

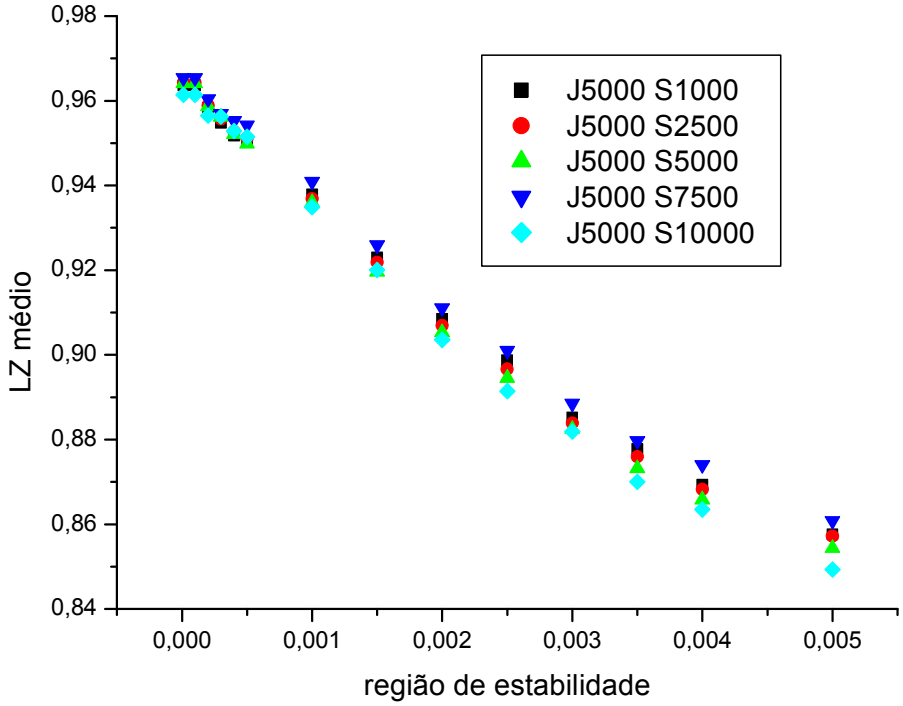

Figura 35: LZ médio *versus* RE TNLP3 JANELA 5000.

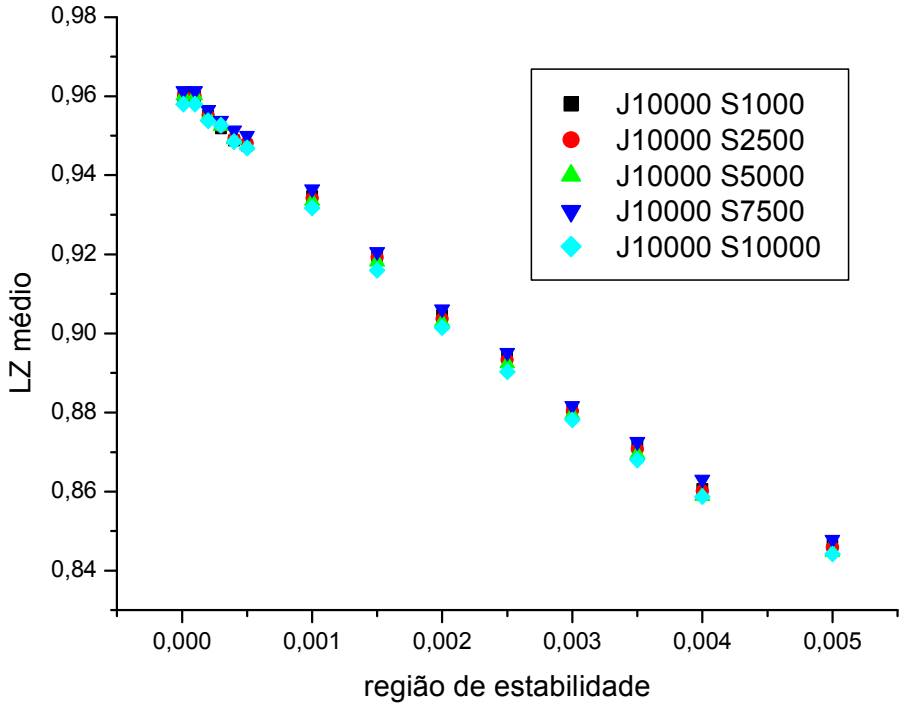

Figura 36: LZ médio *versus* RE TNLP3 JANELA 10000.

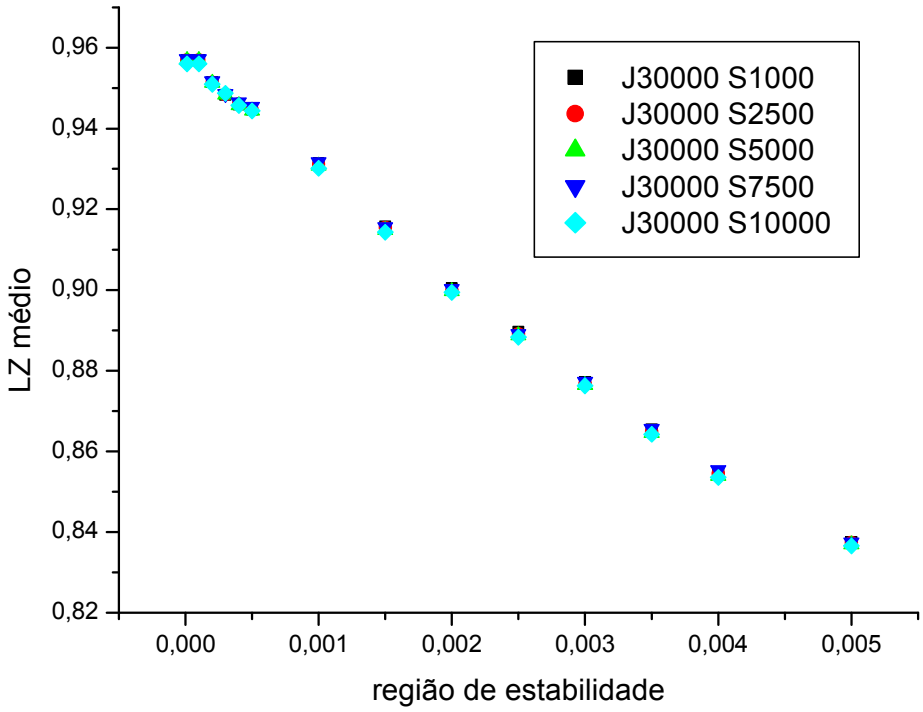

Figura 37: LZ médio *versus* RE TNLP3 JANELA 30000.

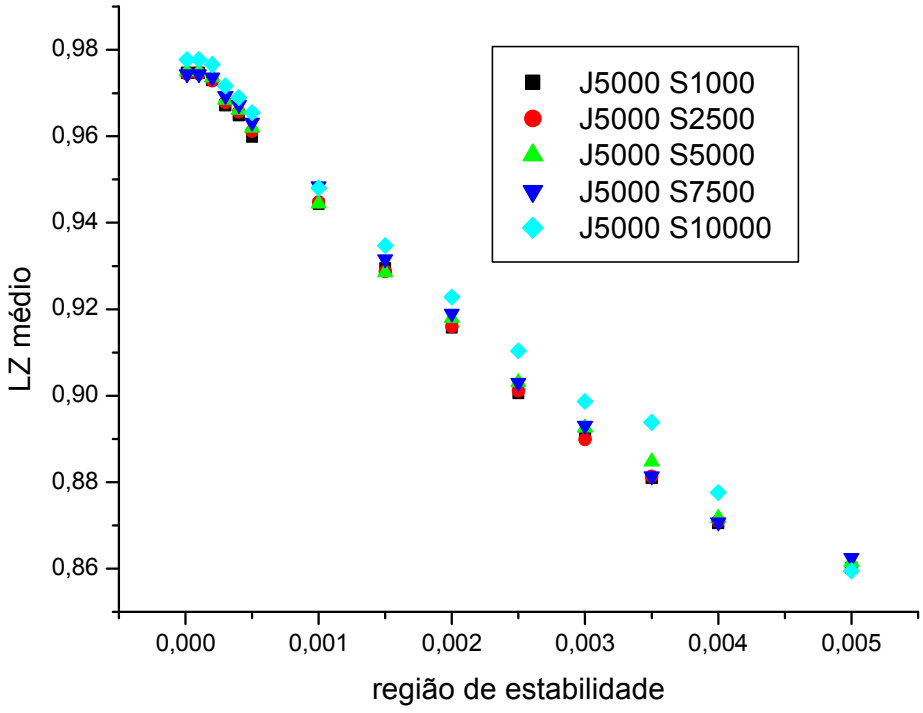

Figura 38: LZ médio *versus* RE TRPL4 JANELA 5000.

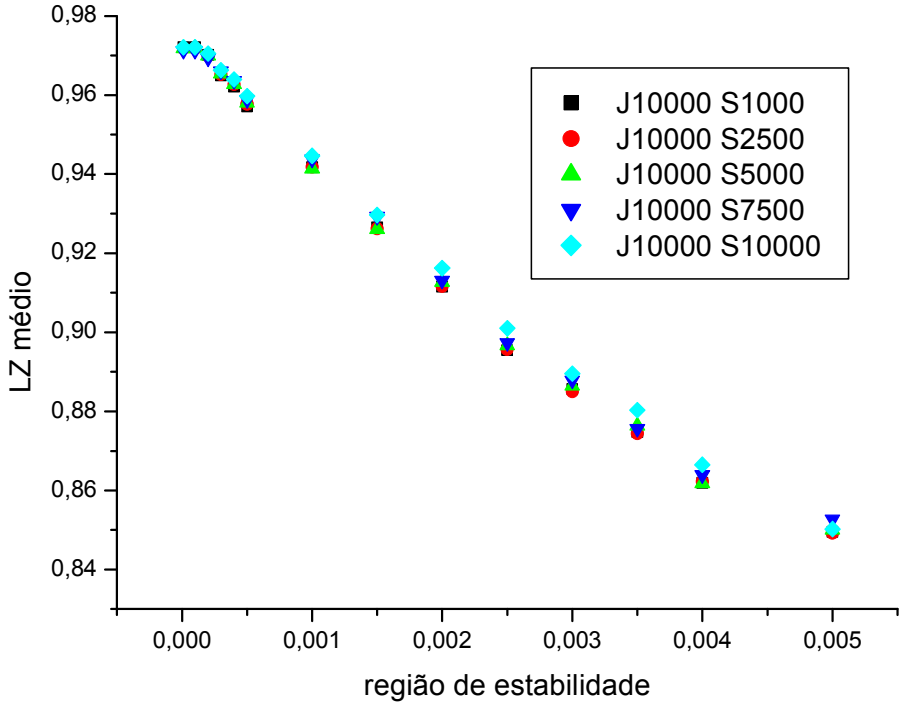

Figura 39: LZ médio *versus* RE TRPL4 JANELA 10000.

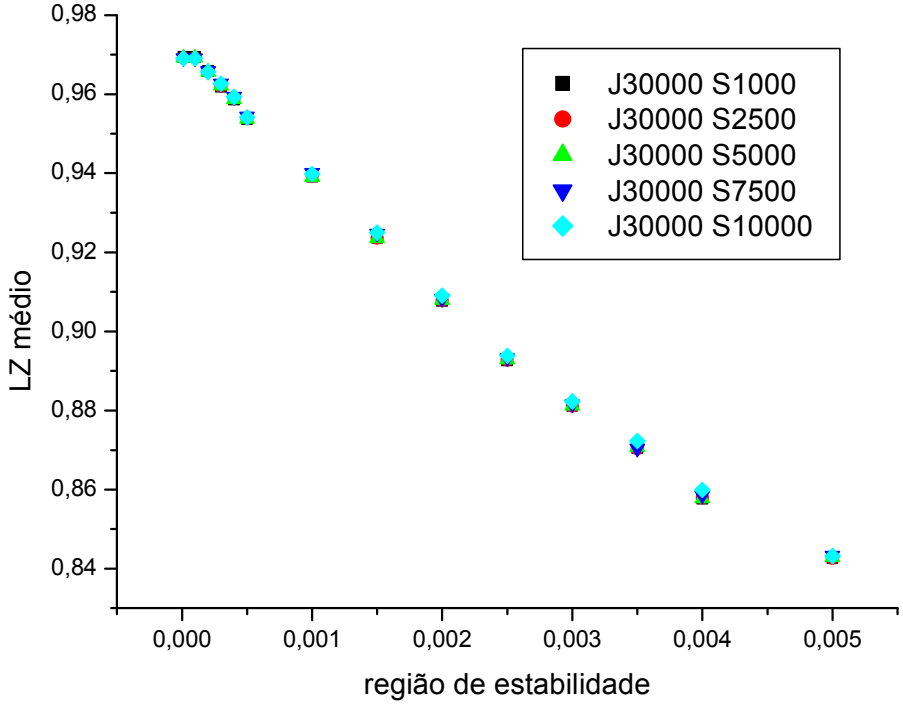

Figura 40: LZ médio *versus* RE TRPL4 JANELA 30000.

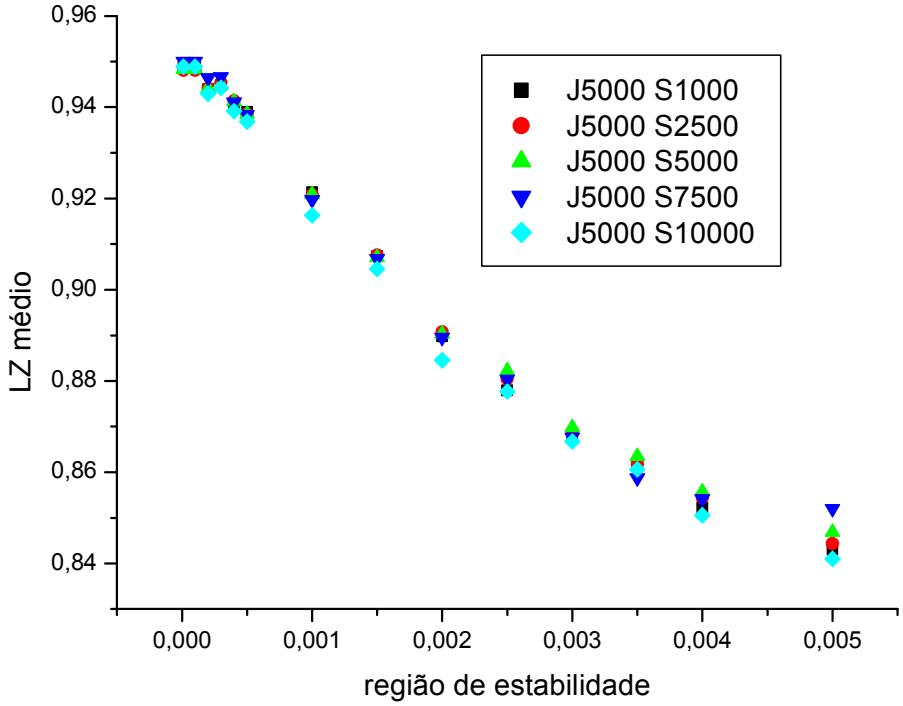

Figura 41: LZ médio *versus* RE UGPA4 JANELA 5000.

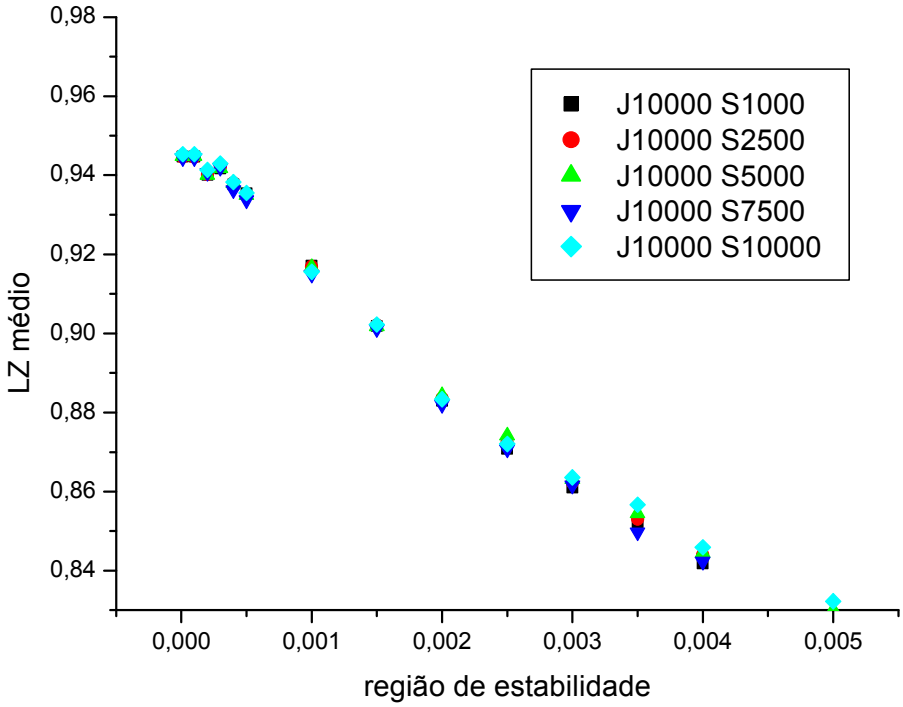

Figura 42: LZ médio *versus* RE UGPA4 JANELA 10000.

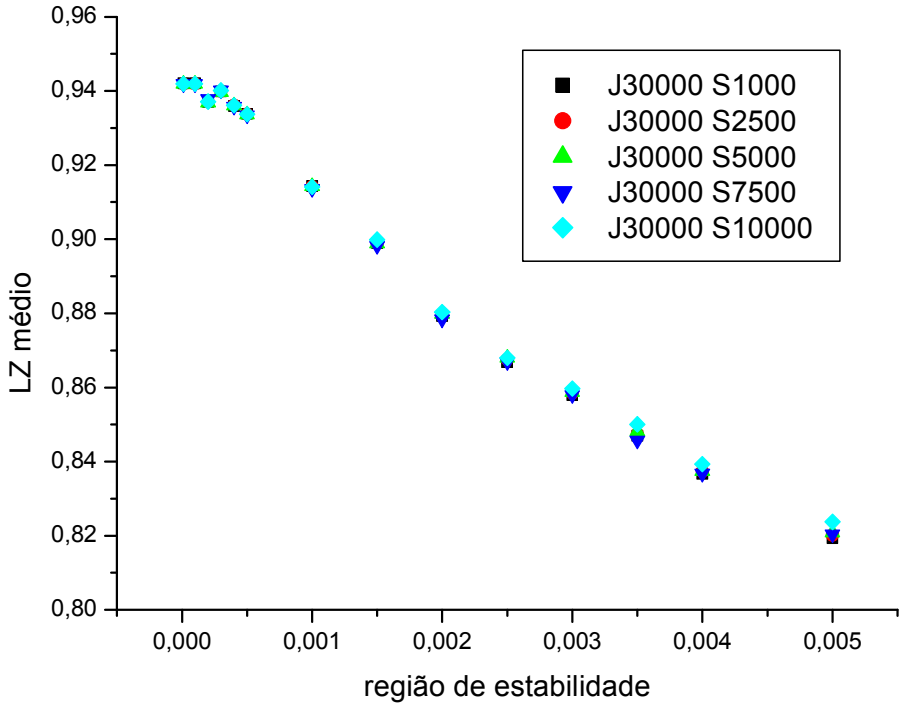

Figura 43: LZ médio *versus* RE UGPA4 JANELA 30000.

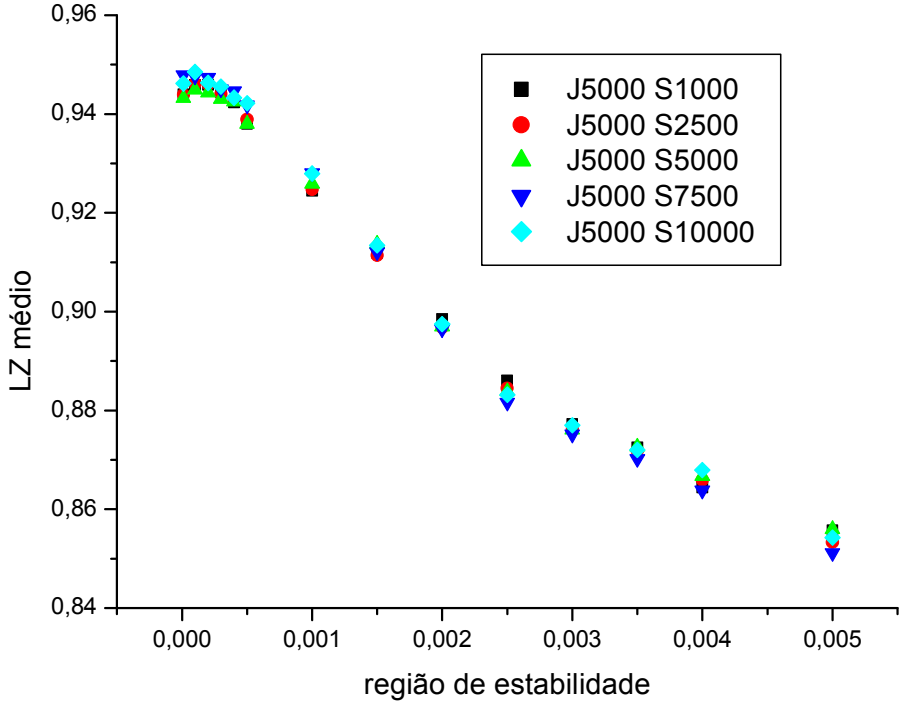

Figura 44: LZ médio *versus* RE USIM3 JANELA 5000.

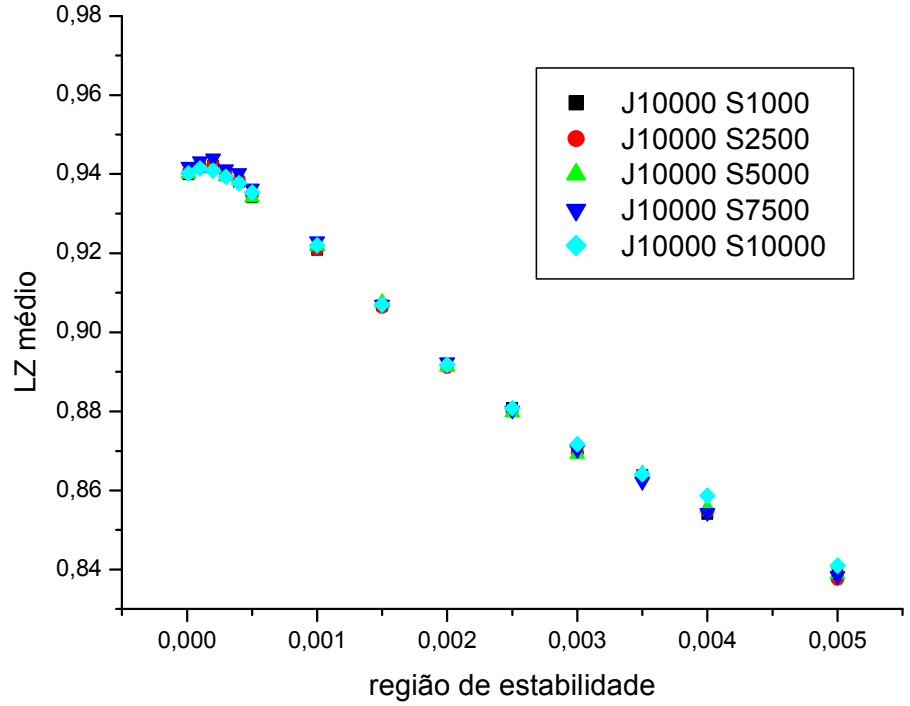

Figura 45: LZ médio *versus* RE USIM3 JANELA 10000.

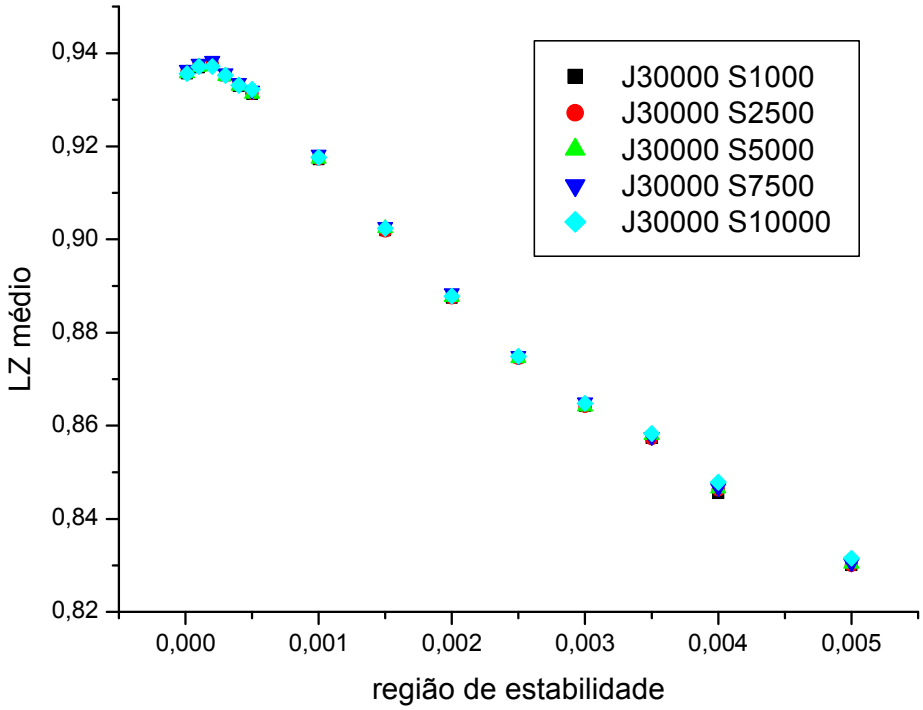

Figura 46: LZ médio *versus* RE USIM3 JANELA 30000.

O comportamento esperado é que o LZ diminua com o aumento da região de estabilidade, pois o tamanho do intervalo de retornos que será convertido em um só código (caractere “2”) aumenta gerando *runs* persistentes, o que possibilita uma rejeição da hipótese de aleatoriedade (RUKHIN 2000). Tal padrão foi evidenciado na maioria dos casos para o LZ médio exceto naqueles ilustrados na Figura 5 (BNCA3 (Nossa Caixa) Janela 5000 – houve crescimento do LZ a partir da região de estabilidade igual a 0,4%), na Figura 20 (SBSP3 (Sabesp) Janela 5000 – houve um leve crescimento do LZ a partir da região de estabilidade igual a 0,4%), na Figura 23 (SDIA4 (Sadia) Janela 5000 – houve crescimento brusco do LZ a partir da região de estabilidade 0,3%. A diferença do valor do LZ entre esta região e a de 0,5% foi de, aproximadamente, 11 centésimos), na Figura 24 (SDIA4 Janela 10000 – houve crescimento brusco do LZ a partir da região de estabilidade 0,3%. A diferença do valor do LZ entre esta região e a de 0,5% foi de, aproximadamente, 40 centésimos) e na Figura 25 (SDIA4 Janela 30000 – houve um leve crescimento do LZ a partir da região de estabilidade igual a 0,4%).

Observando os retornos das ações destacadas no parágrafo anterior, foram percebidos retornos positivos ou negativos persistentes fora da região de estabilidade onde foi verificado o menor valor de LZ médio, sendo esta uma possível explicação para o valor do LZ aumentar com o aumento da região de estabilidade. Um exemplo simples: observe as sequências de dados apresentadas na Figura 47 As barras horizontais mostram os limites de codificação dos valores. Na Figura 47 (a), a região de estabilidade é menor, gerando um padrão de código no trecho que fica inteiro fora daquela região. Ao ampliá-la (Figura 47 (b)), o padrão vai sendo descaracterizado na região da persistência de retornos propiciando aumento no LZ. Mas, deve-se verificar o fato de que a freqüência dos caracteres inseridos não pode ser grande, pois também afetaria o valor da entropia da fonte de forma significativa, aumentando-a e provocando alterações no valor do LZ de tal forma que não se pode definir sua direção.

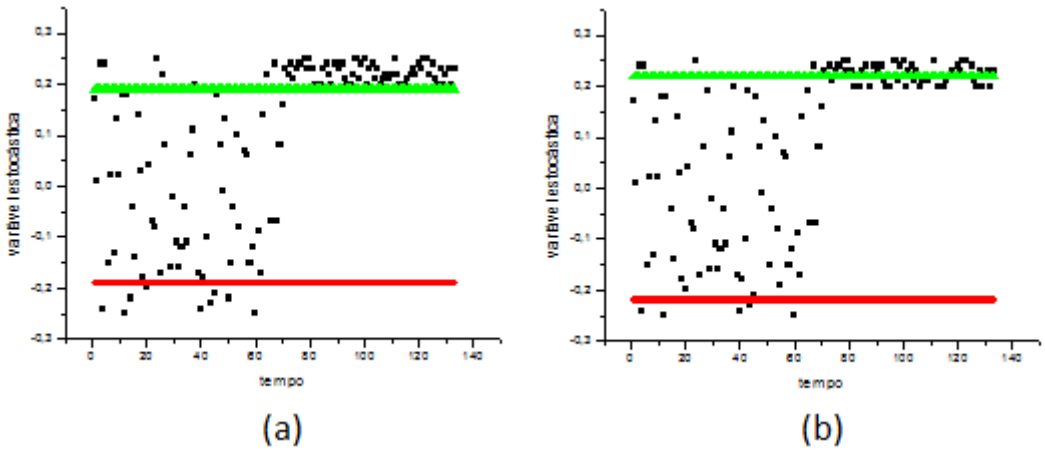

Figura 47: Variável estocástica *versus* tempo. (a) Região de estabilidade 0,19. (b) Região de estabilidade 0,22.

A ação SDIA4 pertence a uma companhia do gênero alimentício, a única desta categoria no rol de ações que foram instrumento de estudo deste trabalho.

Com base na propriedade de convergência do LZ, janelas maiores implicam convergência para a unidade. Mas, para todos os dados analisados, quanto maior o tamanho da janela, menor o valor do LZ médio encontrado para todas as regiões de estabilidade e para todos os tamanhos de salto avaliados neste trabalho. Isto é um possível indicativo de que a série de retornos de alta freqüência dos ativos testados é não aleatória quando comparada à sequência de de Bruijn.

Com relação aos saltos, foi observado que o aumento no tamanho da janela proporcionou menores diferenças de LZ médio entre a categoria de salto que possuía maior LZ médio e a que apresentava menor valor deste para uma dada região de estabilidade. A Tabela 4 ilustra os valores obtidos. Há indício de uma possível convergência no valor do LZ médio obtido para cada salto, que independe da ação analisada e da região de estabilidade.

Tabela 4: Maiores diferenças de LZ médio observadas entre saltos de tamanhos distintos

Tamanho de Janela	Maior diferença de LZ médio observada	Ação observada	Região de estabilidade (%)
5000	0,0155	BNCA3	0,4
10000	0,0094	TMAR5	0,5
30000	0,0044	UGPA4	0,35

3.2.2 Evolução da Eficiência relativa (GIGLIO 2008) em relação à região de estabilidade.

A medida de eficiência relativa de mercado, proposta por Giglio (2008) foi igual a zero para a maioria dos experimentos realizados sobre os elementos da base de dados de alta freqüência apresentados na Tabela 2 com as condições impostas pela combinação dos termos apresentados na Tabela 3. As figuras Figura 48 a Figura 53 referem-se às situações de eficiência relativa de mercado diferente de zero.

Em todas as ações analisadas, para janelas de tamanho 30000, a medida de eficiência relativa em Giglio (2008) apresentou valores nulos. Isto foi devido à baixa dispersão existente entre valores de LZ de cada janela, tal fato estando de acordo com o critério de convergência da complexidade normalizada.

Nas situações em que houve comportamento diferente daquela medida (figuras Figura 48 a Figura 53), foi possível verificar que, com o aumento da janela, seu valor diminui para todos os saltos. É importante ressaltar que o processo de convergência da medida do LZ faz com que o desvio padrão dos valores de LZ das janelas diminua provocando redução da quantidade de valores extremos e uma aproximação ao valor médio. Como todos os valores de LZ médio encontrados foram menores que 1, então, a diminuição na eficiência relativa (GIGLIO 2008) foi esperada. Todavia, um LZ médio em torno de 0,97 da ação SDIA4 (janela de 5000, região de estabilidade 0,5% e salto de 1000) produziu uma medida de eficiência relativa (GIGLIO 2008) de 30%. Para a ação BNCA3 (janela 5000, região de estabilidade 0,5% e salto de 1000), um LZ médio de 0,89 gerou uma medida de eficiência relativa de cerca de 12%.

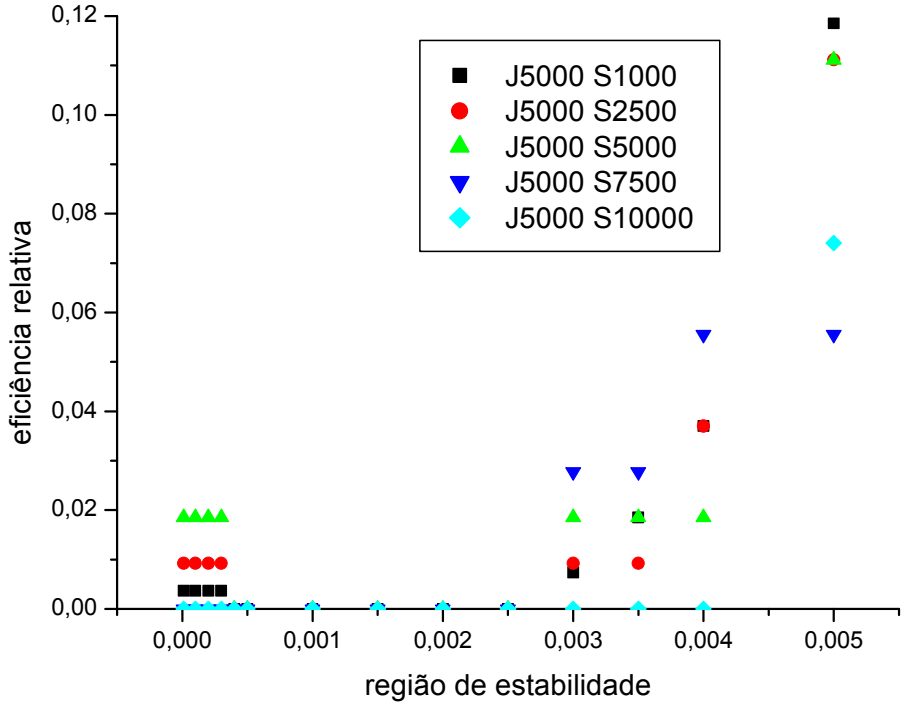

Figura 48: Eficiência Relativa *versus* RE BNCA3 JANELA 5000.

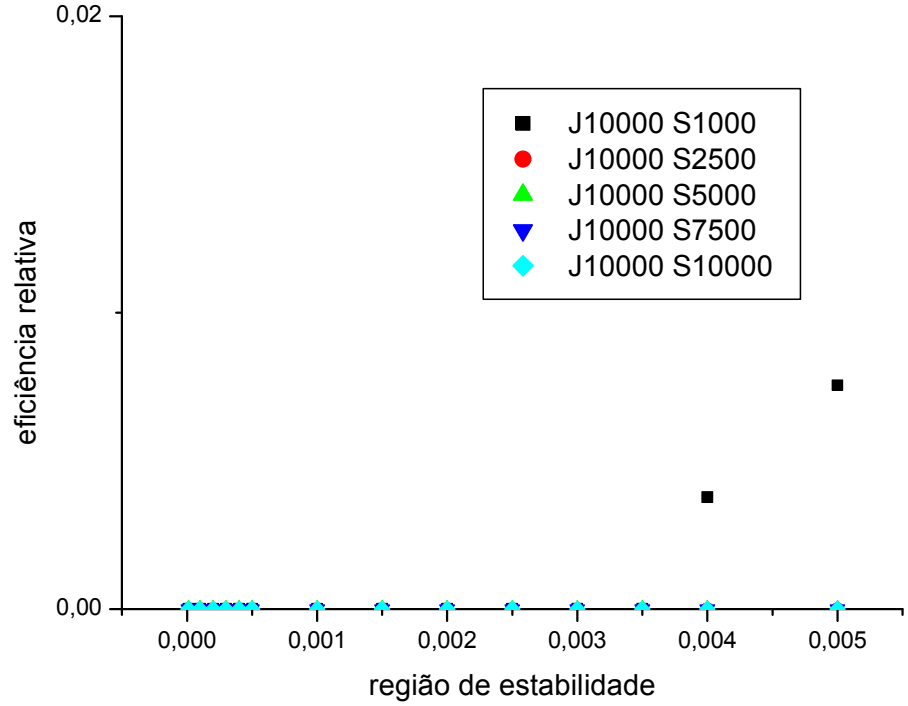

Figura 49: Eficiência Relativa *versus* RE BNCA3 JANELA 10000.

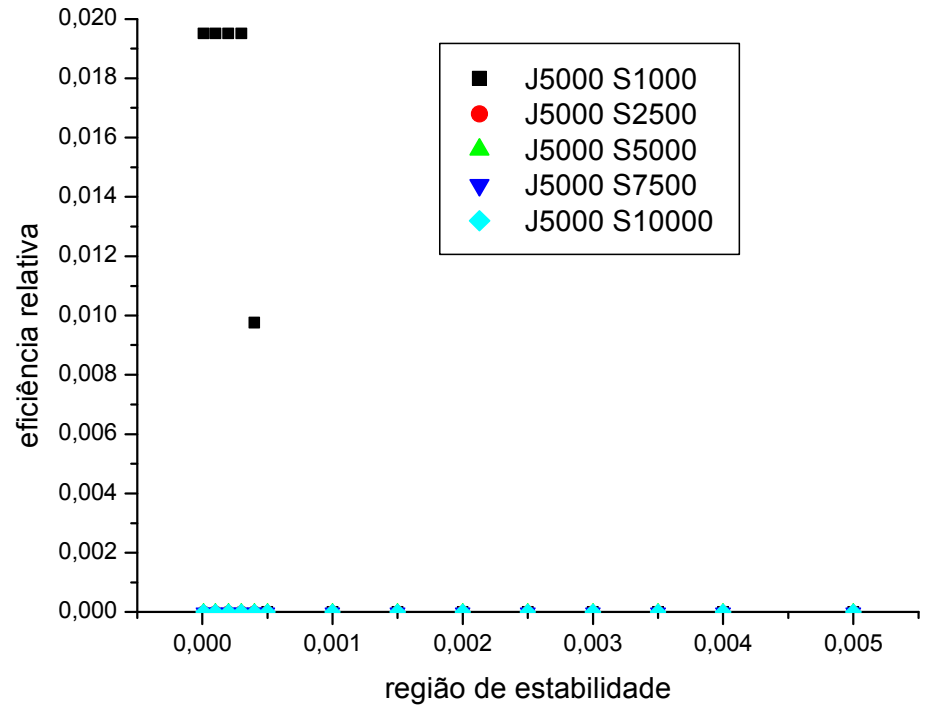

Figura 50: Eficiência Relativa *versus* RE LIGT3 JANELA 5000.

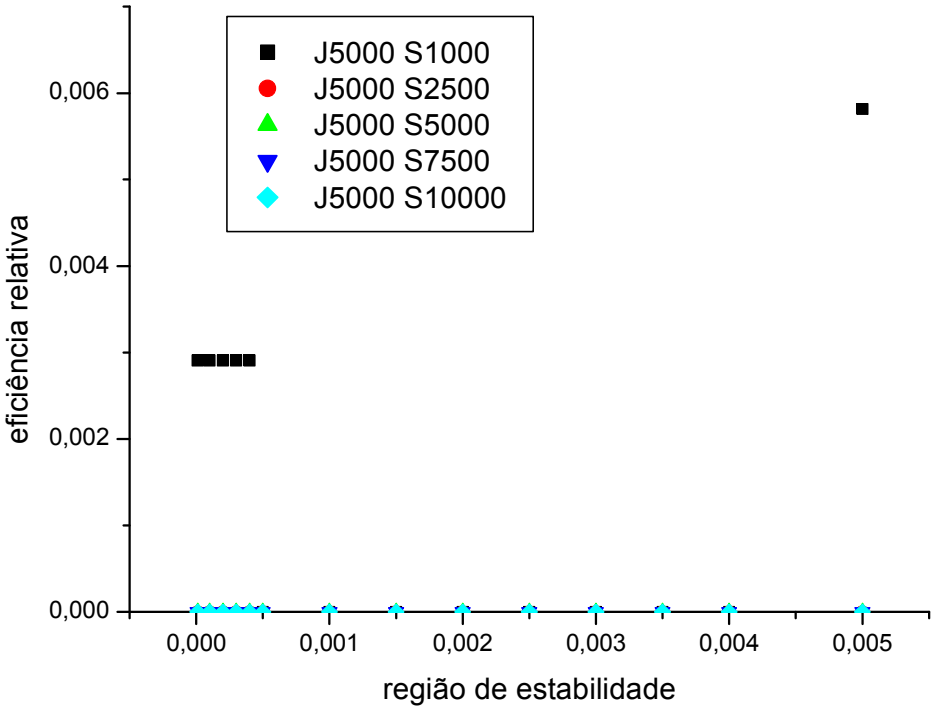

Figura 51: Eficiência Relativa *versus* RE SBSP3 JANELA 5000.

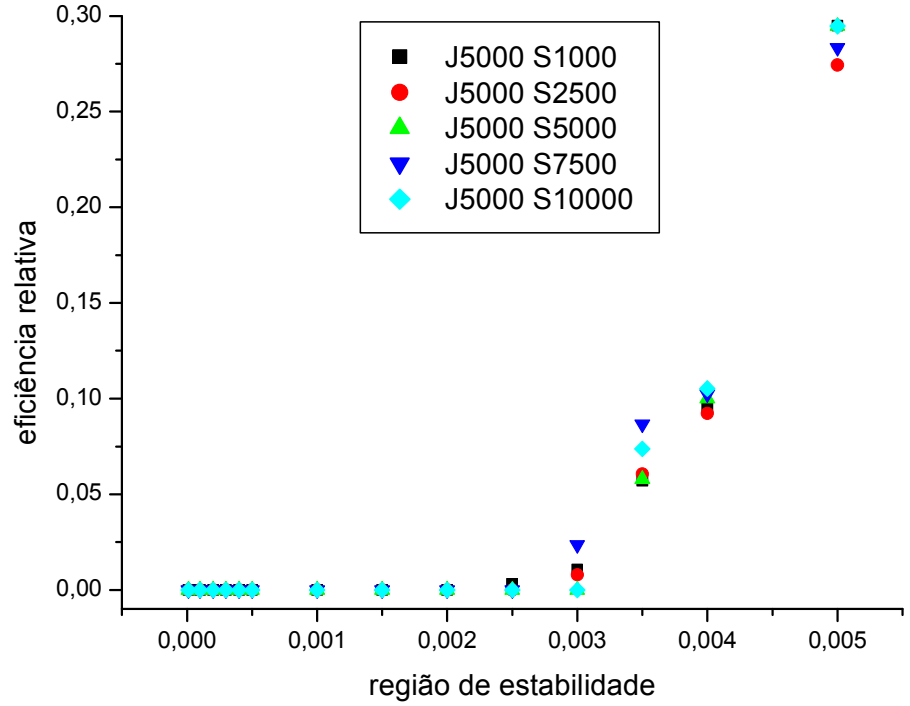

Figura 52: Eficiência Relativa *versus* RE SDIA4 JANELA 5000.

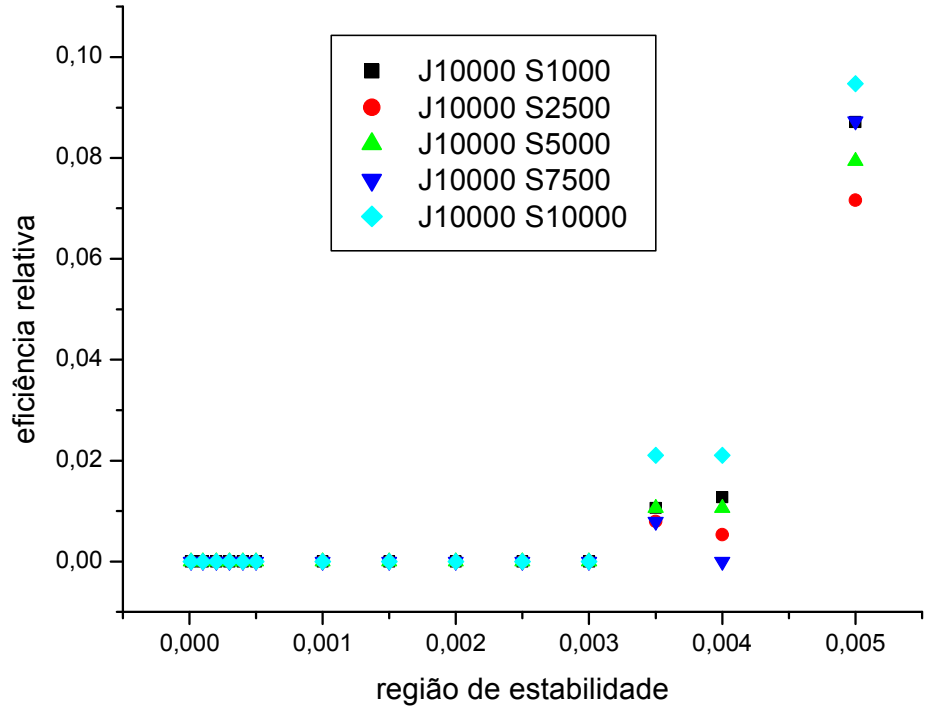

Figura 53: Eficiência Relativa *versus* RE SDIA4 JANELA 10000.

3.3 **Análise dos resultados no caso de intervalos de tempo maior que uma negociação.**

Para os estudos relativos a esta seção foram utilizadas as ações CGAS5 (Comgás) e ELPL6 (Eletropaulo) bem como os tamanhos de janela igual a 1000 (de acordo com Kaspar e Schuster (1987) o LZ converge para seu valor assintótico com tolerância de 5% quando uma sequência de valores possui comprimento igual a 1000) e de salto igual a 1, bem como o valor de 0,25% para a região de estabilidade.

Foram gerados valores de LZ médio, de medida de eficiência relativa de mercado (GIGLIO 2008) e de desvio padrão do LZ para cada intervalo de negociação.

Os comportamentos do LZ médio, da eficiência relativa de mercado e do desvio padrão dos valores de LZ referentes às janelas para cada intervalo entre negociações para as 2 ações podem ser observados nas figuras 54 a 56, 58 a 61 e 63. O eixo das abscissas das ações CGAS5 e ELPL6 possuem escalas diferentes, devido a limitações no número de pontos obtidos na série de retornos.

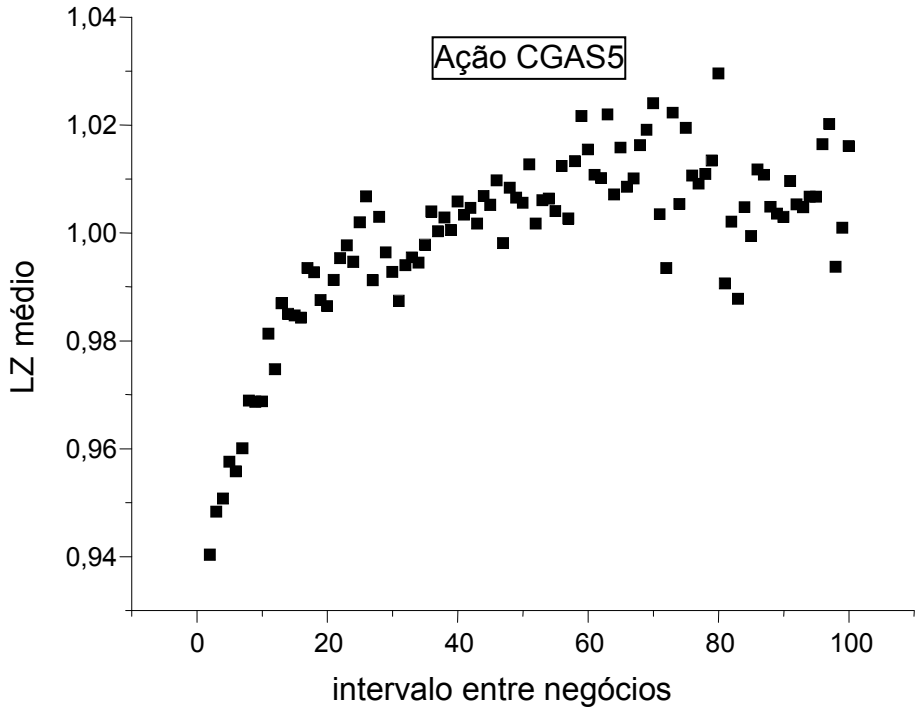

Figura 54: LZ médio *versus* intervalo entre negócios para a ação CGAS5.

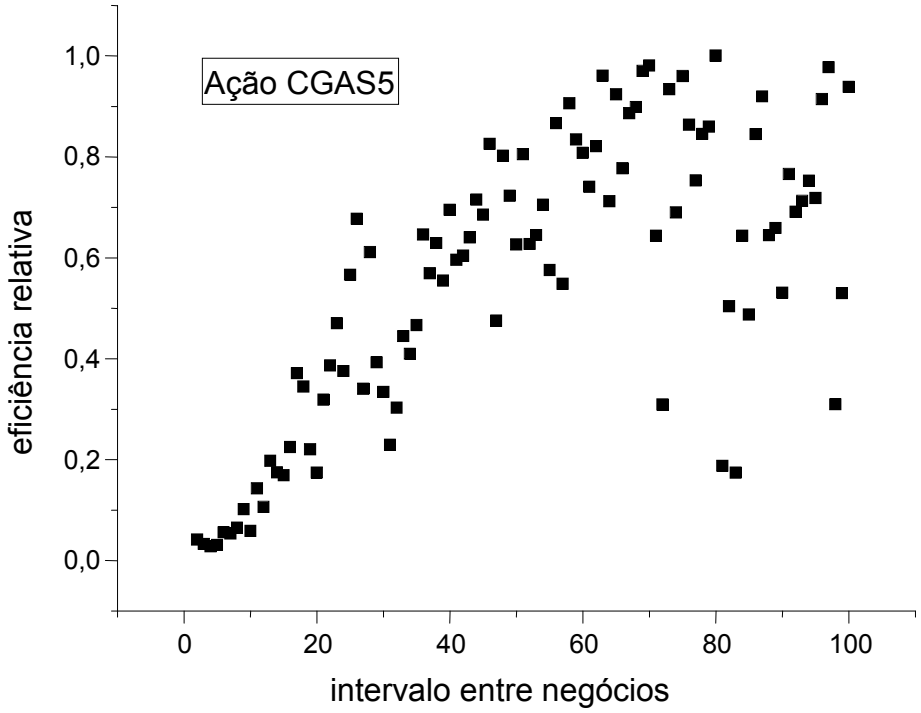

Figura 55: Eficiência relativa *versus* intervalo entre negócios para a ação CGAS5.

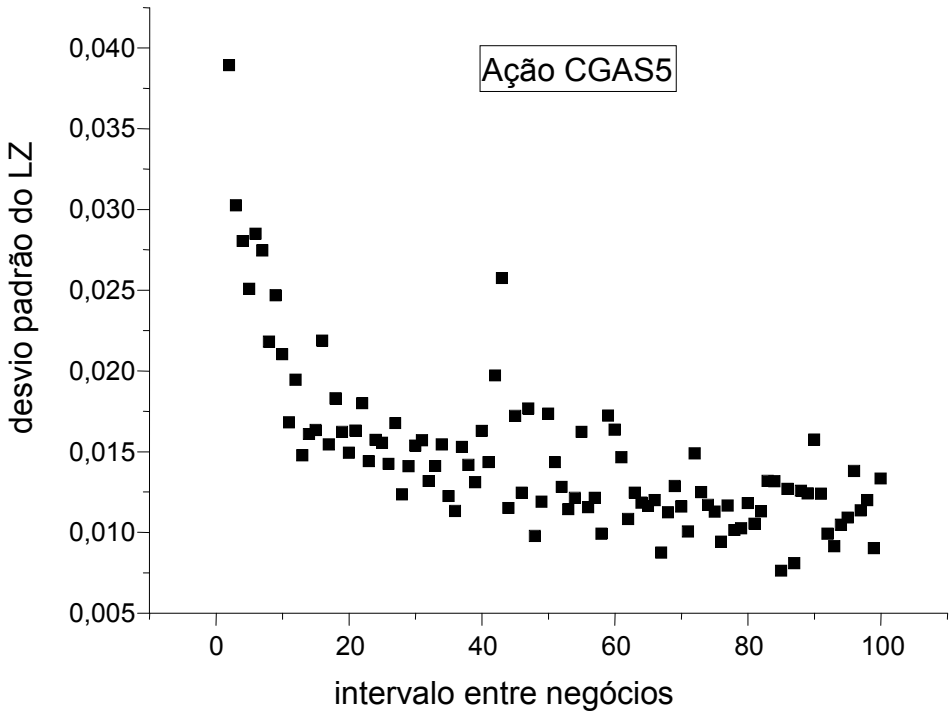

Figura 56: Desvio padrão do LZ médio *versus* intervalo entre negócios para a ação CGAS5.

Para a ação CGAS5, há um acréscimo no valor do LZ médio à medida que o tempo entre as negociações vai aumentando, indicando uma dispersão menor nos dados a partir do intervalo de cerca de 60 negociações. O desvio padrão entre as janelas pertencentes a um mesmo intervalo de tempo, possui tendência decrescente, tornando-se assintótica a partir do intervalo de tempo por volta de 40 negociações. Já a medida de eficiência relativa de mercado (GIGLIO 2008) apresentou elevado grau de dispersão entre os intervalos de tempo 20 e 40 e após 50 negociações apesar da aparente tendência de crescimento.

Para cada intervalo de negociação determinamos as autocorrelações correspondentes utilizando como estimativa a correlação de Pearson e observamos o número de ocorrências de valor significativo desta estimativa (cujo valor absoluto foi definido neste trabalho como superior a 0,05) ao longo das defasagens entre os retornos para cada intervalo fixo entre negócios.

No caso da ação CGAS5, o gráfico da Figura 57 mostra o número de valores absolutos de autocorrelação superiores a 0,05 entre as 20 primeiras defasagens, entre as 100 primeiras defasagens, entre as 1000 primeiras defasagens e no intervalo de 800 a 1000 defasagens para todos os intervalos entre negócios. Para as primeiras 20 defasagens e para as primeiras 100 defasagens foram encontrados, em média, 2 valores que satisfazem a condição estabelecida no início do parágrafo.

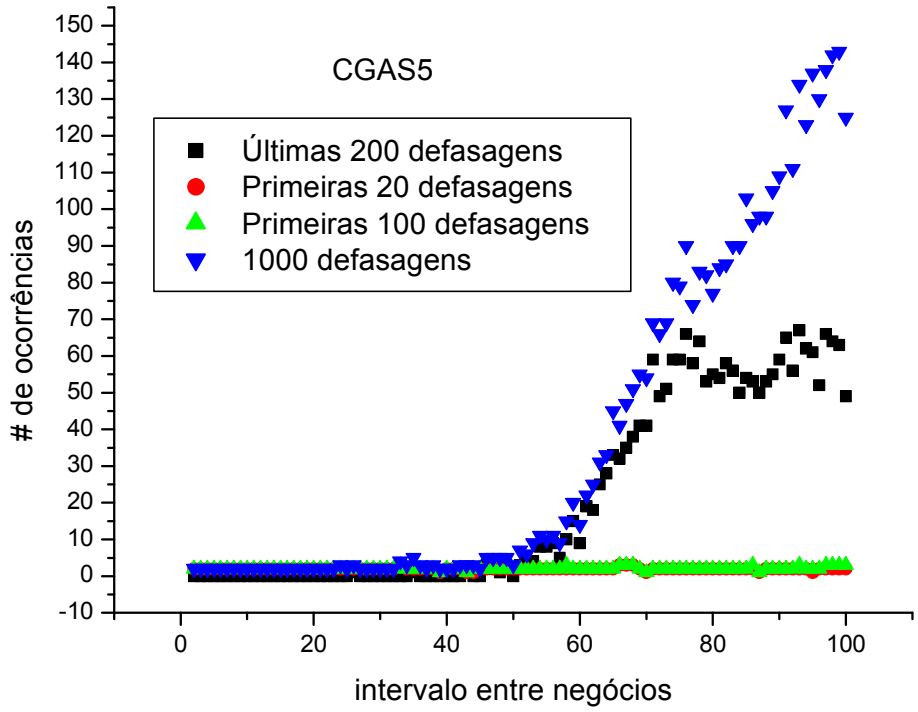

Figura 57: Número de valores absolutos de autocorrelação superiores a 0,05 *versus* intervalo entre negócios para a ação CGAS5.

Foi percebido, conforme a Figura 57, aumento da autocorrelação entre retornos defasados de períodos longos de tempo para intervalos maiores entre negociações. Tal evidência pareceu alinhar-se com as conclusões obtidas em Lo e MacKinlay (1988) no sentido de existir indício que, em intervalos distantes de negociação pode haver dependência entre os retornos. Para intervalos pequenos entre negociações as autocorrelações tiveram queda rápida conforme evidenciado por Mantegna e Stanley (2000).

A Figura 58 mostra, para a ação CGAS5 um diagrama contendo os pares ordenados eficiência relativa (GIGLIO 2008) – LZ médio para cada negociação. Aplicando o índice de correlação de Pearson para extrair uma medida que associasse o crescimento do LZ médio ao da eficiência relativa de mercado, o valor obtido foi de 0,92 sendo este um indício forte de que a medida de eficiência relativa de mercado acompanha a evolução do LZ médio restrita, neste caso, à condição dos parâmetros janela, salto e região de estabilidade mantidos constantes e para o ativo CGAS5.

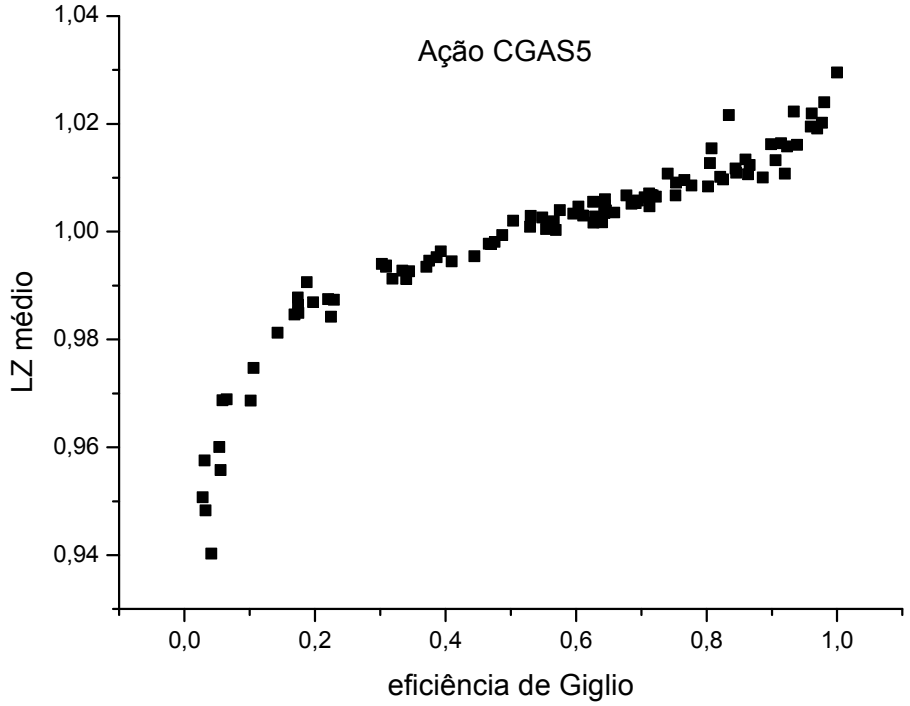

Figura 58: LZ médio *versus* eficiência relativa de mercado (GIGLIO 2008).

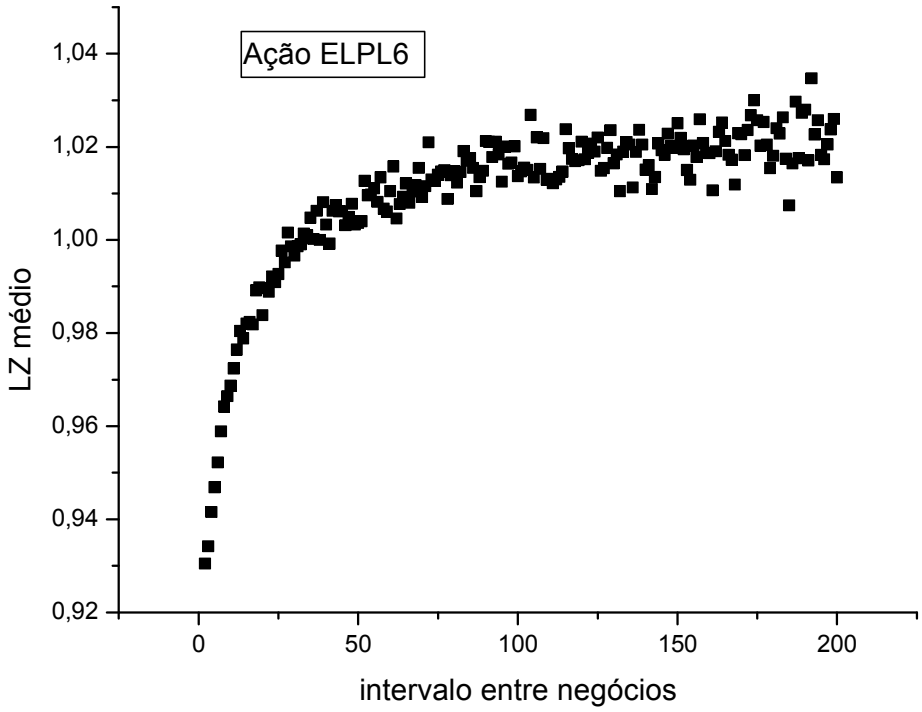

Figura 59: LZ médio *versus* intervalo entre negócios para a ação ELPL6.

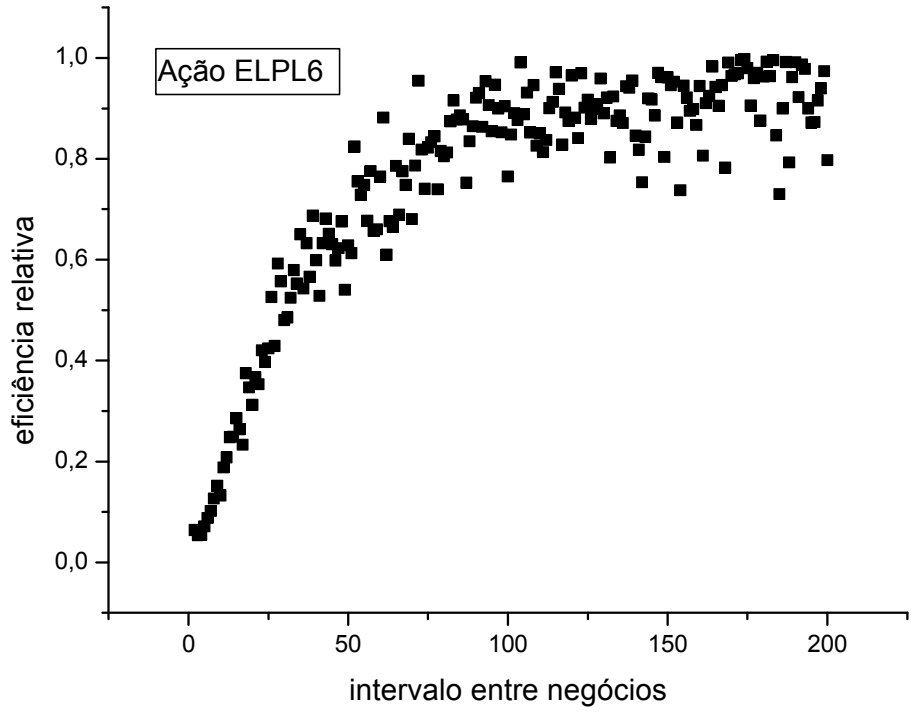

Figura 60: Eficiência relativa *versus* intervalo entre negócios para a ação ELPL6.

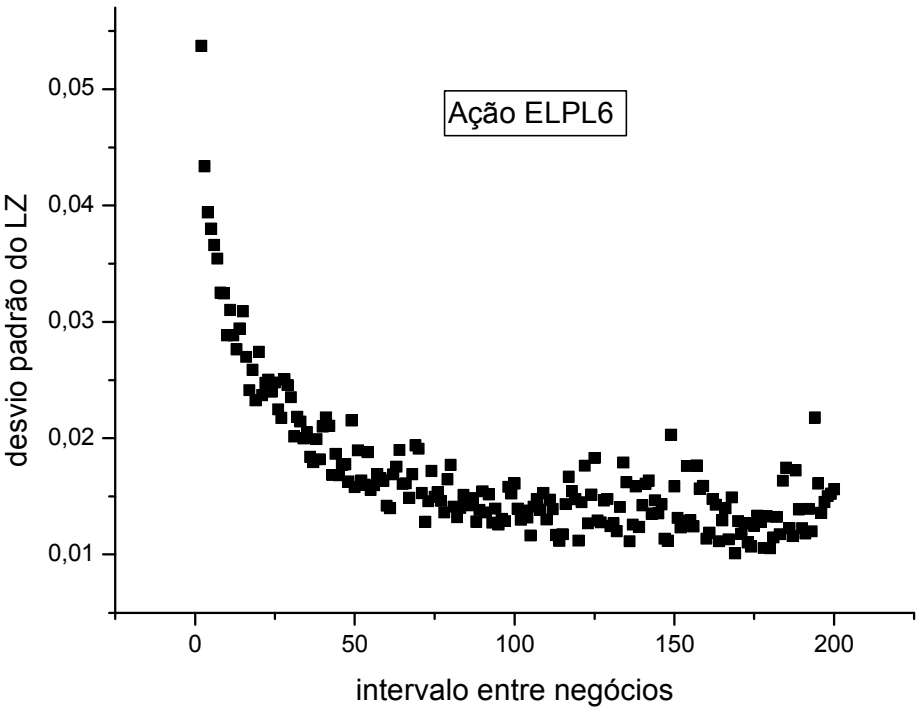

Figura 61: Desvio padrão do LZ médio *versus* intervalo entre negócios para a ação ELPL6.

Para a ação ELPL6, há um acréscimo no valor do LZ médio à medida que o tempo entre as negociações vai aumentando, indicando uma tendência assintótica no intervalo de aproximadamente 100 negociações. O desvio padrão entre as janelas pertencentes a um mesmo intervalo de tempo possui tendência decrescente, tornando-se assintótica a partir do intervalo de tempo de aproximadamente 80 negociações. O grau de dispersão da medida de eficiência (GIGLIO 2008) foi menor do que o visualizado para a ação CGAS5. Para cada intervalo de negociação, determinaram-se as autocorrelações correspondentes utilizando como estimativa a correlação de Pearson e foram observados o número de ocorrências de valor significativo desta estimativa (cujo valor absoluto foi definido neste trabalho como superior a 0,05) ao longo das defasagens entre os retornos para cada intervalo fixo entre negócios.

No caso da ação ELPL6, o gráfico da Figura 62 mostra o número de valores absolutos de autocorrelação superiores a 0,05 entre as 20 primeiras defasagens, entre as 100 primeiras defasagens, entre as 1000 primeiras defasagens e no intervalo de 800 a 1000 defasagens para todos os intervalos entre negócios. Para as primeiras 20 defasagens e para as primeiras 100 defasagens foram encontrados, em média, 2 valores que satisfazem a condição estabelecida no início do parágrafo.

Foi percebido, conforme a Figura 62, aumento da autocorrelação entre retornos defasados de períodos longos de tempo para intervalos maiores entre negociações em um grau menor que para a ação CGAS5. Para intervalos pequenos entre negócios observa-se o comportamento de rápido decaimento de autocorrelação observado por Mantegna e Stanley (2000).

A Figura 63 mostra, para a ação ELPL6 um diagrama contendo os pares ordenados eficiência (GIGLIO 2008)– LZ médio para cada negociação. Aplicando o índice de correlação de Pearson para extrair uma medida que associasse o crescimento do LZ médio ao da eficiência relativa de mercado, o valor obtido foi de 0,96, sendo este um indício forte de que aquela medida acompanha a evolução do LZ médio restrita, neste caso, à condição dos parâmetros janela, salto e região de estabilidade mantidos constantes e para o ativo ELPL6.

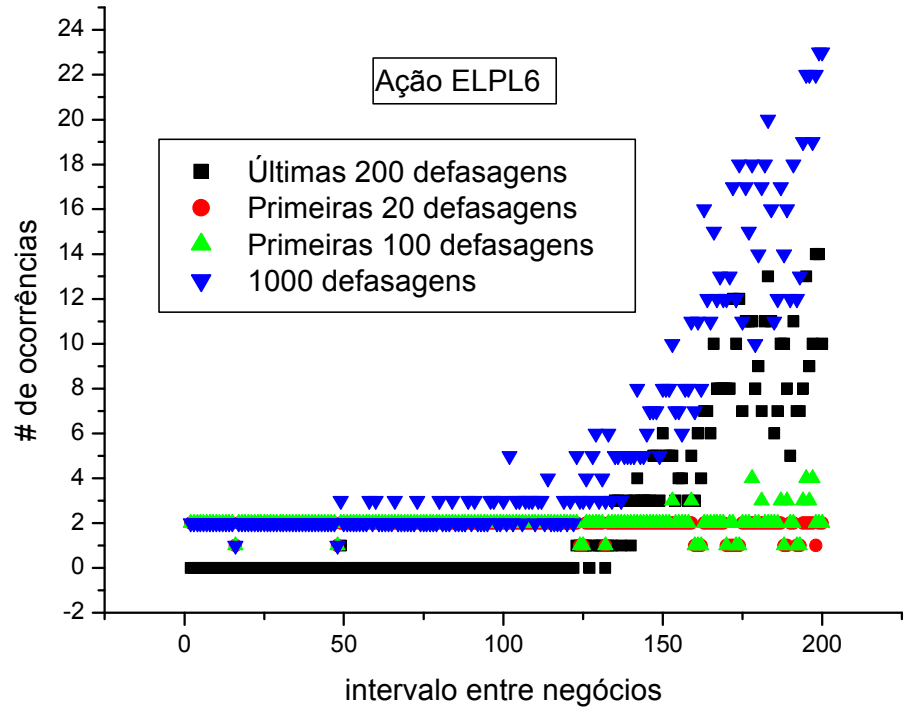

Figura 62: Número de valores absolutos de autocorrelação superiores a 0,05 *versus* intervalo entre negócios para a ação ELPL6.

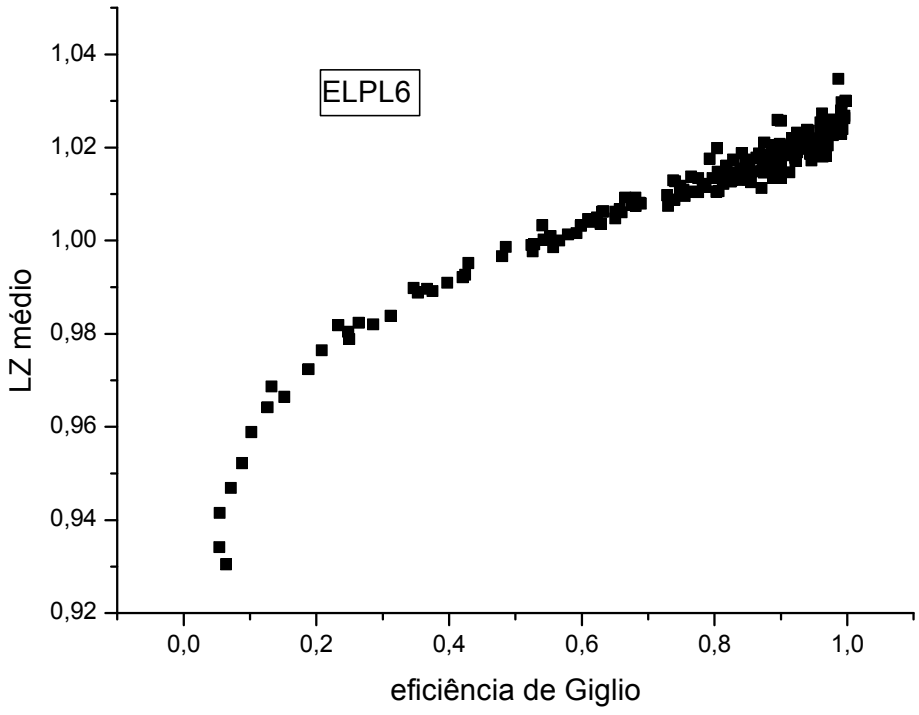

Figura 63: LZ médio *versus* eficiência relativa de mercado (GIGLIO 2008).

4 CONCLUSÃO

Este trabalho apresentou uma teoria que, de acordo com Mantegna e Stanley (2000) verifica a hipótese de eficiência de mercado na sua forma fraca com base na premissa de que esta é valida se os preços flutuarem aleatoriamente: a teoria da informação algorítmica. Nesta área do conhecimento é possível verificar a quantidade de informação existente em uma sequência de dados. Uma das medidas numéricas para se verificar aproximadamente esta quantidade é a medida de complexidade proposta por Lempel e Ziv (1976) que é baseada na idéia de complexidade de Kolmogorov.

A medida foi aplicada a séries de retornos de alta freqüência para 15 ativos negociados na BM&F BOVESPA. Foi utilizado o procedimento computacional de Kaspar e Schuster (1987) com a codificação proposta por Shmilovici et. al. (2003). Foi também analisada a medida de eficiência relativa de mercado proposta por Giglio (2008).

Os dados de alta freqüência não apresentaram, de acordo com o critério do LZ médio, aleatoriedade, desta forma, negando o comportamento aleatório de retornos de ações em séries de alta freqüência. No entanto, a medida eficiência relativa de mercado (GIGLIO 2008), para alguns casos, aponta no sentido de existir aleatoriedade local nesta série. De acordo com Giglio (2008) a medida de eficiência relativa proposta em seu trabalho verifica a quantidade de informação não redundante contida em uma série financeira em relação ao total de informações. Foi observado que tal medida é extremamente sensível ao tamanho da janela, ao tamanho de salto e à região de estabilidade em alguns ativos.

Por meio da realização da análise de sensibilidade do LZ médio, percebeu-se que, para valores pequenos de tamanho da janela e no domínio de região de estabilidade de até 0,5% não é possível definir um procedimento que vise extrair a região de estabilidade que maximize aquela medida. Tampouco é possível encontrar, um intervalo de negociações que gere um valor ótimo de LZ e da medida de eficiência relativa que não dependa do trio de parâmetros: janela, salto e região de estabilidade.

O estudo das autocorrelações das ações CGAS5 e ELPL6 evidenciaram memória curta para intervalos pequenos de negociação, observando a independência de curto prazo, porém, para intervalos grandes de negociação, foi percebido aumento do índice

de correlação para grandes defasagens assim, havendo um maior grau de dependência entre retornos muito afastados, alinhando-se, de certo modo, com o proposto em Lo e MacKinlay (1988).

Sugestões para novos trabalhos podem ser: tentar verificar a aleatoriedade dos preços de mercado para uma nova codificação da sequência de retornos diferente da proposta em Shmilovici et. al. (2003), analisar a sensibilidade com relação ao tamanho da janela, ao do salto e ao da região de estabilidade um conjunto de ativos de um mesmo setor, dado que, para as 5 ações analisadas do setor de telefonia (TMAR5, BRTP3, TLPP4, TCSL3 e TNPL3) o comportamento do LZ em relação a mudanças nos parâmetros mencionados foi semelhante, assim como para 3 ações do setor de energia elétrica (CLSC6, LIGT3 e TRPL4) e de avaliar, com outros instrumentos, a robustez da medida em Giglio (2008), bem como seu domínio de validade caso seja definido o LZ médio como parâmetro de aleatoriedade.

REFERÊNCIAS

BAK, P.; PACZLUSKI, M. “Complexity, Contingency and Criticality”, in: Physics: The Opening to Complexity, 1994, Irvine, **Colóquio**. Irvine: Proc. Natl. Acad. Sci. USA, 1995. v. 92 p. 6689-6696.

BANDT, C.; POMPE, B. “Permutation entropy: A natural complexity measure for time series”. *Physical Review Letters*, v. 88, n. 17, 2002.

BAK,P.; TANG C.; WIESENFELD, K. “Self Organized Criticality: An Explanation of *1/f* noise”. *Physical Review Letters,* v. 59, n. 4, p. 381-384, 1987.

BENDAT, J. S.; PIERSOL, A. G. *Random data: Analysis and measurement procedures*. 2 ed. John Willey and Sons, 1986.

CAMPANI, C. A. P.; MENEZES, P. B. “Teorias da aleatoriedade”. *Revista de Informática Teórica e Aplicada*, v. 11, n. 2, p. 75-98, 2004.

CALUDE C. S. *“A glimpse into algorithmic information theory”*. Research Report Series Centre for Discrete Mathematics and Theoretical Computer Science, n. 93, Department of Computer Science. University of Auckland, 1999

CHAITIN, G. J. “On the length of programs for computing finite binary sequences”. *Journal of the ACM*, v. 13, n. 4, p. 547-569. 1966.

De BRUIJN, N. G."Acknowledgement of Priority to C. Flye Sainte-Marie on the counting of circular arrangements of 2n zeros and ones that show each n-letter word exactly once", T.H.-Report 75-WSK-06, Technological University Eindhoven, 13 p., 1975.

FAMA, E. F. “Efficient Capital Markets: A review of Theory and Empirical Work.” *Journal of Finance*. v. 25, n. 2, p. 383-417, 1970.

FAMA, E. F. “Efficient Capital Markets II.” *Journal of Finance*. v. 46, n. 5, p. 1575-1617, 1991.

GIGLIO, R. *Eficiência relativa de mercado sob a perspectiva da teoria da informação algorítmica*. Dissertação de Mestrado, Economia. Universidade Federal de Santa Catarina, 2008.

GLÉRIA, I. M.; MATSUSHITA, R.; da SILVA, S. ”Sistemas complexos, criticalidade e leis de potência”. Revista Brasileira de Ensino de Física, v. 26, n. 2, p. 99-108, 2004

GROSSMAN, S.; STIGLITZ, J. “On the impossibility of informationally efficient markets”. *The American Economic Review*, v. 70, n. 3, p. 393-408, 1980.

HEYLIGHEN, F.: "Building a Science of Complexity", in: Annual Conference of the Cybernetics Society, 1988, Londres. **Proceedings**. Londres: King's College: H.A. Fatmi, 1988. p. 1-22. Disponível em <http://pespmc1.vub.ac.be/papers/PapersFH2.html>. Acesso em 27 de maio de 2010.

HURLBERT, G.; ISAAK, G. "On the De Bruijn torus problem". *J. Comb. Th. (A). v. 64, n. 1,* p. 50–62, 1993

ISNARD, C. *Introdução à Medida e Integração*. 1 ed. Rio de Janeiro: IMPA, 2007.

JENSEN, M. C. “Some anomalous evidence regarding market efficiency”. *Journal of Financial Economics*, v.6, n. 2/3, p. 96-101. 1978

KASPAR, F.; SCHUSTER, H. “Easily calculable measure for the complexity of spatiotemporal patterns”. *Physical Review A*, v/ 36, n. 2, p. 842-848, 1987.

KOLMOGOROV, A. N. “Three approaches to the quantitative definition of information”. *Problems of Information Transmission*. v. 1, n. 1, p. 3-11, 1965.

LEMPEL, A.; ZIV, J. “On the complexity of finite sequences”. *IEEE Transactions on Information Theory*, v. 22, n. 1, p. 75-81, 1976.

LI, M.; VITÁNYI, P. *An Introduction to Kolmogorov Complexity and Its Applications*, Springer, 1997. Disponível em < <http://books.google.com.br/books?id=25fue3UYDN0C&printsec=frontcover&dq=An+Introduction+to+Kolmogorov+Complexity+and+Its+Applications&source=bl&ots=U49l7fU7ej&sig=MWK0ogQ0Mk8o9RUpazGmSSU-A2s&hl=pt-BR&ei=e4IBTMr6H8yQuAeGl4j3DQ&sa=X&oi=book_result&ct=result&resnum=5&ved=0CDwQ6AEwBA#v=onepage&q&f=false>. Acesso em 27 de maio de 2010.

LO, A. *Efficient Markets Hypothesis*. The New Palgrave: A Dictionary of Economics; Nova Iorque. Palgrave McMillan, 2 ed., p. 1-24, 2007

LO, A.; MACKINLAY, C. “Stock market prices do not follow random walks: evidence from a simple specification test”. *Review of Financial Studies*. v. 1, n. 1, p. 41-66, 1988.

MALKIEL, B. G. “The efficient market hypothesis and its critics”. CEPS Working Paper n. 91. 2003, 47 p.

MANDELBROT, B. B. “The variation of certain speculative prices”, *Journal of Business* v. 36, p. 394-419, 1963.

MANTEGNA, R. N.; STANLEY, H. E. *An introduction to Econophysics: correlation and complexity in finance*. Cambridge University Press, 2000.

MARTIN-LÖF, P. “The definition of random sequences”. *Information and Control*, v. 9, p. 602-609, 1966.

MOREIRA, N. *Computabilidade: uma introdução.* Notas de Aula. Departamento de Ciência de Computadores, Universidade do Porto, 2003. Disponível em <http://www.ncc.up.pt/~nam/publica/compdec.pdf>. Acesso em: 27 de maio de 2010.

RUKHIN, A. L. “Testing randomness: a suite of statistical procedures”. *Theory of Probability and its Applications*, v. 45, n. 1, p. 111-132, 2000.

SAMUELSON, P. “Proof that properly anticipated prices fluctuate randomly”. *Industrial Management Review*. v. 5, n. 1, p. 41-49, 1965.

SCHIMIDT, A. B. *Quantitative Finance for Physicists: An Introduction*. Elsevier Academic Press, 2005

SHANNON, C. E. “A Mathematical Theory of Communication”. *Bell System Technical Journal*, v. 27, p. 379-423, 1948.

SHMILOVICI, A.; ALON-BRIMER, Y.; HAUSER, S. “Using a stochastic complexity measure to check the efficient market hypothesis”. *Computational Economics*, v. 22, p. 273-284, 2003.

TEIXEIRA, J. F. *Mentes e Máquinas: Uma introdução à ciência cognitiva*. Artes Médicas, 1998.

VOLCHAN, S. B. “What is a random sequence?”. *The American Mathematical Monthly*, p. 46-63, Janeiro de 2002.