1.03.02 Cincia da Computao / Matemtica da Computao

TEORIA DA INFORMAO E ESTATSTICA COMPUTACIONAL NO PROCESSAMENTO E ANLISE DE SINAIS IMPLEMENTAES OTIMIZADAS EM R

Eduarda T. C. Chagas[1], Alejandro C. Frery[2]

[1]*Estudante de IC de Ciência da Computação, Ufal*
[2]*Laboratório de Computação Científica e Análise Numérica, Ufal*

Resumo:
Este trabalho relata o processo de desenvolvimento de uma plataforma de análise dos descritores causais de uma série temporal oriundos da Teoria da Informação. A plataforma visa facilitar a análise dessas séries nos mais variados ramos da ciência. O sistema foi implementado na linguagem de programação R, que além de fornecer ferramentas gráficas, também possui uma grande precisão numérica. Ambas as características de extrema importância ao longo deste trabalho. Após comentar brevemente a respeito de conceitos da Teoria da Informação necessários no processo de análise e modelagem de uma série temporal, expomos os resultados alcançados no decorrer do projeto e sugestões para futuros trabalhos.

Palavras-chave: Séries temporais; teoria da informação; linguagem R.

Apoio financeiro: CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

Trabalho selecionado para a JNIC pela instituição: UFAL.

Introdução:
Séries temporais são conjunto de dados obtidos a partir de um processo observacional ao longo de um determinado período de tempo, não necessariamente dividido em espaços iguais, sendo caracterizadas pela dependência serial existente entre as observações.

A análise de séries temporais tem vasta aplicabilidade na análise de dados bancários e de finanças, na caracterização de redes de computadores e veiculares, na descrição de sinais biológicos, além de inúmeras outras áreas.

A hipótese subjacente a toda análise de séries temporais é que os dados observados são o resultado da operação de um sistema causal sujeito a ruído observacional. Esse sistema, ou dinâmica, é responsável pela criação de padrões através de cuja observação deseja-se inferir a respeito da dinâmica.

A análise de séries temporais é um ramo clássico da Estatística [1] que se divide, tipicamente, na análise no domínio do tempo e no domínio da frequência. Ambas abordagens empregam diretamente os valores observados e, portanto, são suscetíveis ao efeito danoso de diversos tipos de contaminação. Uma forma de tornar as análises mais imunes a contaminação é através de técnicas robustas [2]. Outra, mais modernae que complementa as anteriores, é pelo uso de métodos não-paramétricos.

Há diversas ferramentas que auxiliam na análise clássica de séries temporais; na data de redação deste trabalho havia, para a plataforma R [3], 234 bibliotecas para essa finalidade (ver https://cran.r-project.org/web/views/TimeSeries.html). Para essa mesma plataforma, apenas três bibliotecas trabalham exclusivamente com técnicas não paramétricas.

O projeto aqui relatado tomou como ponto de partida a identificação das necessidades dos pesquisadores que trabalham com estas ferramentas: uma ferramenta gráfica amigável e funcionalidades rápidas, eficientes e numericamente confiáveis. Outro requisito foi o da portabilidade para diversos sistemas operacionais e arquiteturas de hardware, e o uso de ferramentas FLOSS (*Free/Libre Open Source Software*).

Apresentamos, assim, o desenvolvimento de uma ferramenta portável, rápida e de boa qualidade numérica que possibilita análises interativas e

Advisor: Dipl.–Ing. Firstname Lastname, Lehrstuhl für Nachrichtentechnik, TUM, WS 2050/2051.

exploratórias dos dados de uma série temporal através de técnicas provenientes da Teoria da Informação. Com ela, o usuário dispõe de uma conjunto técnicas de análise presentes na literatura para processar e examinar seus dados de modo eficiente e com um mínimo período de aprendizado. A ferramenta é extensível.

Metodologia

A primeira parte do projeto consistiu da apropriação do referencial teórico. Seja a série temporal $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. Ao invés de analisarmos os valores, transformaremos grupos de N valores (não necessariamente adjacentes) e padrões ordinais, e analisaremos a sua distribuição de frequência. Por exemplo e sem perda de generalidade, com $N = 3$ e para qualquer i viável, se $x_i < x_{i+1} < x_{i+2}$ assinaremos a esta tripla o padrão π_0; caso $x_i > x_{i+1} > x_{i+2}$ o padrão será π_1 e assim por diante. Com isso, há $N!$ possíveis padrões. Esta é conhecida como *simbolização de Bandt & Pompe* [4].

Esta simbolização é muito resistente a vários tipos de contaminação, por exemplo, o padrão π_0 não será alterado para qualquer $k > 1$ que afete multiplicativamente x_{i+2}. Ainda que o padrão seja alterado, por exemplo se $k = -1$, a mudança será local e afetará, no máximo, N padrões.

Forma-se, então, um histograma e, a partir dele, extraem-se quantificadores como, por exemplo, entropia, distância estocástica a uma distribuição de equilíbrio, e complexidade estatística.

Seja, assim, $\boldsymbol{h} = (h_1, \ldots, h_{N!})$ o histograma de proporções dos $N!$ padrões observados a partir da série temporal $\boldsymbol{x}$. Calculamos a entropia de Shannon

$$H(\boldsymbol{h}) = \sum_{i=1}^{N!} (-\log h_i) h_i, \qquad (1)$$

com a convenção $-\infty 0 = 0$. A entropia de Shannon é o primeiro elemento a descrever a nossa série temporal. Ela mede a desordem do sistema que deu origem aos dados $\boldsymbol{x}$.

Calculamos logo a distância de Jensen-Shannoon à distribuição uniforme $\boldsymbol{u} = (1/N!, \ldots, 1/N!)$

$$D(\boldsymbol{h}, \boldsymbol{u}) = \sum_{i=1}^{N!} \left(h_i \log \frac{h_i}{u_i} + u_i \log \frac{u_i}{p_i} \right), \qquad (2)$$

em que $u_i = 1/N!$. Esta é uma medida de quão perto ou longe a dinâmica subjacente está de um processo sem informação nenhuma.

Finalmente, calculamos o segundo descritor da nossa série temporal: a sua Complexidade Estatística:

$$C(\boldsymbol{h}, \boldsymbol{u}) = H(\boldsymbol{h}) D(\boldsymbol{h}, \boldsymbol{u}). \qquad (3)$$

Cada série temporal pode então ser descrita por um ponto $(H(\boldsymbol{h}), C(\boldsymbol{h}, \boldsymbol{u}))$. O conjunto de todos os pares $(H(\boldsymbol{h}), C(\boldsymbol{h}, \boldsymbol{u}))$ para qualquer série temporal descrita por padrões de comprimento N jaz em um subconjunto compacto $\mathbb{R}^2$: o plano Entropia-Complexidade. A forma desse conjunto pode ser calculada (ver Fig. 1).

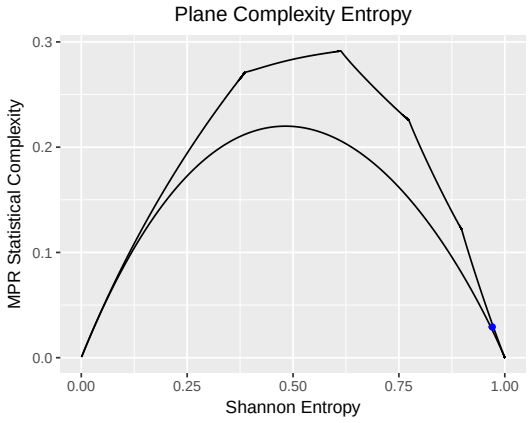

Figura 1. Representação da série no plano Entropia-Complexidade.

A localização do ponto obtido a partir da série no plano Entropia-Complexidade é um quantificador da dinâmica que produziu os dados. +Dois exemplos importantes são os pontos $(0, 0)$, que corresponde a dinâmicas totalmente determinísticas, e $(0, 1)$, que identifica ruído branco. Pontos interiores do plano medem processos em que se misturam dinâmicas determinísticas, caóticas e aleatórias [5].

Embora aqui relatemos apenas o uso da entropia de Shannon e da distância de Jensen-Shannon, o sistema oferece outras entropias [6] e distâncias estocásticas [7]. Com essa contribuição do nosso sistema, as análises podem ser enriquecidas por outros descritores.

A análise da dinâmica subjacente a uma série temporal utilizando a simbolização de Bandt & Pompe tem sido usada com sucesso em diversas áreas como, por exemplo, a discriminação entre fenómenos estocásticos e caóticos [8], a identificação de padrões de comportamento em redes veiculares [9], a classificação e verificação de assinaturas *online* [10], na análise da robustez de redes [11], e a classificação de padrões de consumo de energia elétrica [12]. Tal como antecipamos, o objetivo deste

trabalho é uma ferramenta de apoio a pesquisas como essas.

Durante o desenvolvimento deste trabalho foram estudadas diversas técnicas de análise de séries temporais, com foco nas ferramentas disponíveis na plataforma R. Após o período inicial de aprendizagem, seguido do levantamento dos requisitos do software, foi iniciada a implementação em R, usando o software livre de desenvolvimento integrado RStudio Desktop.

Resultados e Discussões

O sistema for projetado e desenvolvido de forma modular, composto pelas seguintes unidades:

- Módulo de simbolização;
- Módulo de análise;
- Modulo de visualização e interação;

Esses módulos foram desenvolvidos seguindo um cronograma, e depois passaram pelas seguintes etapas:

- Integração dos módulos em um sistema;
- Teste e validação do sistema;
- Geração da interface gráfica.

Para aumentar a aplicabilidade do sistema, permite-se a tanto a geração de séries quanto a leitura de dados em vários formatos (TXT, CSV ou XLSX), e o usuário a seguir escolhe:

- Gerar o gráfico da série (ver Figura 1);
- Calcular diversos tipos de Entropia;
- Calcular diversos tipos de Distâncias Estocásticas;
- Calcular complexidades estatísticas;
- Gerar o histograma de padrões (ver Figura 2);
- Identificar o ponto característico da série no plano Entropia-Complexidade (ver Figura 3)

Um elemento original do sistema é a vinculação entre o histograma de padrões e a série temporal. Escolhendo um ou mais elementos do histograma, os valores correspondentes na série temporal aparecem realçados. Esta funcionalidade permite a análise visual da distribuição temporal dos padrões, possibilitando futuramente a realização de outros testes.

O teste e a validação do sistema são tarefas contínuas, bem como o desenvolvimento de novas funcionalidades.

Um elemento original do sistema é a vinculação entre o histograma de padrões e a série temporal. Escolhendo um ou mais elementos do histograma, os valores correspondentes na série temporal aparecem realçados (ver Figura 4). Esta funcionalidade

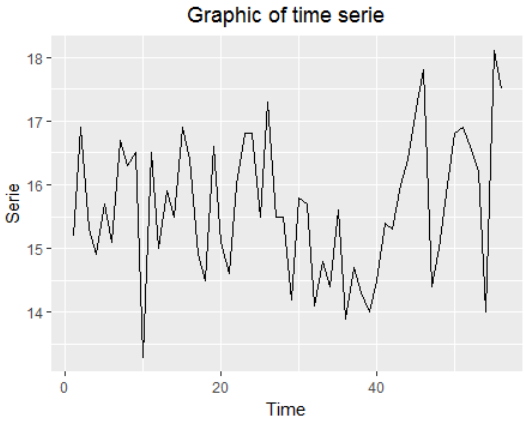

Figura 2. Gráfico de uma série temporal de produção anual de cevada por acre.

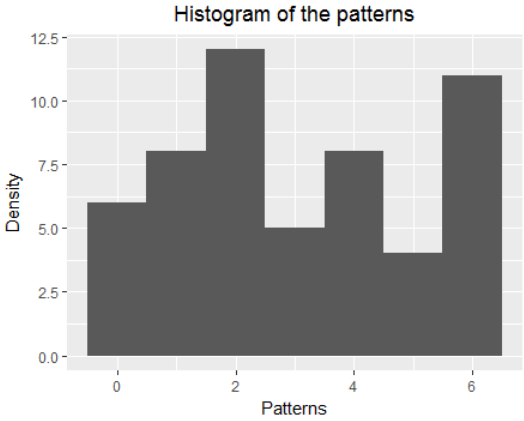

Figura 3. Histograma de densidade dos padrões formados na série.

permite a análise visual da distribuição temporal dos padrões, possibilitando futuramente a realização de outros testes.

O teste e a validação do sistema são tarefas contínuas, bem como o desenvolvimento de novas funcionalidades.

Conclusões

Através do desenvolvimento de tal plataforma por meio da linguagem R, fornecemos a base de geração de inúmeros outros modelos que tenham como objetivo a implementação de sistemas confiáveis que tornem mensuráveis as variadas propriedades presentes na teoria da informação, facilitando não apenas o estudo de séries temporais, como também todo o ramo atuante de análise de dados estatísticos.

Referências

[1] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer-Verlag, Berlin, 2 edition, 1991.

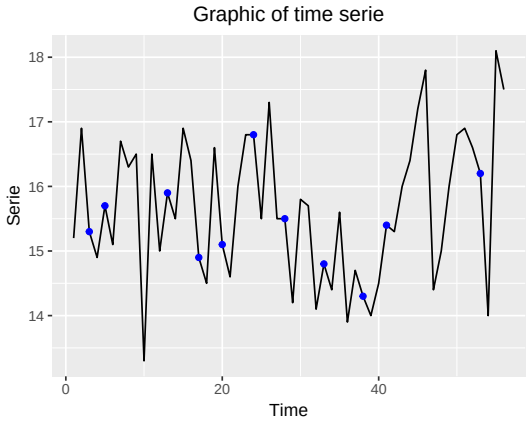

Figura 4. Ilustração da representação na série dos pontos iniciais correspondentes ao padrão $\pi = (2, 1, 3)$

[2] O. Bustos and R. Fraiman. *Robust and nonlinear time series*, volume 29 of *Lectures Notes in Statistics*, chapter Asymptotic behavior of the estimates based on residual autocovariances for ARMA models, pages 26–49. Springer, New York, 1984.

[3] M. Almiron, B. L. Vieira, A. L. C. Oliveira, A. C. Medeiros, and A. C. Frery. On the numerical accuracy of spreadsheets. *Journal of Statistical Software*, 34(4):1–29, 2010.

[4] C. Bandt and B. Pompe. Permutation entropy: A natural complexity measure for time series. *Physical Review Letters*, 88:174102–1–174102–4, Apr 2002.

[5] O. A. Rosso, H. A Larrondo, M. T. Martín, A. Plastino, and M. A. Fuentes. Distinguishing noise from chaos. *Physical Review Letters*, 99:154102, 2007.

[6] M. Salicrú, M. L. Mendéndez, and L. Pardo. Asymptotic distribution of (h, ϕ)-entropy. *Communications in Statistics - Theory Methods*, 22(7):2015–2031, 1993.

[7] L. Pardo. *Statistical Inference Based on Divergence Measures*. Number 185 in Statistics, textbooks and monographs. Chapman & Hall/CRC, Boca Raton, 2006.

[8] M. G. Ravetti, L. C. Carpi, B. A. Gonçalves, A. C. Frery, and O. A Rosso. Distinguishing noise from chaos: objective versus subjective criteria using Horizontal Visibility Graph. *PLOS One*, 9(9):1–15, 2014.

[9] A. L. L. Aquino, T. S. G. Cavalcante, E. S. Almeida, A. C. Frery, and O. A. Rosso. Characterization of vehicle behavior with information theory. *The European Physical Journal B: Condensed Matter and Complex Systems*, 88(10):257–269, Oct 2015.

[10] O. A. Rosso, R. Ospina, and A. C. Frery. Classification and verification of handwritten signatures with time causal information theory quantifiers. *PLoS ONE*, 11(12):e0166868, 2016.

[11] T. A. Schieber, L. Carpi, A. C. Frery, O. A. Rosso, P. M. Pardalos, and M. G. Ravetti. Information theory perspective on network robustness. *Physics Letters A*, 380:359–364, 2016.

[12] A. L. L. Aquino, H. S. Ramos, A. C. Frery, L. P. Viana, T. S. G. Cavalcante, and O. A. Rosso. Characterization of electric load with information theory quantifiers. *Physica A*, 465:277–284, 2017.