XXI ENMC
ENCONTRO NACIONAL DE MODELAGEM COMPUTACIONAL
IX ECTM
ENCONTRO DE CIÊNCIA E TECNOLOGIA DE MATERIAIS

08 a 11 de Outubro de 2018
Instituto Federal Fluminense
Búzios - RJ

ANÁLISE TOPOLÓGICA DE DADOS PARA CARACTERIZAÇÃO DE PERIODICIDADE EM SÉRIES TEMPORAIS DE DADOS PLUVIOMÉTRICOS

Marcella Feitosa dos Santos[1,2] - marcella.feitosa.ufrpe@gmail.com
Marcelo Amorim[2] - marc.amorim@gmail.com
Wilson Rosa de Oliveira[2] - wilson.rosa@ufrpe.br
Tatijana Stosic[2] - tastosic@gmail.com
[1] Instituto Federal de Educação, Ciência e Tecnologia Baiano - BA, Brasil
[2] Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - PE, Brasil

Abstract. *O presente trabalho traz a aplicação de métodos desenvolvidos na Análise Topológica de Dados e propõe uma nova abordagem para classificar séries temporais contendo dados de precipitação. Especificamente aplicamos o método "Sliding Windows Embeddings" e "Maximum Persistence", que combina topologia persistente e mergulhos de janelas móveis para caracterizar e criar um ranking da periodicidade de séries temporais. Comparamos os resultados obtidos com os resultados da aplicação do método Sample Entropy, que mede a taxa de geração de novas informações examinando séries temporais. Quando a entropia é alta dizemos que o fenômeno é de alta complexidade, isto é, trata-se de uma difícil predição. É esperado que em séries com alta entropia tenhamos menor periodicidade e foi exatamente o que encontramos, pudemos caracterizar através do "Score" quais séries, e portanto regiões do estado de Pernambuco, tem regime de chuva menos periódico, coincidindo com regime de menor complexidade.*

Keywords: *Topologia Algébrica, Homologia Persistente, Diagrama de Persistência*

1. INTRODUÇÃO

Uma característica importante da ciência moderna é que dados de vários tipos estão sendo produzidos a uma taxa sem precedentes. Isto é proporcionado devido aos novos métodos experimentais e também pelo aumento na disponibilidade da tecnologia de alta potência para gerá-los. Nessa perspectiva, se faz necessário utilizar e criar novos métodos para a análise dos dados, mais ainda, métodos que possam lidar com os problemas da alta dimensão e da grande quantidade dos dados, assim como a velocidade com que são obtidos e/ou produzidos, além de seus diferentes tipos e formatos (SNÁŠEL, 2017; CARLSSON, 2009). Tais dados são chamados Big Data e possuem como características: volume, velocidade, variedade, veracidade e valor. Volume refere-se ao tamanho dos dados para processamento e análise. Velocidade relaciona-se com a taxa de crescimento e uso desses dados. Variedade significa os diferentes

tipos e formatos utilizados para processamento e análise. Veracidade diz respeito à precisão dos resultados e análise dos dados. Valor, o que é acrescentado e a contribuição oferecida pelo processamento e análise de dados (SNÁŠEL, 2017; CHEN, CHIANG e STOREY, 2012).

Geometria e Topologia são ferramentas naturais, em certo sentido, para esse tipo de análise, pois é possível considerar a geometria como o estudo das funções de distância e a topologia como o estudo da forma, no sentido que as propriedades dos objetos geométricos não dependem das coordenadas escolhidas. Essa insensibilidade à métrica é útil no estudo de situações em que a métrica é entendida apenas de forma grosseira. Por exemplo, no contexto biológico, noções de distância são construídas usando algumas medidas intuitivamente atraentes de similaridade, como o algoritmo BLAST (*Basic Local Alignment Search Tool*) ou seus similares, que permite que o investigador compare uma sequência de consulta com uma biblioteca (ou banco de dados de sequências), e identifique as sequências da biblioteca que se assemelham à sequência de consulta, considerando um determinado limiar (CARLSSON, 2009).

Para a análise de *Big Data* utilizar a metodologia de criação de modelo, simulação e então avaliação, não é viável. O processo descrito é útil e adequado para resolver os problemas clássicos, como os problemas físicos, porque a base teórica para tais problemas tem sido pesquisada e compreendida o suficiente de modo que pode ser reconstruída para ajustar o modelo criado. Para o processamento de Big Data o primeiro problema é definir uma hipótese concreta das características dos dados que podem ser testadas (SNÁŠEL, 2017). Assim, o principal objetivo da pesquisa não é um modelo, mas ser capaz de obter características interessantes do conjunto de dados. É nesse sentido que se justifica utilizar a topologia como ferramenta que possibilite uma visão geral da organização dos dados e verificar que há regiões de interesse, afinal os dados podem estar estruturados em formas que não são fáceis de capturar utilizando métodos tradicionais. Aqui podemos pensar na forma da organização como uma nuvem de pontos amostrados de uma variedade de dimensão n, para então estudar as propriedades do conjunto de pontos sob a perspectiva da topologia.

2. MATERIAIS E MÉTODOS

Considerando sua localização, vegetação e topografia, o estado de Pernambuco tem condições climáticas diversificadas, apresentando dois tipos climáticos, precipitação no outono e inverno em parte do litoral e semiárido quente no sertão, com médias de temperatura acima de $18°C$, (SILVA, 2011). O estado pernambucano é subdividido em cinco mesorregiões: Metropolitana do Recife, Zona da Mata, Agreste, Sertão e São Francisco, (SILVA, 2011).

Os dados que utilizamos foram obtidos através do Laboratório de Meteorologia de Pernambuco (LAMEP), órgão que pertence ao Instituto Tecnológico de Pernambuco (ITEP). As séries analisadas são registros históricos de precipitação mensal ferentes ao período Janeiro de 1950 à Dezembro de 2012, adquiridos de seis pontos pluviométricos, que estão distribuídos em três das Mesorregiões do estado de Pernambuco. Foram escolhidas duas estações da Região Metropolitana, duas estações do Sertão do São Francisco e duas estações do Sertão Pernambucano.

Para diminuir a intervenção da sazonalidade nas séries (também chamadas de anomalias) utilizaremos a transformação proposta por Costa (2005), que retira a tendência anual. Tal transformação é dada por:

$$\tilde{X}_{ij} = \frac{(X_{ij} - \overline{X}_i)}{\sigma_i} \tag{1}$$

Em que, X_{ij} é a i-ésima observação mensal no j-ésimo ano; i é o indicativo do mês independente do ano; j é o indicativo do ano; X_i é a média amostral do i-ésimo mês ao longo dos anos; σ_i é o desvio padrão amostral do i-ésimo dia ao longo dos anos.

Para aplicação dos métodos e geração das imagens utilizamos os softwares:

- R 3.5, pacotes **TDA**, **pracma**, **zoo**;
- Python 2.7.

2.1 Análise Topológica de Dados

O objetivo básico é aplicar a topologia, para desenvolver ferramentas para estudar características geométricas de dados. Chamamos "dados" um conjunto finito de pontos no espaço. Em geral, o espaço em que os pontos se encontram são de dimensão elevada, para entender a ideia da utilização nesse tipo de análise podemos pensar nos dados como pertencentes ao espaço bidimensional ou tridimensional. O que se pretende com a Análise Topológica de Dados (*Topological Data Analysis*) é criar um resumo ou uma representação comprimida de todas as características dos dados para ajudar a desvendar padrões e relacionamentos existentes no conjunto dados. O formalismo matemático que foi desenvolvido para a incorporação de técnicas geométricas e topológicas, lida com nuvens de pontos, ou seja, conjuntos finitos de pontos, (CARLSSON, 2009). Para tanto, são adaptadas ferramentas Topologia Algébrica para o estudo destes conjuntos que são amostras finitas, tomadas a partir de um objeto geométrico, talvez com ruído. A topologia fornece uma linguagem formal para a matemática qualitativa, onde as relações de proximidade (ou vizinhança) são estudadas, sem o uso de distâncias. A ideia de construção de resumos das características dos dados envolve a compreensão da relação entre objetos topológicos e geométricos (SNÁŠEL, 2017).

2.2 Homologia Persistente e Diagrama de Persistência

A extração de informações de bancos de dados de alta dimensão, incompletos e com ruídos é um desafio geral e com essa nova abordagem podemos contribuir com a superação de tais dificuldades, devido à **funtorialidade**, considerada uma das chaves para a matemática moderna por sua natureza topológica. O principal conceito utilizado é o de homologia persistente, um conceito modificado de grupo de homologia. Segundo Hatcher (2002), a Topologia Algébrica pode ser definida como o estudo das técnicas para a formação de imagens algébricas de espaços topológicos. Na maioria das vezes essas imagens algébricas são grupos, mas as estruturas mais elaboradas, tais como anéis, módulos e álgebras também surgem. Os mecanismos que criam essas imagens - as "lanternas" da topologia algébrica - são conhecidas formalmente como funtores e têm a característica de formar imagens não só de espaços, mas também de mapas. Assim, os mapas contínuos entre os espaços são projetados sobre homomorfismos entre suas imagens algébricas, então espaços topologicamente relacionados têm imagens algebricamente relacionadas. Com lanternas adequadamente construídas espera-se ser capaz de formar imagens com detalhe suficiente para reconstruir com precisão as formas de todos os espaços, ou pelo menos grandes e interessantes classes de espaços.

A ideia é que para uma nuvem de pontos, temos uma variedade subjacente da qual os pontos da nuvem são uma amostra capaz de descrever as propriedades geométricas e características topológicas da variedade. As informações topológicas consideradas são as componentes conexas (número de partes que a nuvem possui ou *clusters*), números de furos e noção

similar para dimensão superior. Através da homologia persistente conseguimos computar tais características, marcando nascimento e morte de cada uma delas, quando "engordamos" cada ponto da nuvem, o que persiste à medida que esse espessamento é realizado é o que consideramos como característica pertencente à nuvem analisada.

Podemos pensar no diagrama de persistência como o marcador do nascimento e morte de cada característica topológica que apareceu/desapareceu/fundiu à medida que o espessamento foi feito. Topologicamente uma circunferência possui como características uma única componente e um furo. Já com duas circunferências podemos ter diferentes configurações, para o número de componentes podemos ter uma ou duas componentes; para o número de furos podemos ter um, dois ou três. Nas Figura 1 temos diferentes espaços e seus respectivos diagramas de persistência.

2.3 Mergulhos de Janelas Móveis e Persistência Máxima

Para caracterização da periodicidade em séries temporais, a novidade trazida na abordagem de Perea (2015) é estabelecer a *persistência máxima* como uma medida de quão redonda é uma nuvem de pontos após o mergulho em dimensão superior das janelas móveis, ele demonstra que isso ocorre quando o tamanho da janela corresponde à frequência natural do sinal. Isso significa que a persistência de dimensão 1 é um quantificador efetivo de periodicidade e pode ser usada para inferir propriedades do sinal, (PEREA, 2015). Em um diagrama de persistência, denotado por dgm, sejam $\mathbf{x} = (x, y) \in dgm$, define-se $pers(x, y) = y - x$ para $(x, y) \in \mathbb{R}^2$, a máxima persistência, denotada por mp é dada por:

$$mp(dgm) = \max_{\mathbf{x} \in dgm} pers(\mathbf{x}) \tag{2}$$

Para a quantificar a periodicidade estamos computando o *Score da periodicidade*, dado por:

$$\frac{mp(dgm\ (X_s))}{\sqrt{3}} = Score(S) \tag{3}$$

3. RESULTADOS E DISCUSSÃO

Apresentamos, discutimos e comparamos os resultados obtidos utilizando a Máxima Persistência, com os resultados obtidos com a utilização da Sample Entropy para dados mensais de precipitação dos seis postos pluviométricos, distribuídos em três Mesorregiões do estado de Pernambuco. Também apresentamos os valores das médias e desvios padrão para cada uma das séries temporais sem anomalias. Na Figura 2 apresentamos as séries temporais originais da precipitação, assim como as novas séries obtidas quando retiramos as anomalias. Registramos na Tabela 1 as estatísticas descritivas para cada série.

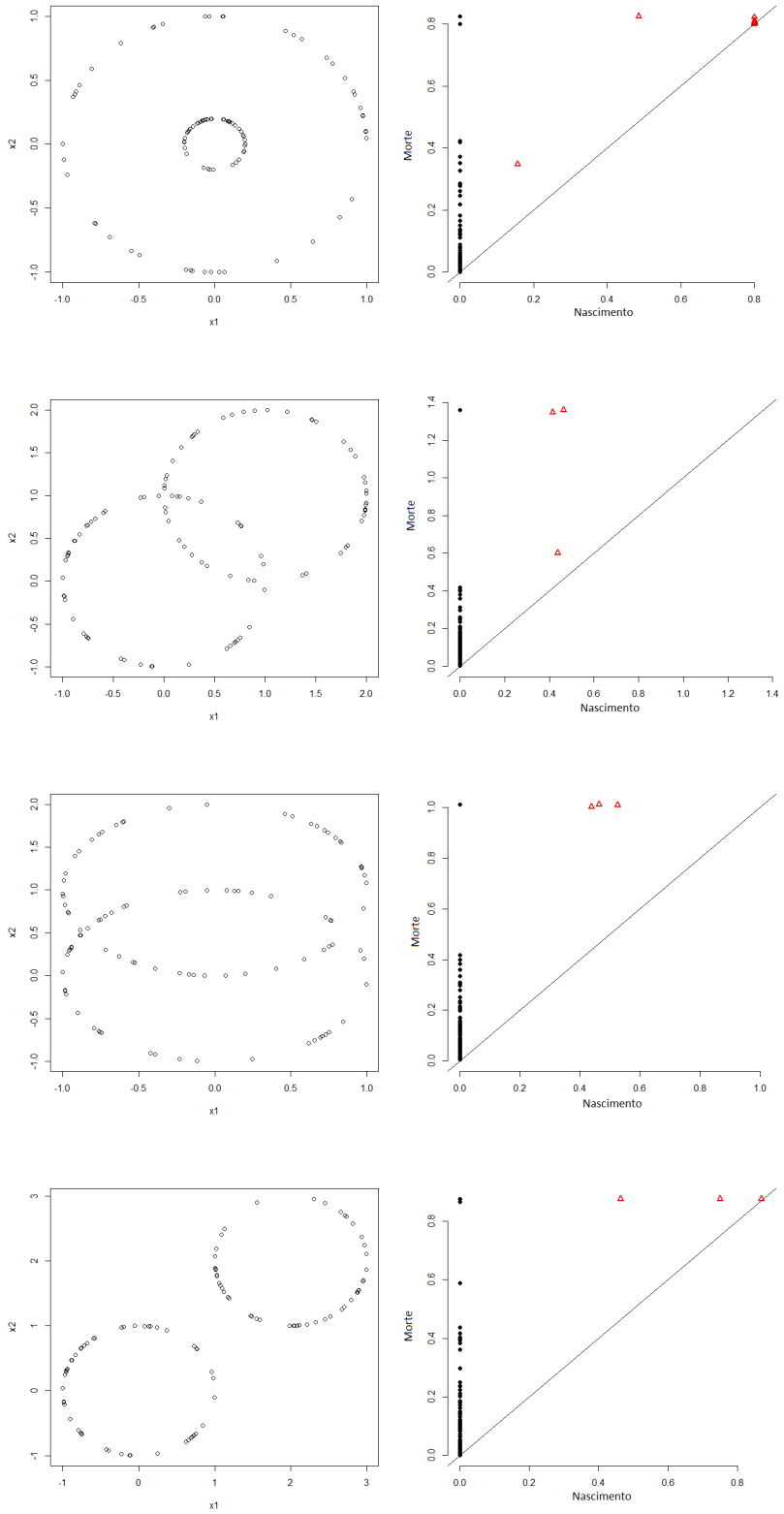

Figure 1- À esquerda temos os pontos amostrados de circunferência com diferentes posições relativas e à direita o respectivo diagrama de persistência. No diagrama de persistência os pontos em preto contam as componentes conectadas e os triângulos em vermelhos computam os furos de dimensão 1.

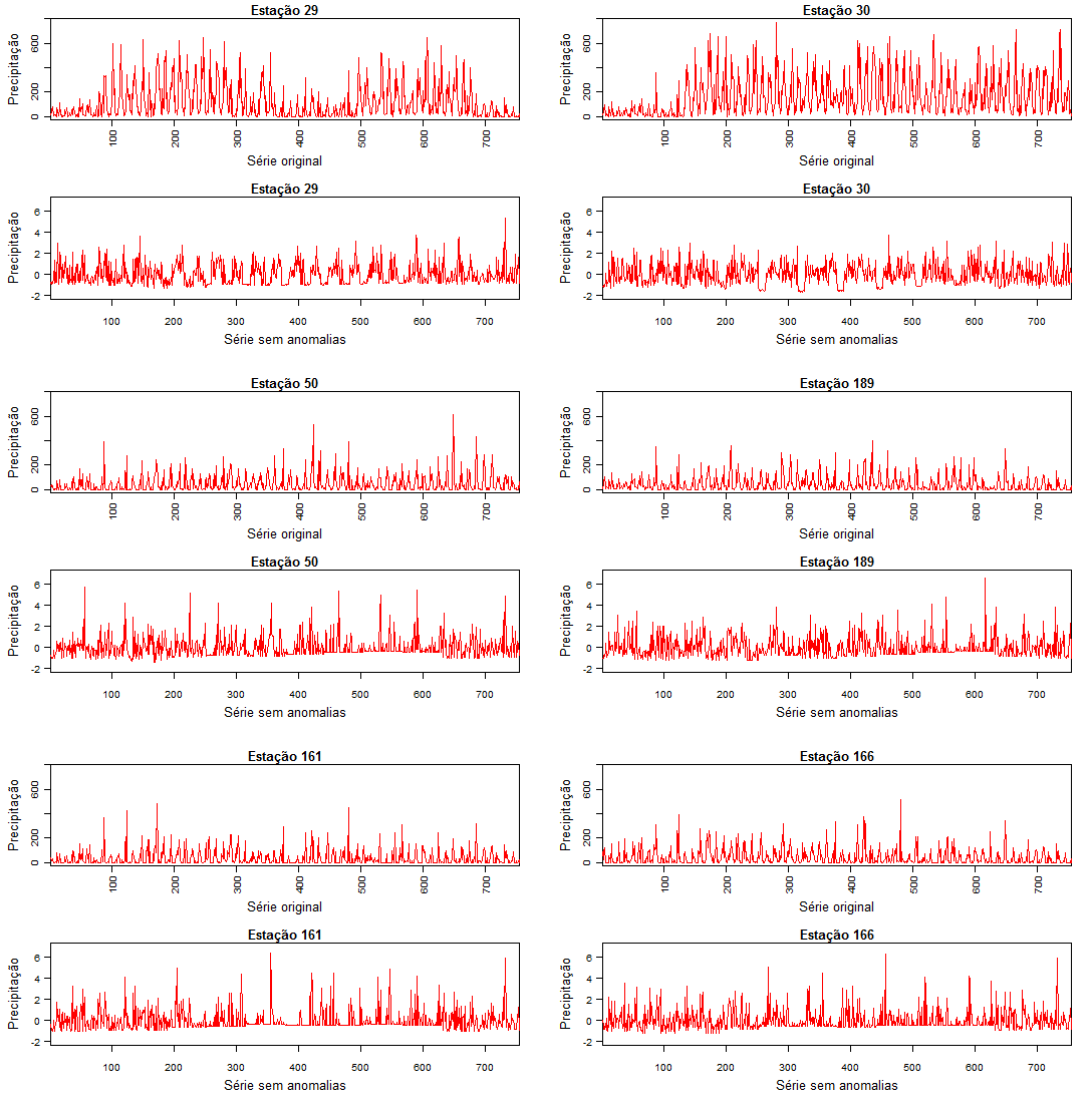

Figure 2- Séries temporais da precipitação das estações pluviométricas, de Janeiro de 1950 à Dezembro de 2012.

É perceptível que as séries originais possuem uma variabilidade considerável, que está concentrada em torno de 300 mm com a presença de picos acentuados. Já as estações da Região Metropolitana possuem uma variabilidade diferente das demais, com concentração em torno de 400 mm.

Table 1- Estatística Descritiva

Localização/Estação	Média	Desvio Padrão	Coeficiente de Variação
Região Metropolitana - Estação 29	111.9673	137.1638	1.225034
Região Metropolitana - Estação 30	164.9812	159.9831	0.9697048
Sertão Pernambucano - Estação 50	47.488	70.38422	1.482148
Sertão Pernambucano - Estação 189	44.68978	63.4902	1.420687
Sertão do São Francisco - Estação 161	42.58101	65.3853	1.535551
Sertão do São Francisco - Estação 166	34.63313	60.30561	1.741269

Na Tabela 2 apresentamos o *Score da Periodicidade* para cada série, destaque para a Estação 30 que possui maior Score, maior Média mensal e menor Coeficiente de Variação, isto é, o regime de chuvas é mais periódico na região que mais chove e a variabilidade em relação à média é a menor dentre as demais Mesorregiões. Temos situação análoga para a Estação 29, também situada na Região Metropolitana, tendo o segundo maior dos Scores. Não é surpresa a Estação 189, localizada no Sertão Pernambucano, ter obtido o menor Score, devido à sua a Média mensal baixa a variabilidade que não é das menores. Destacamos nesse sentido o que também ocorreu com a Estação 166, com menor Média e maior Coeficiente de Variação, possui Score mais baixo que os demais.

Table 2- Score da Periodicidade

Localização	Score
Região Metropolitana - Estação 29	0.540186
Região Metropolitana - Estação 30	0.583412
Sertão Pernambucano - Estação 50	0.473161
Sertão Pernambucano - Estação 189	0.233717
Sertão do São Francisco - Estação 161	0.368799
Sertão do São Francisco - Estação 166	0.396345

Na Tabela 3 apresentamos a comparação entre o Score da Periodicidade e os valores de Sample Entropy obtidos para cada série. O que ocorreu na Estação 189 corrobora com nossa suspeita inicial, a menor periodicidade é acompanhada do alto valor da entropia, nos informando que o regime de chuvas na região não é periódico e é complexo, isto é é mais irregular. Outro aspecto que chama a atenção são os maiores scores obtidos, pertencentes às Estações 29 e 30 ambas situadas na Região Metropolitana, mas com valores da entropia bem diferentes, significando que na Estação 29 o regime de chuvas possui periodicidade maior e menor grau de aleatoriedade, enquanto na Estação 30 temos uma periodicidade marcada por elevado grau de complexidade.

Table 3- Periodicidade *versus* Complexidade

Localização	Score	SampEn
Região Metropolitana - Estação 29	0.540186	0.7939959
Região Metropolitana - Estação 30	0.583412	1.393248
Sertão Pernambucano - Estação 50	0.473161	0.9701715
Sertão Pernambucano - Estação 189	0.233717	1.280525
Sertão do São Francisco - Estação 161	0.368799	1.008903
Sertão do São Francisco - Estação 166	0.396345	0.6770704

4. CONCLUSÕES

Apresentamos resultados preliminares para a caracterização da periodicidade de séries temporais em dados de precipitação. Foram utilizadas séries pertencentes a diferentes Mesorregiões do estado de Pernambuco. Nas estações localizadas na Região Metropolitana tivemos os maiores *scores*, informando que nessa Mesorregião o regime de chuvas é mais periódico do

que na Mesorregião Sertão Pernambucano, onde obtivemos o menor *score*. Fizemos ainda a comparação do Score de Periodicidade com os valores de Sample Entropy para cada série e pudemos observar como as duas informações a respeito de cada série podem se complementar para o entendimento do regime de chuvas nas diferentes Mesorregiões: Na Região Metropolitana a periodicidade é maior em ambas as séries, com entropia caracterizando diferente complexidade; No Sertão Pernambucano encontramos a menor periodicidade e uma das maiores entropias, isto é, o regime de chuvas na região além de menos periódico dentre todos é o segundo de menor regularidade. Trazemos na Tabela 4 um *ranking* das séries utilizando o *Score da Complexidade* em ordem decrescente, e o valor da *Sample Entropy*, em ordem crescente.

Table 4- Ranking: Periodicidade *versus* Complexidade

Localização	Score	Localização	SampEn
Estação 189	0.233717	Estação 30	1.393248
Estação 161	0.368799	Estação 189	1.280525
Estação 166	0.396345	Estação 161	1.008903
Estação 50	0.473161	Estação 50	0.9701715
Estação 29	0.540186	Estação 29	0.7939959
Estação 30	0.583412	Estação 166	0.6770704

A partir da obtenção desses primeiros resultados, trabalharemos para ampliar a quantidade de séries do estado de Pernambuco e criar um *rank* maior, para enxergarmos como a periodicidade do regime de chuvas é caracterizado. Também consideramos a possibilidade de comparar a metodologia da Máxima Persistência com outras metodologias, além da Sample Entropy.

Agradecimentos

Ao Professor Doutor Samuel Sousa, por auxiliar na indicação dos dados utilizados no presente trabalho e na orientação prestada para a análise dos dados utilizando a Entropia quando contribuiu com a disciplina Métodos Quantitativos Aplicados às Ciências Agrárias no semestre 2017.1, no Programa de Pós-Graduação em Biometria e Estatística Aplicada (PPGBEA) da Universidade Federal Rural de Pernambuco. O terceiro autor agradece o financiamento de pesquisa do CNPq processos números 421849/2016-9 e 310086/2015-9.

REFERÊNCIAS

Hatcher, A. (2002), *"Algebraic topology*", Cambridge University Press.

Carlsson, G. (2009), "Topology and data", *"Bulletin of the American Mathematical Society*", 46(2), 255-308.

Snášel, V., Nowaková, J., Xhafa, F., & Barolli, L. (2017), "Geometrical and topological approaches to Big Data", *"Future Generation Computer Systems*", 67, 286-296.

Richman, Joshua S.; Moorman, J. Randall. (2000), "Physiological time-series analysis using approximate entropy and sample entropy", *"American Journal of Physiology-Heart and Circulatory Physiology*", v. 278, n. 6, p. H2039-H2049.

Costa, M.; Goldberger, A. L.; Peng, C.-K. (2005), "Multiscale entropy analysis of biological signals", *"Physical review E*", v. 71, n. 2, p. 021906.

da Silva, A. O., de Albuquerque Moura, G. B., de França, Ê. F., Lopes, P. M. O., da Silva, A. P. N. (2011). "Análise espaço-temporal da evapotranspiração de referência sob diferentes regimes de precipitações em Pernambuco", *"Revista caatinga*", 24(2), 135-142.

Perea, J. A., & Harer, J. (2015), "Sliding windows and persistence: An application of topological methods to signal analysis", *"Foundations of Computational Mathematics*", 15(3), 799-838.
Chen, H., Chiang, R. H., & Storey, V. C. (2012), "Business intelligence and analytics: from big data to big impact", "*MIS quarterly*", 1165-1188.

TOPOLOGICAL DATA ANALYSIS FOR CHARACTERIZATION OF PERIODICITY IN TIMES SERIES OF PLUVIOMETRIC DATA

Abstract. *The present work presents the application of methods developed in Topological Data Analysis and proposes a new approach to classify time series containing precipitation data. Specifically we apply the "Sliding Windows Embeddings" and "Maximum Persistence" methods, which combines persistent topology and sliding window embeddings for characterizing and creating a ranking of the periodicity of time series. We compared the results obtained with the application of the Sample Entropy method, which measures the rate of generation of new information by examining time series. When the entropy is high we say that the phenomenon is highly complex, it is difficult to predict. It is expected that in series with high entropy we have less periodicity and it was exactly what we have found, we were able to characterize through the "Score", which series, and therefore regions of the state of Pernambuco, has a less periodic rain regime, coinciding with a regime of lesser complexity.*

Keywords: *Algebraic Topology, Persistent Homology, Persistence Diagram.*