Estudo da Influência de diversas Medidas de Similaridade na Previsão de Séries Temporais utilizando o Algoritmo kNN-*TSP*

Jorge Aikes Junior[1], **Huei Diana Lee**[1], **Carlos Andrés Ferrero**[1], **Feng Chung Wu**[12]

[1]Laboratório de Bioinformática – LABI – Universidade Estadual do Oeste do Paraná
Parque Tecnológico Itaipu – PTI
Caixa Postal 39, 85869-970 – Foz do Iguaçu, PR, Brasil

[2]Faculdade de Ciências Médicas – Universidade Estadual de Campinas
Pós-Graduação em Ciências da Cirurgia
CEP 13083-887 – Campinas, SP, Brasil

{jorgeaikes, hueidianalee, anfer86, wufengchung}@gmail.com

Abstract. *This paper presents a study on the influence of different similarity measures in time series forecasting on artificial data with seasonal and chaotic features and real time series related to transport's flow. The results show that the L_p Norm's measures, especially the Manhattan one, may present an advantage over the other evaluated measures, due to, in general, a greater accuracy in the temporal data forecasting using the kNN-TSP algorithm, as also as for its lower computational cost.*

Resumo. *Este trabalho*[1] *apresenta um estudo sobre a influência de diversas medidas de similaridade na previsão de dados em séries temporais artificiais, de características sazonais e caóticas, e séries temporais reais relacionadas a fluxo de transportes. Os resultados demonstram que as medidas da Norma L_p, especialmente a Manhattan, apresentam vantagem sobre as demais medidas avaliadas devido a apresentar, em geral, maior acurácia na previsão de dados temporais utilizando o algoritmo kNN-TSP e ao seu menor custo computacional.*

1. Introdução

As Séries Temporais (ST) podem ser entendidas como qualquer conjunto de observações que se encontram ordenadas no tempo [Morettin e Toloi 2006]. Assim, pode-se definir uma ST Z de tamanho m como um conjunto ordenado de valores, ou seja, $Z = (z_1, z_2, \ldots, z_m)$ onde $z_t \in \Re$ e representa uma observação z em um instante t [Chiu et al. 2003].

A previsão de dados temporais, que busca por meio da exploração de dados conhecidos do passado da série projetar dados futuros dessa série, é uma tarefa que tem atraído a atenção de pesquisadores de diversas áreas do conhecimento. Ao longo do tempo foram desenvolvidas diversas abordagens para a previsão de dados. Como exemplos dessas abordagens têm-se as paramétricas, que assumem que os dados respeitam alguma distribuição conhecida e modelam parâmetros de funções que se ajustem a essa distribuição dos

[1]Agradecimentos à Fundação Parque Tecnológico Itaipu — FPTI-BR — pelo apoio por meio da linha de financiamento de bolsas.

dados; e as não-paramétricas, que buscam modelar o comportamento sem o conhecimento prévio da distribuição dos dados [Morettin e Toloi 2006, Aguirre 2007].

As abordagens não-paramétricas podem ainda ser divididas em globais e locais. As primeiras empregam funções de aproximação global, utilizando toda a série temporal; as segundas dividem a série em conjuntos menores de séries, apresentando funções de aproximação para cada um dos subconjuntos, sendo que a função de aproximação de um subconjunto é válida apenas para esse subconjunto [Karunasinghe e Liong 2006].

Dentre as abordagens para a previsão de séries temporais não-paramétricas de aproximação local existentes, pode-se citar a adaptação do algoritmo de Aprendizagem de Máquina (AM) k-*Nearest Neighbor* (kNN). Neste trabalho foi empregada a adaptação desse algoritmo denominada k-*Nearest Neighbor - Time Series Prediction* (kNN-*TSP*) proposta por Ferrero (2009). No kNN-*TSP* o objetivo é encontrar as k sequências mais similares dentro da série, a partir de uma sequência de referência, e utilizar os valores dessas sequências similares e uma função de previsão, para realizar o cálculo do valor futuro da série.

Nessa abordagem, um dos parâmetros que influencia a acurácia do algoritmo consiste na medida de similaridade selecionada para identificar as sequências similares na série. Assim, determinar a influência desse parâmetro constitui tema de importância. Desse modo, neste trabalho é apresentado um estudo da influência de diversas medidas de similaridade na previsão de dados em diversas séries temporais, tanto artificiais quanto reais, utilizando como método de previsão o algoritmo não-paramétrico kNN-*TSP*.

Este trabalho faz parte do projeto de Análise Inteligente de Dados em uma parceria entre o Laboratório de Bioinformática da Universidade Estadual do Oeste do Paraná (UNIOESTE)/Foz do Iguaçu, o Serviço de Coloproctologia da Faculdade de Ciências Médicas da Universidade Estadual de Campinas (UNICAMP)/Campinas, o Laboratório de Inteligência Computacional do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP)/São Carlos e o Grupo Interdisciplinar em Mineração de Dados e Aplicações da Universidade Federal do ABC (UFABC).

O presente trabalho está organizado do seguinte modo: na Seção 2 são apresentados conceitos de previsão de ST, métodos paramétricos e não-paramétricos de previsão, em especial o algoritmo kNN-*TSP*; na Seção 3 são descritas diversas medidas de similaridade entre ST; na Seção 4 são apresentadas as ST reais e artificiais usadas neste trabalho, bem como a configuração experimental; na Seção 5 é apresentada a avaliação experimental e a discussão dos resultados; e, na Seção 6 são apresentadas as conclusões deste trabalho.

2. Previsão de Séries Temporais

Como mencionado, existem diversos métodos para previsão de ST, que utilizam desde complexos modelos estatísticos a modelos intuitivos e simples, cada um apresentando suas próprias capacidades e limitações. Uma mesma série pode ser analisada e prevista por vários desses métodos. Assim, para uma melhor seleção do método de previsão a ser empregado, é necessário não somente se ter conhecimento do comportamento do fenômeno observado, mas também da natureza e do objetivo da análise bem como do método utilizado [Morettin e Toloi 2006].

Uma das possíveis formas de classificação das abordagens de previsão é a divisão entre métodos paramétricos e não-paramétricos [Aguirre 2007].

2.1. Métodos Paramétricos

Métodos nessa categoria necessitam do conhecimento prévio da distribuição dos dados para a estimação de seus parâmetros. Entre esses métodos, podem ser destacados: Auto-regressivos (AR), Médias Móveis (MA), Auto-regressivos de Médias Móveis (ARMA), Auto-regressivos de Médias Móveis Integrados (ARIMA) e Auto-regressivos de Médias Móveis Integrados Sazonais (SARIMA) [Morettin e Toloi 2006, Aguirre 2007].

Dentre esses métodos, um que apresenta grande destaque é o modelo ARIMA. Esse modelo pode, em teoria, descrever séries temporais de qualquer natureza já que ele está preparado para lidar com ST não-estacionárias. O modelo ARIMA de ordem (h, d, q), ARIMA(h, d, q), é descrito pela Equação 1 [Morettin e Toloi 2006, Aguirre 2007]:

$$z_t = \phi_1 M_{t-1} + \phi_2 + M_{t-2} + \ldots + \phi_h M_{t-h} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q} \quad (1)$$

onde $M_t = z_t - z_{t-d}$; d representa o grau do operador de diferença; ϕ_h e θ_q são os parâmetros dos processos Auto-regressivos e de Médias Móveis, possuindo ordem h e q respectivamente (ARMA(h, q)); e e_t corresponde ao elemento "inovativo" que não pode ser explicado pelo modelo.

Para a utilização desse modelo, supõe-se que a d-ésima diferença entre as observações da ST pode ser representada por um processo estacionário capaz de ser estimado por um modelo ARMA. Desse modo, séries que apresentam tendência não-explosiva, ou seja, não-estacionaridade homogênea, bem como séries estacionárias podem ser descritas por esse modelo.

2.2. Métodos Não-Paramétricos: O algoritmo *kNN-TSP*

Os métodos não-paramétricos não necessitam do conhecimento prévio da distribuição dos dados. Dentre os exemplos que podem ser destacados estão as Redes Neuronais Artificiais (RNA) e variações do algoritmo de aprendizagem de máquina k-*Nearest Neighbor* (*kNN*) [Haykin 1999, Han e Kamber 2006, Ferrero 2009].

O *kNN* é um algoritmo de aprendizagem supervisionada que consiste em encontrar, segundo alguma medida de similaridade, os *k* exemplos mais próximos de um exemplo ainda não-rotulado e, baseado nos rótulos desses *k* exemplos próximos, rotular o novo exemplo [Han e Kamber 2006].

Desse modo, quando se utiliza, por exemplo, 1-*NN*, o novo exemplo recebe o mesmo rótulo da classe do vizinho mais próximo encontrado; caso sejam considerados $k > 1$ vizinhos, deve-se então definir como será determinado o rótulo do novo exemplo. Uma das abordagens mais simples consiste em utilizar a classe majoritária, ou seja, a classe predominante entre esses k exemplos. Outra possível abordagem é a utilização de pesos para cada um dos vizinhos próximos de acordo com algum critério, como a proximidade desse vizinho. Assim, é possível perceber que a decisão de quantos vizinhos próximos devem ser considerados para a classificação, pode exercer influência no resultado do funcionamento do algoritmo. Esse valor é particular a cada problema, levando à necessidade de avaliação dos possíveis valores de k a serem considerados.

O algoritmo *kNN* foi adaptado em Ferrero (2009) para utilização com ST, resultando no algoritmo k-*Nearest Neighbor - Time Series Prediction* (*kNN-TSP*). Ele busca

estimar um valor z_{t+1} de uma ST Z, utilizando os valores anteriores dessa ST, ou seja, $z_t, z_{t-1}, z_{t-2}, \ldots, z_{t-m+1}$ onde m corresponde à quantidade de valores do passado da ST a serem considerados. Para esse cálculo utilizam-se as k sequências de tamanho w mais próximas à sequência final da ST, sendo essa denominada de sequência de referência. As sequências próximas são selecionadas através de alguma medida de similaridade.

Um exemplo da aplicação do algoritmo *kNN-TSP* para a previsão da ST de Mackey-Glass [McNames 1999] é apresentado na Figura 1. Nessa figura, a linha mais clara representa os valores reais observados na ST; a linha de cor preta, a sequência dos cinco últimos valores ocorridos; a linha de cor cinza escuro com asteriscos, as sequências similares encontradas pelo algoritmo; e o quadrado, o valor estimado calculado pela função de previsão. Nesse exemplo, deseja-se prever o último valor da série, baseado na última sequência de cinco pontos ($w = 5$), em uma medida de similaridade e duas sequências similares. Baseado no valor do ponto futuro dessas duas sequências similares foi calculado o valor previsto para o final da série.

Como pode ser observado, a previsão com o algoritmo *kNN-TSP* depende de alguns parâmetros os quais são descritos a seguir e ilustrados pela Figura 1: **(a) Tamanho *w* da janela para extrair as sequências**: tamanho das sequências consideradas para o cálculo do valor futuro na ST; **(b) Conjunto de exemplos de treinamento**: conjunto de sequências pertencentes à ST que constituem o conjunto de treinamento; **(c) Medida de similaridade**: utilizada para quantificar a similaridade entre os exemplos; **(d) Cardinalidade do conjunto de sequências similares**: quantidade (k) de sequências mais próximas consideradas para a previsão do valor futuro e **(e) Função de previsão**: utilizada para determinar a maneira como serão considerados os valores das sequências mais próximas para estimar o valor futuro.

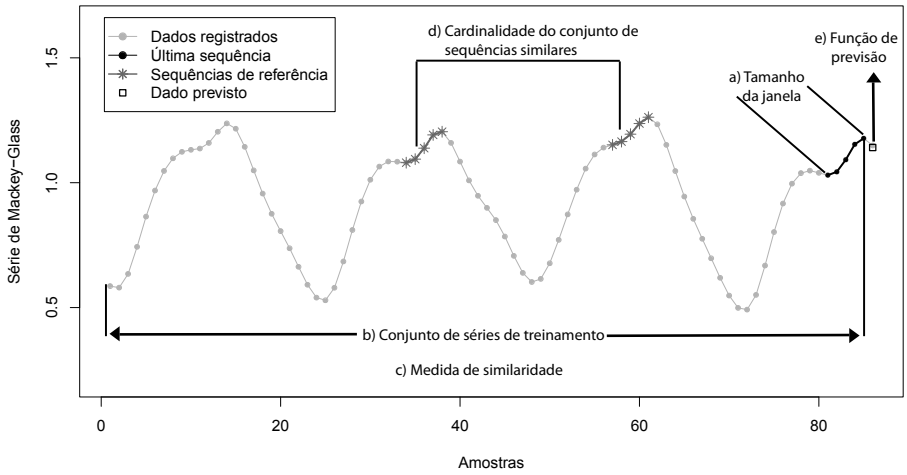


Figura 1. Parâmetros do algoritmo *kNN-TSP* (Modificado de [Ferrero et al. 2009]).

3. Medidas de Similaridade entre Séries Temporais

A medida de similaridade define o critério para quantificar quão similares são duas sequências e decidir se serão denominadas como pertencentes ou não a determinado padrão. A aplicação dessa definição em ST é bastante subjetiva, pois é dependente de diversos fatores. Entre esses fatores estão o domínio da aplicação e as características do método escolhido para o cálculo da similaridade [Aggarwal et al. 2001]. A seguir são descritas algumas das principais medidas de similaridade aplicadas em ST.

3.1. Norma L_p

As medidas da Norma L_p, em especial a distância Euclidiana, estão entre as medidas de distância mais conhecidas e exploradas na literatura, geralmente tendo sua aplicação expandida para dados bidimensionais, tridimensionais, ou de maior número de dimensões. Para o cálculo da distância baseado na Norma L_p, cada sequência é considerada um ponto no espaço W-dimensional. Desse modo, a similaridade entre essas sequências é dada pela diferença entre esses pontos (Equação 2) [Aggarwal et al. 2001]:

$$L_p(x, y) = \left(\sum_{i=1}^{W} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{2}$$

onde x e y são os vetores W-dimensionais, correspondentes as sequências em uma ST; e p define a medida de distância a ser utilizada. Neste trabalho, as medidas da Norma L_p foram subdividas em dois grupos: L_p Inteiras, que representam valor de p igual ou superiores a um (1) sem casas decimais, e L_p Fracionárias, representando valores de p superiores a zero (0) e inferiores a um (1), indicando assim valores fracionários.

Dentre as mais conhecidas estão a distância Euclidiana e a distância Manhattan. As medidas L_p Inteiras são nomeadas de acordo com o valor de p: p=1: Manhattan, também conhecida como *City Block* (L_1); p=2: Euclidiana (L_2) e p=3: Métrica L_3 (L_3).

Em Aggarwal et al (2001) foi demonstrado que valores de p da Norma L_p menores que um, em conjuntos de grande dimensionalidade, apresentam vantagens se comparados às medidas com valores inteiros. Essa vantagem se dá pelo fato das medidas fracionárias atribuírem mais peso a pequenas variações entre os dados.

Enquanto as medidas L_p Inteiras tendem a formar espaços de busca quadriculares de cantos arredondados, a medida que se aumenta o valor de p de um a L_∞ (infinito), até uma figura quadrática, as medidas fracionárias tendem a retrair os espaço de busca. Assim, o losango formado por p=1 tende a ter suas laterais atraídas ao centro da figura.

3.2. Canberra

Essa distância assemelha-se à distância Manhattan, calculando a diferença absoluta entre dois vetores. No entanto, na distância Canberra a diferença absoluta é dividida pela soma dos valores absolutos, antes de realizar a soma. Assim, dados dois vetores, x e y, W-dimensionais, a distância Canberra é dada pela Equação 3 [Deza e Deza 2006]:

$$c_a(x, y) = \sum_{i=1}^{W} \frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{3}$$

3.3. Geodésica

Um espaço de métrica pode ser considerado geodésico se quaisquer dois pontos do espaço estão unidos por um segmento geodésico, que é a menor curva entre dois pontos. Desse modo, essa distância consiste no comprimento de um segmento geodésico entre dois pontos em um espaço geodésico [Deza e Deza 2006]. Assim, dados dois vetores, x e y, W-dimensionais, a distância Geodésica é dada pela Equação 4 [Meyer e Bucht 2011]:

$$g_e(x, y) = \arccos\left(\frac{xy}{\sqrt{xx \times yy}} \right) \tag{4}$$

onde xy indica a multiplicação escalar dos vetores x e y; e xx e yy indicam a multiplicação escalar do vetor x com ele próprio e do vetor y com ele próprio, respectivamente. A distância entre os vetores é dada pelo ângulo desses vetores no espaço W-dimensional.

3.4. *Dynamic Time Warping*

Essa distância busca alinhar, da maneira mais adequada possível, os valores das séries a serem comparadas. Isso permite que duas ST globalmente similares, mas que estejam fora de alinhamento no eixo temporal, possam ser alinhadas para posterior comparação ponto-a-ponto [Ratanamahatana e Keogh 2005]. Para alinhar duas séries Z e T, de tamanhos m e n, respectivamente, o algoritmo constrói uma matriz $m \times n$, na qual cada elemento (i, j) corresponde ao valor da distância entre os pontos (Z_i, T_j), então busca-se uma rota R, que alinhe as séries Z e T, conforme a Equação 5.

$$R = (r_1, r_2, \ldots, r_L) \tag{5}$$

onde cada r_l corresponde a um mapeamento $(i, j)_L$ para $l = 1, \ldots, L$ e L, representando o tamanho da rota, está restrito à condição $max(m, n) \leq L < m + n$ [Salvador e Chan 2007].

4. Configuração Experimental

De maneira a avaliar o comportamento do algoritmo *kNN-TSP* e a influência das medidas de similaridade no algoritmo frente a uma grande quantidade de situações, foram selecionadas séries artificiais e reais para a realização dos experimentos.

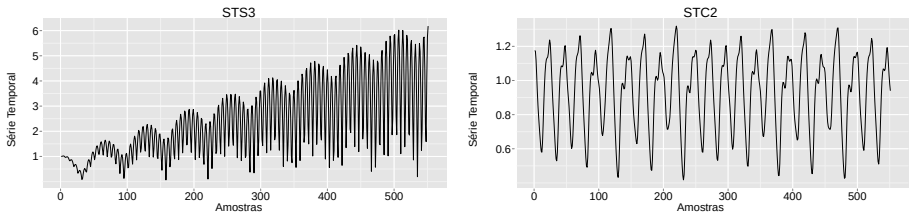
4.1. Séries Artificiais

Foram consideradas cinco séries temporais artificiais para avaliar o comportamento do algoritmo utilizando as diferentes medidas de similaridade [Kulesh et al. 2008]. Essas séries estão agrupadas em duas famílias de acordo com as suas características: ST de modelos sazonais (STS) e ST de modelos caóticos (STC) (Tabela 1).

Tabela 1. Características das ST artificiais.

Séries Temporais de Modelos Sazonais		
Id	**Série Temporal**	***Tamanho (m)***
STS1	Dependência sazonal	2200
STS2	Sazonalidade multiplicativa	590
STS3	Alta frequência	550
Séries Temporais de Modelos Caóticos		
STC1	Lorenz	551
STC2	Mackey-Glass	551

As Séries Temporais de Modelos Sazonais (STS) permitem avaliar o algoritmo de previsão em comportamentos previsíveis. Em geral, apresentam tendência definida e mudança de amplitude de modo sazonal, como exemplificada na Figura 2(a).



(a) (b)

Figura 2. Séries temporais artificiais geradas através de modelos sazonais e caóticos: (a) STS3, (b) STC2 (Modificado de [McNames 1999] e [Kulesh et al. 2008]).

As STS utilizadas neste trabalho foram: **Série temporal de dependência sazonal (STS1)**: sazonalidade constante e tendência linear; **Série temporal de sazonalidade multiplicativa (STS2)**: variação de tendência não-linear e sazonalidade multiplicativa; **Série temporal de alta frequência (STS3)**: possui sazonalidade multiplicativa e constante aumento de amplitude.

As ST de Modelos Caóticos (STC) permitem avaliar algoritmos de previsão, frente a comportamentos pouco previsíveis e que apresentam ciclos não-repetitivos, aumentando a dificuldade de previsão dessas séries em relação às de modelos sazonais [McNames 1999, Kulesh et al. 2008]. Uma dessas séries é exemplificada na Figura 2(b).

As STC utilizadas neste trabalho foram: **Sistema de Lorenz (STC1)**: comportamento não-periódico e imprevisível gerado por meio de um sistema de equações diferenciais; **Sistema de Mackey-Glass (STC2)**: características de sistemas caóticos, geradas por meio de um sistema de equações originalmente desenvolvido para modelar a formação de linfócitos.

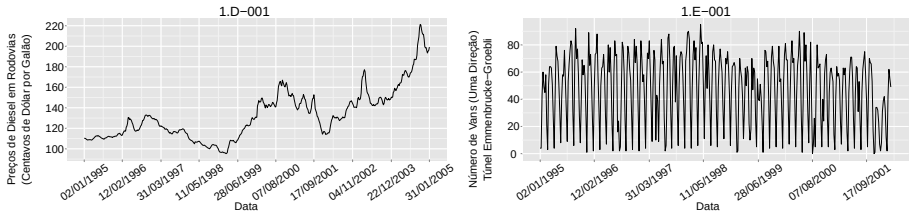
4.2. Séries Reais

Netes trabalho, foram utilizadas diversas séries disponibilizadas pela edição do ano de 2010 da *Time Series Forecasting Grand Competition for Computational Intelligence (NN GC1)*[2]. Na Tabela 2 são apresentadas as ST da *NN GC1* agrupadas por base de dados, a frequência de aquisição dos dados e o tamanho das séries.

Tabela 2. Características das ST disponíveis pela *NN GC1*.

Séries	Base de Dados	Quantidade de Séries	Aquisição	Tamanho (m)
1.B-001 a 1.B-011	1.B	11	Quaternal	31 a 148
1.C-001 a 1.C-011	1.C	11	Mensal	48 a 228
1.D-001 a 1.D-011	1.D	10	Semanal	437 a 618
1.E-001 a 1.E-011	1.E	10	Diária	377 a 747
1.F-003 a 1.F-011	1.F	09	Horária	902 a 1742

Todas as séries são mensurações de dados relacionados ao transporte, as quais incluem tráfego em rodovias, tráfego de pessoas em metrôs, entre outros, como exemplificado na Figura 3.



(a) (b)

Figura 3. Séries reais disponibilizadas pela *NN GC1*: (a) 1.D-001 e (b) 1.E-001.

4.3. Configuração de Parâmetros

Como mencionado, o algoritmo *kNN-TSP* necessita da configuração de alguns parâmetros, dentre eles a medida de similaridade, o tamanho da janela de busca (w), a quantidade de pontos a serem previstos e a quantidade de vizinhos mais próximos (k). Neste

[2]http://www.neural-forecasting-competition.com/

trabalho, foram considerados os valores de $k = 1, 5$ e 10 para todas as séries do experimento, conforme selecionados em estudos preliminares [Aikes Junior et al. 2011] e como medidas de similaridade: Norma L_p (sendo as inteiras $p = 1, 2$ e 3, e as fracionárias $p = 0.1, 0.3, 0.5$ e 0.7), Canberra, Geodésica e *DTW*. Neste trabalho, o algoritmo *kNN-TSP* foi utilizado para realizar previsões de um valor futuro, considerando, a cada estimativa, todos os valores observados do passado.

O valor de w para as séries artificiais foi selecionado baseado em Kulesh et al. (2008), e para as séries reais foi selecionado de acordo com a sazonalidade sugerida por Lemke e Gabrys (2010). A quantidade de valores previstos foi baseado na *NN GC1* (Tabela 3).

Tabela 3. Configuração dos parâmetros *w*, *m* e número de valores previstos para as ST artificias e da *NN GC1* [Kulesh et al. 2008, Lemke e Gabrys 2010].

Séries Artificiais				
Id	**Série Temporal**	***m***	***w***	**Pontos Previstos**
STS1	Dependência sazonal	2200	100	220
STS2	Sazonalidade multiplicativa	590	15	88
STS3	Alta frequência	550	70	55
STC1	Lorenz	551	25	100
STC2	Mackey-Glass	551	7	100
Séries Reais				
Id	**Série Temporal**	***m***	***w***	**Pontos Previstos**
1.B	Quaternais	31 a 148	4	8
1.C	Mensais	48 a 228	12	12
1.D	Semanais	437 a 618	52	26
1.E	Diárias	377 a 747	7	14
1.F	Horárias	902 a 1742	24	48

A função de previsão utilizada em conjunto com o *kNN-TSP* neste trabalho foi a Média de Valores Relativos (MVR) [Ferrero 2009]. Essa função foi escolhida por apresentar melhores resultados na previsão de ST, se comparadas com abordagens tradicionais da literatura [Ferrero et al. 2009].

Os erros de predição foram medidos por meio do *Mean Absolute Percentage Error* (*MAPE*) [Hyndman e Koehler 2006]. Para a análise quanto à existência de diferença estatisticamente significativa (**d.e.s**) foi utilizado o teste estatístico não-paramétrico de Friedman para dados emparelhados e comparações múltiplas, considerando nível de significância de 5% ($p\text{-}valor < 0,05$), com pós-teste de Dunn [Motulsky 1995].

5. Resultados e Discussão

De modo a avaliar a influência das diversas medidas de similaridade, bem como o número de vizinhos próximos, na previsão de ST utilizando o algoritmo *kNN-TSP*, uma avaliação empírica foi realizada usando séries artificiais [McNames 1999, Kulesh et al. 2008] e séries reais[3]. Devido ao fato de que as séries artificiais são geradas utilizando uma função conhecida, ao contrário das séries reais, e de modo a controlar essa variável, os resultados foram analisados separadamente.

Por meio da aplicação do teste de Friedman obtiveram-se $p\text{-}valores < 0,0001$, tanto para as séries artificiais quanto reais, constatando a existência de **d.e.s** entre os grupos. Após, aplicou-se o pós-teste de Dunn, cujos resultados são apresentados na Tabela 4:

[3]http://www.neural-forecasting-competition.com/

Cada célula preenchida representa a existência de **d.e.s** entre a medida da linha e da coluna de encontro dessa célula. Por exemplo, a notação *** na célula de encontro entre a linha da medida L_1 e a coluna da medida $L_{0.1}$, representa a existência de **d.e.s** com $p\text{-}valor < 0,001$ entre essas medidas, sendo que a distância L_1 apresentou um valor médio de *MAPE* menor que a $L_{0.1}$ (**d.e.s** favorável).

Tabela 4. Comparativo sobre a existência de d.e.s entre as medidas de similaridade para as séries artificiais e reais (*, ** e * para d.e.s com p-valor < 0,001, 0,01 e 0,05, respectivamente).**

Artificiais									
	L_1	L_2	L_3	$L_{0.1}$	$L_{0.3}$	$L_{0.5}$	$L_{0.7}$	***DTW***	**Canberra**	**Geodésica**
L_1	———			***				***	***	**
L_2		———		***	***	**		***	***	***
L_3			———	***	***	***	**	***	***	***
$L_{0.1}$				———		*	***	***	***	
$L_{0.3}$					———			***	***	
$L_{0.5}$						———		***	***	
$L_{0.7}$							———	***	***	
DTW								———		
Canberra								***	———	
Geodésica								***	***	———
Reais										
	L_1	L_2	L_3	$L_{0.1}$	$L_{0.3}$	$L_{0.5}$	$L_{0.7}$	***DTW***	**Canberra**	**Geodésica**
L_1	———							***		
L_2		———		***	**			***	***	
L_3			———	***	***			***	***	
$L_{0.1}$				———						
$L_{0.3}$					———					
$L_{0.5}$						———				
$L_{0.7}$							———			
DTW				***	***	***	***	———		
Canberra	***					**	**	**	———	
Geodésica								***	**	———

Observa-se nessa tabela, que as medidas L_p Inteiras apresentaram **d.e.s** para com a maioria das medidas, sendo que suas médias foram, em geral, inferiores às das demais, em especial, nas séries artificiais. Pode ser observado ainda que não foi possível constatar existência de **d.e.s** entre as medidas L_p Inteiras, indicando que não há evidência estatística que comprove melhor qualidade em termos de acurácia de alguma delas. Assim, devido ao menor custo computacional, a distância Manhattan (L_1) pode ser determinada como a melhor escolha entre essas três. Esse resultado corrobora com o alcançado em Aikes Junior et al. (2011) para as mesmas séries artificiais e reais, considerando valores de k variando entre um e cinco.

Apesar das medidas L_p Fracionárias apresentarem vantagens quanto à acurácia nas tarefas de agrupamento e classificação, quando comparadas às medidas L_p Inteiras em dados com várias dimensões [Aggarwal et al. 2001], essa vantagem não foi constatada para a previsão de ST e as medidas L_p Fracionárias apresentaram maiores valores de *MAPE*, sendo que foi constatada **d.e.s** favorável apenas quando comparadas às distâncias *DTW* e Canberra nas séries artificiais. Entretanto, não houve **d.e.s** favorável para essas medidas nas séries da *NN GC1*, inclusive havendo **d.e.s** desfavorável para algumas medidas L_p Fracionárias quando comparadas a *DTW* e a Canberra. Assim, a característica das medidas L_p Fracionárias de ressaltar a percepção de pequenas diferenças em conjuntos de alta dimensionalidade não foi vantajosa para a seleção dos melhores vizinhos mais próximos nas séries avaliadas neste trabalho. Desse modo, verifica-se que as distâncias L_p Fracionárias podem não ser as mais adequadas para a previsão de dados temporais utilizando o algoritmo k*NN-TSP*, considerando um cenário mais amplo.

Em geral, a distância *DTW* apresentou **d.e.s** com todas as medidas, sendo que, na maior parte dos casos, seus valores de erro médio foram superiores aos das demais medidas. Assim, ao contrário das tarefas de classificação na qual a *DTW* apresentou desempenho superior às L_p Inteiras [Ratanamahatana e Keogh 2005], essa medida aparenta não ser adequada como medida de similaridade empregada na tarefa previsão de dados temporais utilizando o algoritmo kNN-*TSP*. A razão dessa degradação de desempenho pode estar relacionada com a característica dessa medida de buscar o melhor alinhamento entre as séries, o que acaba por diminuir a influência de pequenas diferenças locais durante a etapa de busca dos vizinhos mais próximos. Assim, diferenças locais que poderiam ser verdadeiras e desejadas para o cálculo do valor futuro acabam recebendo menor influência ou mesmo não sendo consideradas. Isso possibilita que a escolha dos vizinhos mais próximos acabe por não ser a mais adequada para a função de previsão.

A medida Canberra permaneceu como medida de desempenho intermediário tanto para as séries artificiais quanto para as séries da *NN GC1*. Assim, a característica de normalização dessa medida demonstra não apresentar vantagens consideráveis quando empregada na previsão de ST com o algoritmo kNN-*TSP*.

Apesar da medida Geodésica ter apresentado valores de *MAPE* baixos, esses não foram os melhores. Essa medida apresentou **d.e.s** desfavorável apenas para as medidas L_p Inteiras e favorável apenas para a *DTW* e Canberra. Entretanto, seu elevado custo computacional, quando comparada às medidas L_p Inteiras e a não constatação de **d.e.s** favorável para a Geodésica nesses casos, indica que essa distância não apresentou vantagem considerável. Dessa forma, reforça-se a suposição de que medidas de menor custo computacional, como a Manhattan, podem ser empregadas para a previsão de ST.

6. Conclusões

Neste trabalho foi apresentado um estudo sobre a influência do parâmetro medida de similaridade na acurácia da previsão de séries temporais utilizando o algoritmo kNN-*TSP*. De maneira a estudar essa influência, foram selecionadas várias medidas de similaridade, conforme descrito na Seção 3. O algoritmo kNN-*TSP*, em conjunto com essas medidas, foi submetido a uma avaliação experimental contendo séries temporais artificiais, de características sazonais e caóticas, e várias séries temporais reais relacionadas a transporte, com características diversas, conforme descrito na Seção 4. Os resultados foram avaliados por meio do *Mean Absolute Percentage Error* (*MAPE*), e foram realizadas análises de estatística analítica, permitindo dessa maneira verificar a existência de diferença estatisticamente significativa (**d.e.s**) que possa apoiar a escolha de medidas de similaridade que apresentem algum benefício.

As medidas L_p Inteiras apresentaram os menores valores de *MAPE* para a maioria das situações, ou, nos piores casos, valores de erro médio intermediários. Em geral, para essas medidas, foram constatadas **d.e.s** favoráveis para com a maioria das medidas avaliadas, sem terem sido evidenciadas **d.e.s** desfavoráveis. Esse resultado demonstra a capacidade de adaptação do algoritmo kNN-*TSP* utilizando essas medidas para as séries temporais avaliadas, favorecendo a sua utilização.

As medidas L_p Fracionárias, por sua vez, alcançaram valores de erro médio elevados para a maioria dos casos, tendo sido encontradas como as medidas de maior erro médio, ou, quando apresentaram erro médio intermediário, em geral, esse estava próximo

do maior. Seu elevado custo computacional, quando comparadas às medidas L_p Inteiras, somado ao fato de, em grande parte das situações ter sido possível comprovar **d.e.s** desfavorável para com as demais medidas, acabam por tornar a utilização das medidas L_p Fracionárias desfavorável para a maioria dos casos estudados neste trabalho.

O desempenho da medida *Dynamic Time Warping* (*DTW*) para as tarefas de classificação, incluindo classificação de dados temporais é, em geral, melhor que os das medidas da Norma L_p. Entretanto, para a previsão de dados temporais utilizando o algoritmo k*NN-TSP*, essa medida apresentou valores de erro médio elevados, em geral acrescidos de **d.e.s** desfavorável. Desse modo, o alto custo computacional dessa medida, somado ao fato de ter alcançado, em alguns casos, os maiores valores de erros, desfavorecem a sua utilização para a previsão de séries temporais com o algoritmo k*NN-TSP*.

A medida Canberra apresentou, para a maioria das situações verificadas neste trabalho, valores de erro médio intermediários. Apesar de seu custo computacional ser inferior aos das medidas Euclidiana e Métrica L_3, devido a sua característica de normalização, é ainda superior ao da medida Manhattan, não tendo evidenciado **d.e.s** favorável para com essas medidas para a maioria dos casos. Esses resultados desfavorecem a utilização dessa medida, em comparação com as medidas L_p Inteiras, para a maioria das situações. Entretanto, seus valores de erro médio geralmente baixos e seu custo computacional superior apenas ao da medida Manhattan, favorecem a utilização dessa medida em comparação com as medidas fracionárias e *DTW*.

Em relação à medida Geodésica, foram encontrados valores de erro médio intermediários, porém, próximos aos menores valores. Entretanto, seu elevado custo computacional, somado ao fato de não ter sido constatada, para a maioria dos casos, **d.e.s** favorável quando comparada à medidas de custo computacional inferior, como as L_p Inteiras, acabam por desfavorecer a utilização dessa medida neste trabalho.

Por meio da avaliação experimental, verificou-se que as medidas pertencentes à Norma L_p, utilizando valores de p inteiros entre um e três, apresentaram os menores valores de erros para a maioria dos casos, tanto para as séries artificiais quanto reais. A não constatação de **d.e.s** entre essas medidas indica que não há possível vantagem de alguma dessas medidas sobre as demais, dessa forma, medidas que apresentam um menor custo computacional podem ser preferenciais. Adicionalmente, verifica-se que a medida Manhattan apresenta-se candidata interessante como medida de similaridade a ser adotada para a maioria dos casos em que se desejam realizar previsões utilizando o algoritmo k*NN-TSP*. A medida Canberra também pode ser considerada interessante, por apresentar um custo computacional baixo, acima somente ao da medida Manhattan, e ter alcançado resultados de previsão próximos aos dessa medida.

Trabalhos futuros incluem a utilização de ST de outros domínios e a investigação da influência do parâmetro k na acurácia da previsão.

Referências

Aggarwal, C., Hinneburg, A., e Keim, D. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. *Proceedings of the Eighth International Conference on Database Theory*, pages 420–434.

Aguirre, L. A. (2007). *Introdução à Identificação de Sistemas: Técnicas Lineares e Não-Lineares Aplicadas a Sistemas Reais*. UFMG, Belo Horizonte.

Aikes Junior, J., Lee, H. D., Ferrero, C. A., Zalewski, W., e Wu, F. C. (2011). Estudo da Influência de Medidas de Similaridade da Norma Lp no Algoritmo kNN-TSP para Previsão de Dados Temporais. In *X Conferência Brasileira de Dinâmica, Controle e Aplicações*, Águas de Lindóia.

Chiu, B., Keogh, E., e Lonardi, S. (2003). Probabilistic Discovery of Time Series Motifs. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 493–498.

Deza, M. e Deza, E. (2006). *Dictionary of Distances*. Elsevier, Amsterdam.

Ferrero, C. A. (2009). Algoritmo kNN para Previsão de Dados Temporais: Funções de Previsão e Critérios de Seleção de Vizinhos Próximos Aplicados a Variáveis Ambientais em Limnologia. Master's dissertation, Instituto de Ciências Matemáticas e Computação (ICMC) - Universidade de São Paulo (USP/São Carlos).

Ferrero, C. A., Monard, M. C., Lee, H. D., e Wu, F. C. (2009). Proposta de uma Função de Previsão de Dados Temporais para o Algoritmo dos Vizinhos mais Próximos. In *Anais do XXXV Confêrencia Latinoamericana de Informática*, pages 1–10, Pelotas.

Han, J. e Kamber, M. (2006). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 2 edition.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall, New Jersey, 2 edition.

Hyndman, R. e Koehler, A. (2006). Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting*, 22(4):679–688.

Karunasinghe, D. S. e Liong, S.-Y. (2006). Chaotic Time Series Prediction With a Global Model: Artificial Neural Network. *Journal of Hydrology*, 323(1-4):92–105.

Kulesh, M., Holschneider, M., e Kurennaya, K. (2008). Adaptive Metrics in the Nearest Neighbours Method. *Physica D: Nonlinear Phenomena*, 237(3):283–291.

Lemke, C. e Gabrys, B. (2010). Meta-learning for Time Series Forecasting in the NN GC1 Competition. In *World Congress on Computational Intelligence*, pages 1–5, Barcelona. IEEE.

McNames, J. (1999). *Innovations in Local Modeling for Time Series Prediction*. PhD thesis, Stanford.

Meyer, D. e Bucht, C. (2011). *Package Proxy: Distance and Similarity Measures*.

Morettin, P. A. e Toloi, C. M. C. (2006). *Análise de Séries Temporais*. Edgard Blücher LTDA, São Paulo, 2 edition.

Motulsky, H. (1995). GraphPad InStat 3.0 User's Guide. http://www.graphpad.com.

Ratanamahatana, C. e Keogh, E. (2005). Three Myths About Dynamic Time Warping Data Mining. In *Proceedings of SIAM International Conference on Data Mining*, pages 506–510. Citeseer.

Salvador, S. e Chan, P. (2007). Toward Accurate Dynamic Time Warping in Linear Time and Space. *Intelligent Data Analysis*, 11(5):561–580.