

**CENTRO FEDERAL DE EDUCACAO TECNOLÓGICA
CELSO SUCKOW DA FONSECA**

**Avaliação de Agregação Temporal na Previsão da
Temperatura de Superfície do Mar do Oceano
Atlântico**

Rebecca Pontes Salles

Prof. Orientador:
Eduardo Soares Ogasawara, D.Sc.

**Rio de Janeiro,
Julho de 2016**

**CENTRO FEDERAL DE EDUCACÃO TECNOLÓGICA
CELSO SUCKOW DA FONSECA**

**Avaliação de Agregação Temporal na Previsão da
Temperatura de Superfície do Mar do Oceano
Atlântico**

Rebecca Pontes Salles

Projeto final apresentado em cumprimento às
normas do Departamento de Educação
Superior do Centro Federal de Educação
Tecnológica Celso Suckow da Fonseca,
CEFET/RJ, como parte dos requisitos para
obtenção do título de Bacharel em Ciência da
Computação.

Prof. Orientador:
Eduardo Soares Ogasawara, D.Sc.

**Rio de Janeiro,
Julho de 2016**

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

S168 Salles, Rebecca Pontes
Avaliação de agregação temporal na previsão da temperatura
de superfície do mar do Oceano Atlântico / Rebecca Pontes
Salles.—2016.
xiii, 58f. : il. (algumas color.) , grafs. , tabs. ; enc.

Projeto Final (Graduação) Centro Federal de Educação
Tecnológica Celso Suckow da Fonseca , 2016.
Bibliografia : f. 51-58
Orientador : Eduardo Soares Ogasawara

1. Computação. 2. Precipitação (Meteorologia) – Previsão. 3.
Análise de séries temporais. 4. Temperatura atmosférica. 5.
Atlântico, Oceano, I. Ogasawara, Eduardo Soares (Orient.). II.
Título.

CDD 004

DEDICATÓRIA

A Deus e à minha família amada que me
ajudaram, apoiaram e guiaram ao longo de
toda a minha vida.

AGRADECIMENTOS

Agradece-se ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo financiamento parcial desta pesquisa.

Agradece-se também as contribuições de Patricia Mattos que deu início a pesquisas sobre o tema abordado.

RESUMO

O crescente aumento de congestionamentos no tráfego rodoviário demanda pesquisas relacionadas a mobilidade urbana. Essas pesquisas modelam o problema do tráfego como de trajetória, a partir da análise individualizada de objetos móveis que continuamente transmitem sua geolocalização. Um exemplo disso são os ônibus do Rio de Janeiro. Tais veículos funcionam como sensores de trajetória e produzem grande quantidade de dados. Algumas questões, no entanto, continuam em aberto, como características do transporte público em horários de pico e comportamentos sistêmicos presentes nas diferentes regiões que interfiram na mobilidade da cidade. A agregação espaço-temporal dos dados das trajetórias dos modais urbanos oferece uma visão sumarizada dos dados com potencial para identificação de padrões sistêmicos. Este trabalho visa a aplicar técnicas de identificação de motifs sobre esses dados agregados. Espera-se encontrar padrões que expliquem as diferentes formações e propagações de atrasos, bem como outros fenômenos escondidos sob esse grande volume de dados.

Palavras-chave: modelos de previsão; séries temporais; agregação temporal; temperatura da superfície do mar; Oceano Atlântico

ABSTRACT

Extreme environmental events such as droughts affect millions of people all around the world. Although it is not possible to prevent this type of event, its prediction under different time horizons enables the mitigation of eventual damages caused by its occurrence. An important variable for identifying occurrences of droughts is the Sea Surface Temperature (SST). In the Tropical Atlantic Ocean, SST data are collected and provided by the Prediction and Research Moored Array in the Tropical Atlantic (*Prediction and Research Moored Array in the Tropical Atlantic* (PIRATA)) Project, which is an observation network composed of sensor buoys arranged in this region. Sensors of this type, and more generally Internet of Things (Internet das Coisas (IoT)) sensors, commonly lead to data losses that influence the quality of data sets collected for adjusting prediction models. In this paper, we explore the influence of temporal aggregation in predicting step-ahead SST considering different prediction horizons and different sizes for training data sets. We have conducted several experiments using data collected by PIRATA Project. Our results point out scenarios for training data sets and prediction horizons indicating whether or not temporal aggregated SST time series may be beneficial for prediction.

Keywords: prediction models; time series; temporal aggregation; sea surface temperature; Atlantic Ocean

SUMÁRIO

1	Introdução	1
2	Preliminares	3
3	Trabalhos Relacionados	4
4	Proposta	6
5	Estado Atual do Trabalho	8
	Referências Bibliográficas	9

LISTA DE FIGURAS

FIGURA 1:	Conversão dos dados de localização dos pontos de ônibus.	8
FIGURA 2:	Estações obtidas aplicando o algoritmo de agrupamento DBSCAN sobre os pontos de ônibus do Rio de Janeiro.	8

LISTA DE TABELAS

TABELA 1:	Comparação dos trabalhos relacionados	5
TABELA 2:	Cronograma de desenvolvimento da dissertação	9

LISTA DE ABREVIACÕES

CNPQ	Conselho Nacional De Desenvolvimento Científico E Tecnológico	v
IOT	Internet Das Coisas	vii
PIRATA	<i>Prediction and Research Moored Array in the Tropical Atlantic</i>	vii
SST	Sea Surface Temperature	vii

Capítulo 1

Introdução

O aumento da população mundial em áreas urbanas associado ao grande número de veículos presentes nas cidades provocam problemas como congestionamentos, acidentes e poluição [Ferreira et al., 2013; Chen et al., 2015]. No Brasil, o aumento de congestionamentos no tráfego rodoviário de regiões metropolitanas é assunto de preocupação e de constante estudo. Em particular, o Rio de Janeiro desponta como a quarta cidade com pior tráfego no mundo [tom, 2016]. Neste contexto, observa-se a necessidade de intensificação de pesquisas relacionadas à análise de dados do transporte e mobilidade urbana visando a identificação de fatores causadores de congestionamento (gargalos).

Para entender a dinâmica do trânsito urbano e identificar problemas comuns à mobilidade, dados referentes a objetos móveis são coletados principalmente por meio de GPS. No Rio de Janeiro, cada ônibus da frota municipal possui um GPS associado que emite dados referentes à sua geolocalização. Desta forma, cada ônibus corresponde a um objeto móvel e se caracteriza como um sensor de trajetória que coleta e transmite sua localização ao longo do tempo.

Para compreender estes dados, faz-se valer de técnicas de análise de séries espaço-temporais. As séries espaço-temporais podem ser compostas por observações de objetos móveis ou de objetos permanentes [Frank, 2003]. Quando uma das propriedades variantes do objeto é a sua posição, a sequência de observações do objeto móvel configura o problema de trajetória. Neste contexto, estudam-se, por exemplo, rotas frequentemente observadas pelos objetos e caminhos com propriedades semelhantes [Spaccapietra et al., 2008].

No contexto do monitoramento de ônibus, parte dessa análise individualizada já é conhecida, uma vez que os mesmos têm linhas associadas e circulam nas rotas dessas linhas. Algumas perguntas importantes ficam em aberto, como, por exemplo, características do transporte público em horários de pico, ou, também, comportamentos sistêmicos presentes nas diferentes regiões, como padrão de formação e propagação de engarrafamentos e identificação de presença de pontos de interesse, tais como shoppings, escolas e hospitais que possam interferir na mobilidade de uma região da cidade.

Visando cobrir tal lacuna, esta dissertação visa aplicar técnicas de mineração de dados de

identificação de *motifs* (padrões não conhecidos) sobre dados agregados de mobilidade urbana. Na agregação espaço-temporal, as informações são agregadas considerando-se um particionamento do tempo (unidade temporal) e do espaço (unidade espacial). Na prática, cada unidade espacial (região) passa a ter uma série espaço-temporal de objeto permanente associada. Esta série é produzida a partir da agregação das informações referentes aos diferentes objetos móveis que percorrem a região na unidade temporal analisada [Tao et al., 2004]. Intuitivamente, estas séries espaço-temporais agregadas funcionam como sensores virtuais distribuídos. A agregação espaço-temporal favorece uma visão sumarizada dos dados por região. Desta maneira, a partir da análise dos dados de cada unidade espacial e das unidades vizinhas, é possível entender como engarrafamentos se propagam pela cidade.

Também é possível identificar padrões diferenciados que se repetem em regiões específicas que possam indicar a presença de pontos de interesse. Sendo assim, o desafio consiste em identificar e analisar *motifs* nestas séries agregadas que promovam as respostas aos questionamentos supracitados. Desta forma, o trabalho pretende contribuir ao estabelecer novos algoritmos e técnicas necessárias para identificação e análise desses *motifs* em agregações de séries espaço-temporais de mobilidade urbana.

Além dessa introdução, o trabalho se divide em mais quatro outras seções. A seção 2 apresenta a revisão da literatura. A seção 3 apresenta os trabalhos relacionados. A proposta propriamente dita é apresentada na seção 4. Finalmente, a seção 5 apresenta o estado atual do trabalho o planejamento da dissertação.

Capítulo 2

Preliminares

Séries espaço-temporais são definidas como sequência de observações que contêm dados sobre o local e momento em que coleta foi feita [Cressie and Wikle, 2015]. As séries espaço-temporais podem ser caracterizadas como observações de objetos móveis ou objetos permanentes. Os objetos que tem dados de sua localização variantes com o tempo são classificados como móveis. A sequência de observações sobre objetos móveis é classificada como trajetória. A trajetória é o modelo de dados mais aplicado a problemas relacionados ao tráfego [Chen et al., 2015]. Entretanto, estudos que se relacionam mais diretamente com o tema deste trabalho são aqueles em que se agrega informações do objeto [Tao et al., 2004], gerando séries espaço-temporais associados a objetos permanentes.

A agregação de dados de objetos móveis pode ser feita de diferentes formas. Os tipos básicos de agregações existentes são Espacial (S), Temporal (T) e Atributiva ou categórica (A). Outros tipos de agregação podem ser feitos a partir da combinação dos tipos básicos, como $S \times T$, $S \times T \times A$ e $S \times S \times T \times T$ [Chen et al., 2015]. Neste trabalho, a agregação aplicada é a Espaço-Temporal ($S \times T$), em que o resultado esperado é composto por séries temporais (T) associadas a pontos geográficos permanentes (S). Tais agregações podem ser interpretadas como sensores virtuais correspondente a região agregada.

As séries temporais apresentam subsequências que se repetem com frequência. Alguns dos padrões que se repetem não são conhecidos, configurando *motifs* [Esling and Agon, 2012]. A identificação de *motifs* podem ser feita sobre séries temporais univariadas ou multivariadas. Neste trabalho, as séries espaço-temporais obtidas após a agregação $S \times T$ são multivariadas [Wang et al., 2016] [Vahdatpour et al., 2009].

Capítulo 3

Trabalhos Relacionados

Diversos trabalhos foram desenvolvidos recorrendo à agregações espaço-temporais para lidar com problemas de mobilidade urbana. Normalmente, as agregações feitas nessas pesquisas são baseadas em regiões (AETR), definidas através da projeção de grades sobre a região observada [Ferreira et al., 2013; Andrienko and Andrienko, 2008; Andrienko and Andrienko, 2011; Ferreira et al., 2013]. Nessa classe de trabalhos, pode-se destacar Andrienko and Andrienko [2008], que adotam a agregação espaço-temporal baseada em regiões, enquanto Andrienko and Andrienko [2011] adotam dois tipos diferentes de agregação - baseada em tempo e em pontos geográficos permanentes (AETP). A agregação baseada em pontos geográficos permanentes adotada por Andrienko and Andrienko [2011] considerou os pontos turísticos da região observada como sensores virtuais.

A identificação de *motifs* em séries espaço-temporais não tem sido explorada na literatura. Em contrapartida, há diversos trabalhos que envolvem identificação de *motifs* sobre séries temporais (IMS) [Li and Nallela, 2009]. Cassisi et al. [2013], por exemplo, aplicou a técnica a problemas relacionados à sísmica, Jiang et al. [2008] ao mercado financeiro e Chi et al. [2012] ao reconhecimento facial. Apesar dos trabalhos listados aplicarem a técnica a problemas reais, a maioria dos trabalhos estudam séries temporais univariadas.

No contexto de mobilidade urbana, a identificação de *motifs* é feita sobre dados de trajetória (IMT). Schneider et al. [2013] faz uma análise a partir de dados de trajetória individuais. Apesar deste trabalho também visar a identificação de *motifs* em problema de mobilidade urbana, o modelo de dados aplicado é o de séries espaço-temporais associadas a pontos geográficos permanentes. Desta forma, as técnicas de identificação de *motifs* em trajetórias não se adequam ao problema proposto nesse trabalho.

A tabela 1 apresenta os trabalhos relacionados e as técnicas aplicadas. Áreas mais exploradas são a identificação de *motifs* em séries temporais e a agregação espaço-temporal baseada em região. Não foram observados trabalhos associados a identificação de *motifs* em séries espaço-temporais associadas a objetos permanentes. Observa-se, portanto, uma lacuna para estudo com amplo potencial de exploração.

Tabela 1: Comparação dos trabalhos relacionados

Trabalho	AETR	AETP	IMS	IMT
Ferreira et al. [2013]	X			
Andrienko and Andrienko [2008]	X			
Adrienko and Adrienko [2011]		X		
Cassisi et al. [2013]			X	
Jiang et al. [2008]			X	
Chi et al. [2012]			X	
Schneider et al. [2013]				X

Capítulo 4

Proposta

Este trabalho tem como objetivo explorar *motifs* em agregações de séries espaço-temporais de mobilidade urbana. O trabalho se divide nas fases de pesquisa, implementação e experimentação. Durante a fase de pesquisa, a fundamentação teórica é desenvolvida e aprimorada. A fase de implementação, por sua vez, se divide em três etapas principais: (i) seleção de objetos permanentes (estações), (ii) agregação espaço-temporal e (iii) identificação de *motifs*. A solução implementada é aplicada a uma coleção de dados na fase de experimentação. Posteriormente, faz-se, então, uma análise do significado e relevância dos *motifs* encontrados, confrontando alguns padrões que possam indicar a presença de pontos de interesse com suas reais localizações.

A seleção de estações tem como objetivo definir objetos permanentes (S) que atuem como sensores virtuais na agregação espaço-temporal. São considerados os dados de localização de todos os pontos de ônibus da cidade do Rio de Janeiro disponíveis no Portal de Dados Abertos da Prefeitura do Rio de Janeiro [dat, 2016]. Como algumas áreas tem-se maior presença de pontos de ônibus do que outras, estes pontos foram agrupados utilizando o algoritmo de agrupamento DBSCAN [Borah and Bhattacharyya, 2004]. Cada estação passa a ser representada por um *medoid* dos pontos presentes em cada grupo identificado. O objetivo desta técnica é evitar a interferência e sobreposição de pontos muito próximos na agregação espaço-temporal.

A etapa de agregação espaço-temporal é iniciada com a associação das observações dos objetos móveis a uma estação. Os objetos móveis que atuam como sensores de trajetória são os ônibus do município do Rio de Janeiro e seus dados também podem ser obtidos no Portal de Dados Abertos da Prefeitura do Rio de Janeiro [dat, 2016]. A agregação espaço-temporal considera a velocidade média, quantidade e lista de ônibus e de linhas próximos às estações, considerando-se as unidades de tempo estabelecidas. As agregações calculam estas grandezas tanto no geral quanto por sentido (norte, sul, leste, oeste). Ao fim dessa etapa, são obtidas séries espaço-temporais multivariadas para cada estação.

Uma vez que as séries espaço-temporais sejam produzidas, aplica-se a identificação de *motifs* sobre essas séries. Em função das propriedades destes dados, pretende-se desenvolver um algoritmo, inspirado em *random projection* [Li and Nallela, 2009], que introduza restrições

espaço-temporais na identificação destes padrões. Estas restrições espaço-temporais visam observar padrões que se propagam ao longo de uma vizinhança. Deste modo, espera-se que sejam emanados padrões capazes de responder as questões de pesquisa previamente mencionadas.

Capítulo 5

Estado Atual do Trabalho

Seguindo o processo de implementação descrito na Seção 4, inicialmente foram definidas as estações (1:Estações) que funcionam como pontos permanentes à agregação espaço-temporal. O primeiro passo é converter os dados geodésicos em um plano cartesiano. Para tanto, foi definido um ponto inicial $P_0(lat, long)$ referente a menor latitude e longitude do espaço amostral. Em seguida, os pontos de ônibus foram convertidos em pares ordenados, de modo que um ponto de ônibus $P_1(lat, long)$ seja convertido em $P(x, y)$ tal que x e y são a distância *Harversine* entre $P_2(lat_{P_0}, long_{P_0})$ e P_0 e a distância *Harversine* entre $P_3(lat_{P_1}, long_{P_1})$, respectivamente. Os valores de x e y estão representados na Figura 1.



Figura 1: Conversão dos dados de localização dos pontos de ônibus.



Figura 2: Estações obtidas aplicando o algoritmo de agrupamento DBSCAN sobre os pontos de ônibus do Rio de Janeiro.

Os pares ordenados obtidos com o processo de conversão dos dados de localização dos pontos de ônibus foram agrupados por DBSCAN, considerando *eps* de 300 metros e o mínimo de um ponto. A partir dos 7020 pontos de ônibus obtidos pelo Portal de Dados Abertos do Rio de Janeiro, foram geradas 307 estações. Na Figura 2, os marcadores vermelhos indicam as estações obtidas após o agrupamento dos pontos de ônibus. O tamanho do marcador é proporcional a quantidade de pontos que foram agrupados.

As observações de GPS sobre os ônibus da cidade do Rio de Janeiro devem ser associadas às estações, de modo que cada observação seja associada a estação mais próxima. Em seguida, os

dados foram agregados às estações, considerando um Δt de 1 minuto. No processo de agregação foram obtidos dados referentes a contagem de linhas e de ônibus, velocidade média de ônibus associados a estação e listagem das linhas e dos ônibus associados a estação no momento. Inicialmente, o processo foi executado sobre dados do dia 26 de junho de 2015. Das mais de 4 milhões de observações individuais de ônibus, foram geradas 278.307 observações agregadas.

Além da agregação considerando o $\Delta t = 1\text{minuto}$, serão realizadas agregações espaço-temporais considerando o $\Delta t = 4\text{minutos}$ e $\Delta t = 8\text{minutos}$. Após o processo de agregação (2:Agr.Temp.), teremos como resultado séries espaço-temporais associadas a pontos permanentes. Com as séries espaço-temporais definidas, o desenvolvimento de algoritmos para identificação de *motifs* (3:Id.Motifs) será realizada. A conclusão prevista para os algoritmos *motifs* é janeiro de 2017, abrindo espaço para experimentação (4:Aval.Exp.) e análise dos resultados obtidos (5:Ana.Result.).

As pesquisas e a produção textual estão sendo realizada em paralelo às demais etapas. A fundamentação teórica (6:Fund.Teo.) se encontra em fase de desenvolvimento e tem conclusão prevista para setembro de 2016. A solução proposta (7:Metodologia) neste trabalho será descrita de forma mais precisa e detalhada até janeiro de 2017. O experimento e seus resultados serão explorados após a etapa de experimentação, levando em conta a produção de artigo (8:Artigo) associado.

A defesa da dissertação (9:Defesa) está prevista para setembro de 2017. O cronograma, indicado na tabela 2, prevê as etapas abordadas neste trabalho. A tabela indica os meses de previsão de conclusão de cada uma das etapas.

Tabela 2: Cronograma de desenvolvimento da dissertação

Atividade	jun- jul	ago- set	out- nov	dez- jan	fev- mar	abr- mai	jun- jul	ago- set
1:Estações	X							
2:Agr.Temp.	X	X						
3:Id.Motifs			X	X				
4:Aval.Exp.				X				
5:Ana.Result.					X	X		
6:Fund.Teo.	X	X						
7:Metodologia		X	X	X				
8:Artigo					X	X	X	
9:Defesa								X

Referências Bibliográficas

- (2016). Portal de dados abertos da prefeitura do Rio de Janeiro. <http://data.rio/>.
- (2016). Tomtom Traffic Index. http://www.tomtom.com/pt_br/trafficindex/.
- Adrienko, N. and Adrienko, G. (2011). Spatial generalization and aggregation of massive movement data. *IEEE Transactions on visualization and computer graphics*, 17(2):205–219.
- Andrienko, G. and Andrienko, N. (2008). Spatio-temporal aggregation for visual analysis of movements. In *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*, pages 51–58. IEEE.
- Borah, B. and Bhattacharyya, D. (2004). An improved sampling-based dbscan for large spatial databases. In *Intelligent Sensing and Information Processing, 2004. Proceedings of International Conference on*, pages 92–96. IEEE.
- Cassisi, C., Aliotta, M., Cannata, A., Montalto, P., Patanè, D., Pulvirenti, A., and Spampinato, L. (2013). Motif discovery on seismic amplitude time series: The case study of mt etna 2011 eruptive activity. *Pure and Applied Geophysics*, 170(4):529–545.
- Chen, W., Guo, F., and Wang, F.-Y. (2015). A survey of traffic data visualization. *Intelligent Transportation Systems, IEEE Transactions on*, 16(6):2970–2984.
- Chi, L., Feng, Y., Chi, H., and Huang, Y. (2012). Face image recognition based on time series motif discovery. In *Granular Computing (GrC), 2012 IEEE International Conference on*, pages 72–77. IEEE.
- Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.
- Esling, P. and Agon, C. (2012). Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):12.
- Ferreira, N., Poco, J., Vo, H. T., Freire, J., and Silva, C. T. (2013). Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2149–2158.

- Frank, A. U. (2003). Ontology for spatio-temporal databases. In *Spatio-Temporal Databases*, pages 9–77. Springer.
- Jiang, T., Feng, Y., Zhang, B., Shi, J., and Wang, Y. (2008). Finding motifs of financial data streams in real time. In *International Symposium on Intelligence Computation and Applications*, pages 546–555. Springer.
- Li, L. and Nallela, S. (2009). Probabilistic discovery of motifs in water level. In *Information Reuse & Integration, 2009. IRI'09. IEEE International Conference on*, pages 388–393. IEEE.
- Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., and González, M. C. (2013). Unraveling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246.
- Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., and Vangenot, C. (2008). A conceptual view on trajectories. *Data Knowl. Eng.*, 65(1):126–146.
- Tao, Y., Kollios, G., Considine, J., Li, F., and Papadias, D. (2004). Spatio-temporal aggregation using sketches. In *Data Engineering, 2004. Proceedings. 20th International Conference on*, pages 214–225. IEEE.
- Vahdatpour, A., Amini, N., and Sarrafzadeh, M. (2009). Toward unsupervised activity discovery using multi-dimensional motif detection in time series. In *IJCAI*, volume 9, pages 1261–1266.
- Wang, L., Wang, Z., and Liu, S. (2016). An effective multivariate time series classification approach using echo state network and adaptive differential evolution algorithm. *Expert Systems with Applications*, 43:237–249.