Análise Espacial de Dados Geográficos

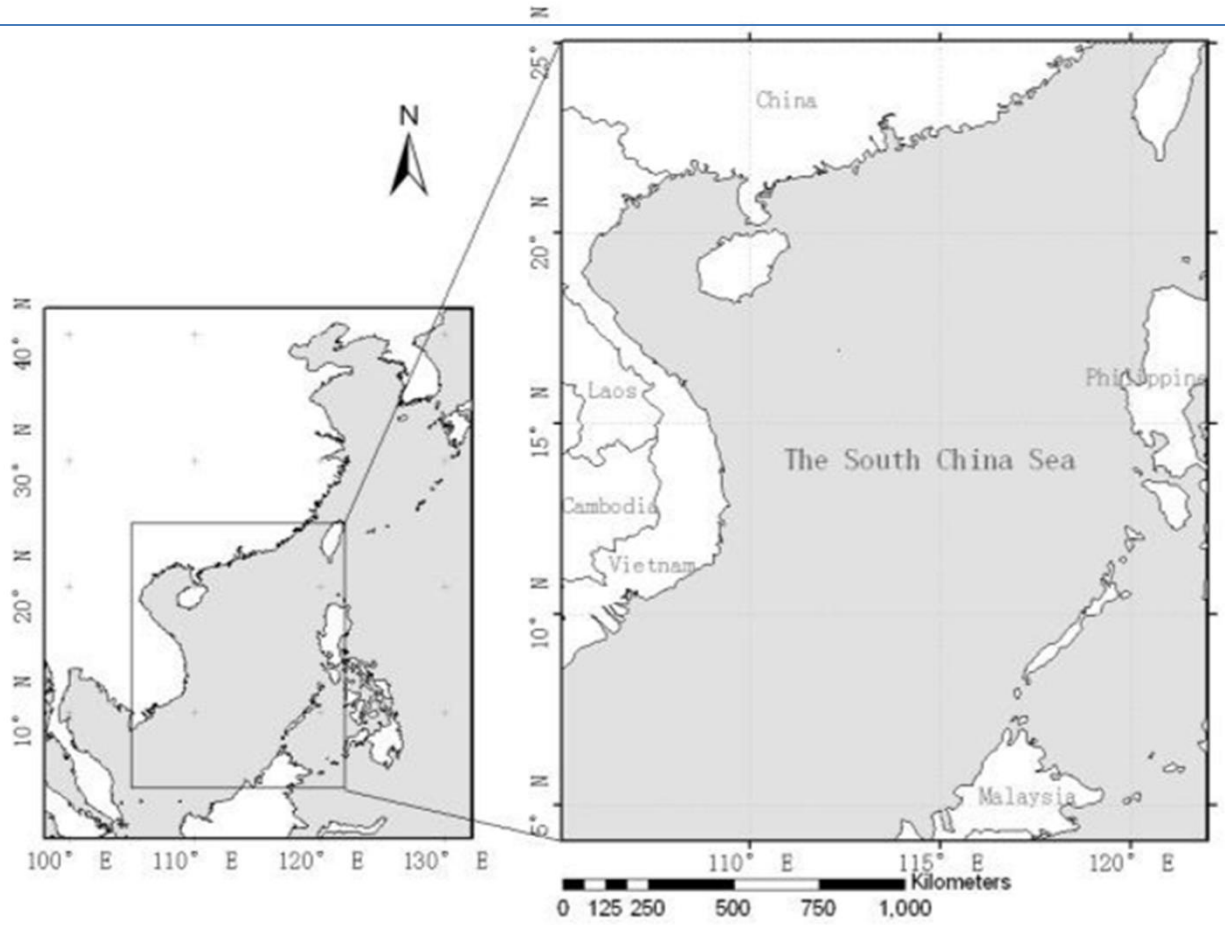
SER301-2018

Aluno: Gabriel Moiano Cesar

Satellite Remotely-sensed Analysis of Temporal spatial Variations of Chlorophyll-a Concentration in South China Sea

WANG et al. 2011

Área de Estudo



Posição Geográfica: 4 ° N ~ 25° N, 105 ° E ~ 122 °

Banco de dados

Dados de Clorofila-a do sensor Sea-viewing Wide Field-of view (SeaWiFS)

Média Mensal (Level-3)

Resolução espacial de 9km

Período de Setembro 1997 a Dezembro de 2010.

HDF → Registro +correções → Raster

Softwares: SeaDAS MATLAB

Análise por agrupamento (Clustering)

ISODATA (Iterative Self-Organizing Data Analysis Technique)

-Baseado na densidade do GRID;

-Descobre a distribuição espacial de objetos;

-Usando informações múltiplas contidas em dados espacialmente distribuídos;

PARÂMETROS:

número de iterações: 100 - tamanho mínimo da classe: 20 - intervalo da amostra: - fração de rejeição: 0 - ponderação de probabilidade a priori: igual.

Análise de tendências

Weatherhead et al., 1998

$$Y_t = A_t + B_t * X_t + C_t \sin(2\frac{\pi}{D_t} X + E_t) + N_t.$$

Usado para quantificar as características de expressão da distribuição espacial da concentração de clorofila-a;

Onde:
Y = média mensal de concentração de clorofila-a;
X = mês;
t= diferentes zonas;
A_t, B_t, C_t, D_t e E_t = parâmetros do modelo.

Análise de tendências

$$Y_t = A_t + B_t * X_t + C_t \sin(2\frac{\pi}{D_t} X + E_t) + N_t.$$

Calcula a tendência da variação

Análise de tendências

$$Y_t = A_t + B_t * X_t + C_t \sin(2\frac{\pi}{D_t} X + E_t) + N_t.$$

Calcula a variação cíclica mensal

Análise de tendências

$$Y_t = A_t + B_t * X_t + C_t \sin(2\frac{\pi}{D_t} X + E_t) + N_t.$$

Erro residual da simulação

Dinâmica espaço-temporal da concentração de clorofila-a

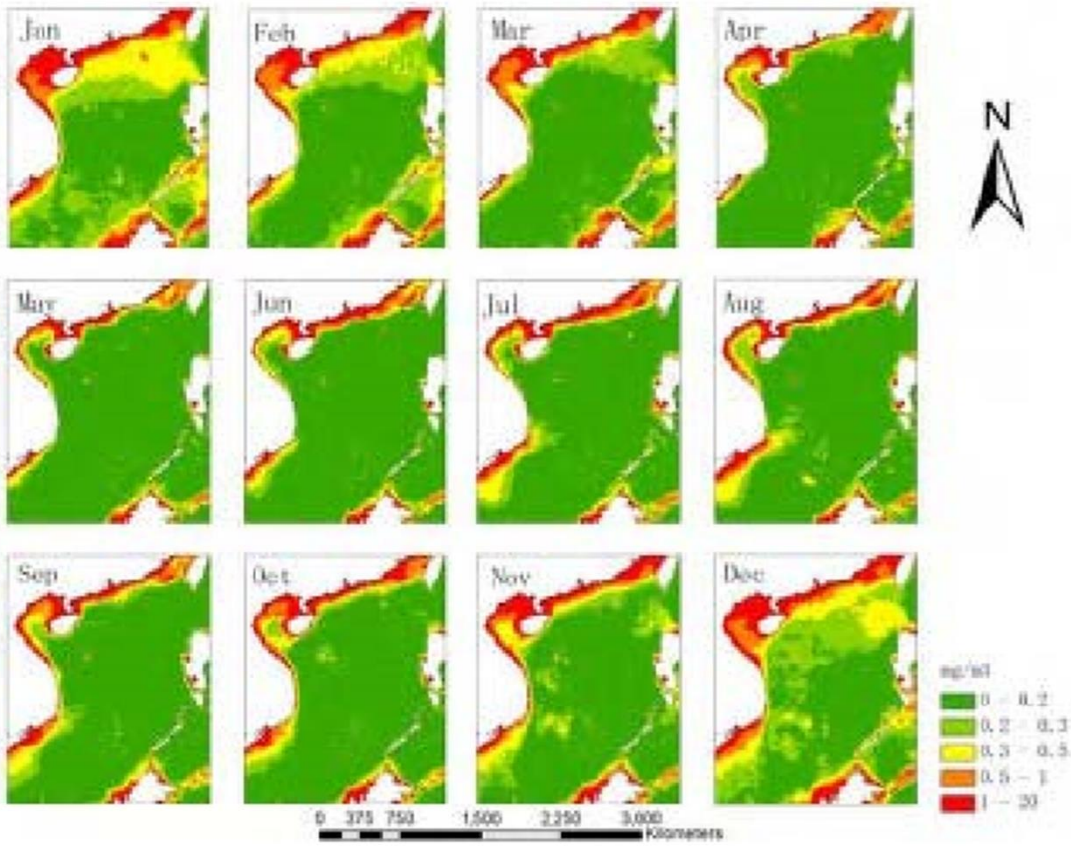


Figure 2. Monthly average of chlorophyll-a concentration of many years.

Tendência de variação média

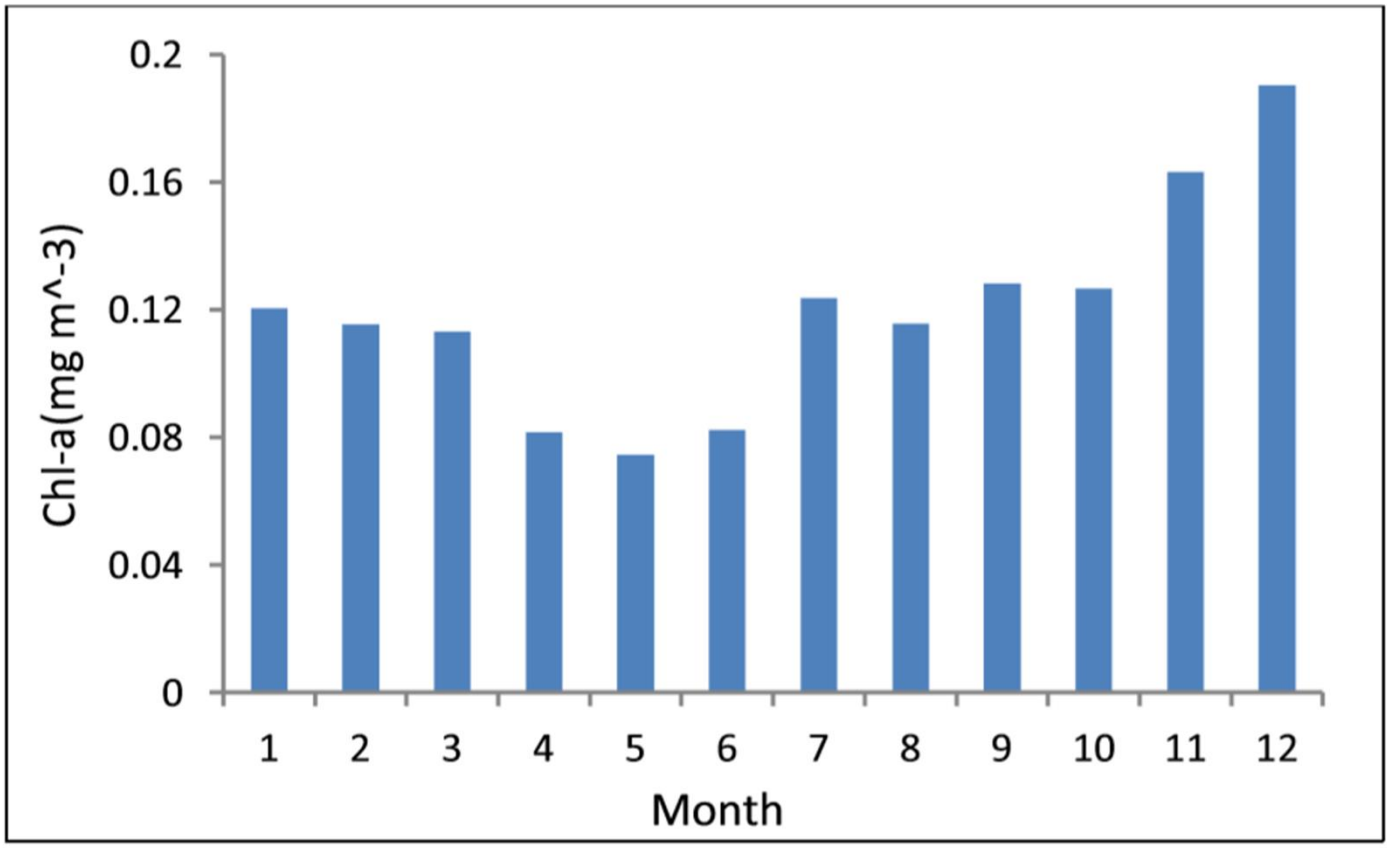


Figure 3. Monthly average variation trend of chlorophyll-a concentration.

Análise de zoneamento

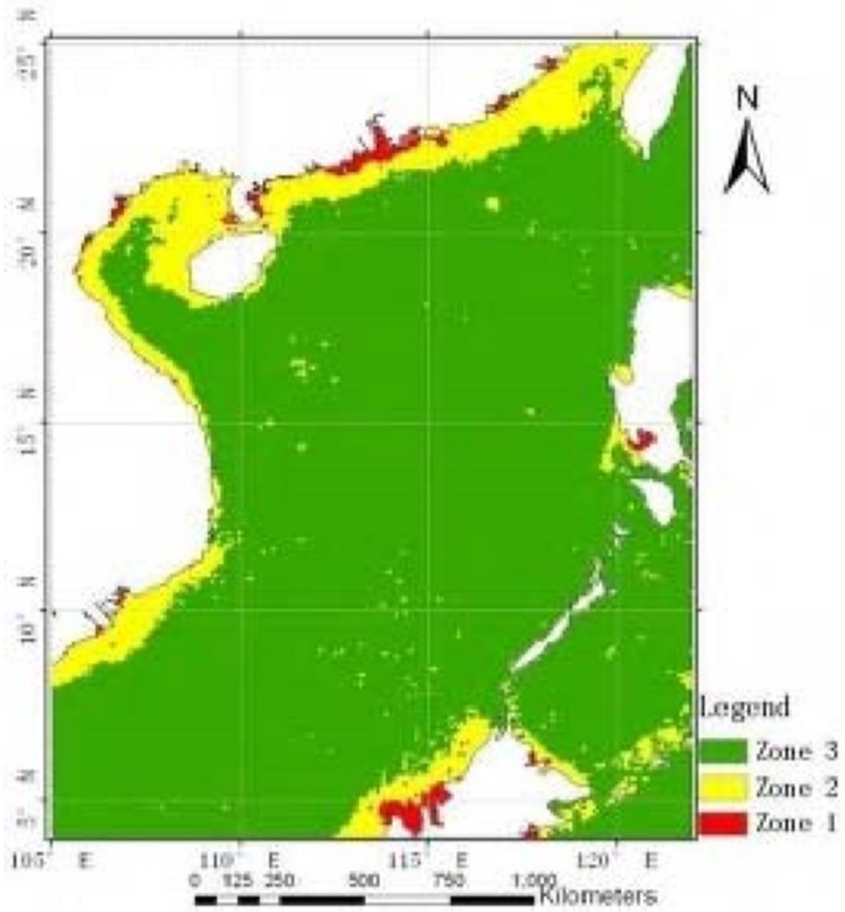


Figure 4. Spatial clustering result of chlorophyll-a concentration.

Análise de zoneamento

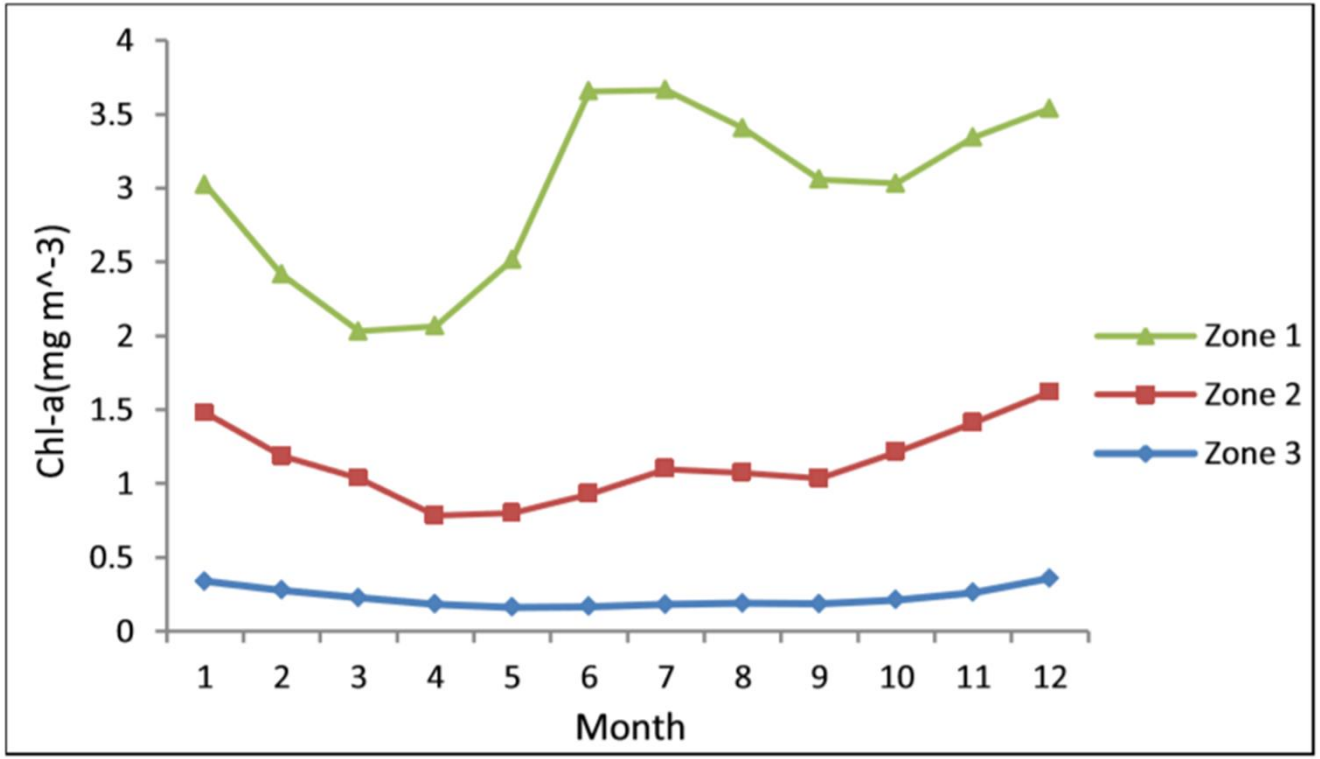


Figure 5. Monthly average of chlorophyll-a concentration in each zone.

Variação mensal da concentração de clorofila-a por zona

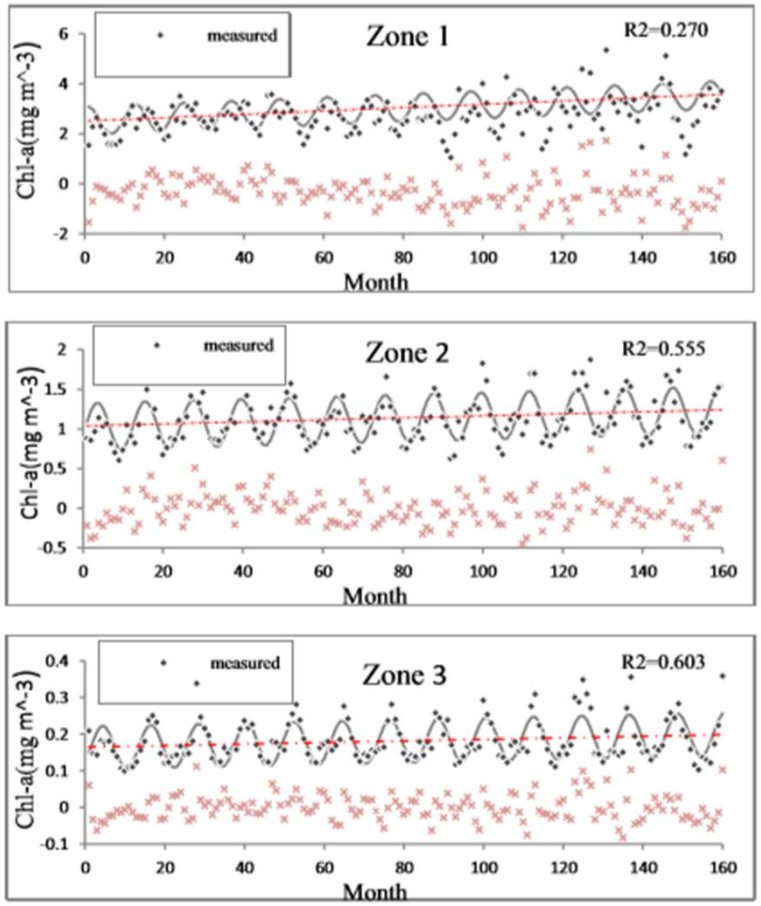


Figure 6. Fitting results of monthly average chlorophyll-a concentration in five zones.

2.517 mg m-3

1.028 mg m-3

0.162 mg m-3

Críticas

1 Citação do algoritmo usado para o cálculo da concentração de clorofila-a

2 ISODATA é a melhor técnica de agrupamento espacial?

3 Explorar melhor os resultados e discussão.

4 Retirada de outliers poderia “melhorar” o resultado do modelo.

Críticas

5

used in the present paper is the Level-3 monthly average products of 9km-resolution from September 1997 to December 2010.

C. Pocedure

1) Data processing

The raw data was in the HDF format. Through the coordinate registration, mask, computing and other treatments, we got the raster files which are more suited for spatial analysis. Influencing by environment such as cloud, the remote sensing data have many outliers. Compared by test, the inverse distance weighted interpolation method (IDW) can eliminate the outliers best (The correlation of IDW interpolation results of test points and in situ data: R^2=0.807).

The raw data of chlorophyll-a with the HDF format were converted to raster through coordinates correction, mask, computing and other treatments, we got the raster files which are more suited for spatial analysis. Influencing by environment such as cloud, the remote sensing data have many outliers. Compared by test, the inverse distance weighted interpolation method (IDW) can eliminate the blank value best (The correlation of IDW interpolation results of test points and in situ data: R^2=0.807).

2) Cluster analysis

Spatial clustering is a kind of unsupervised data mining. According to a large number of location-related space information contained in data, it could discover the distribution of spatial objects. This paper analysis the research area used one of the spatial clustering methods based on grid density named ISODATA (Iterative Self-Organizing Data Analysis Technique).

In view of the overall situation of Chinese coast, we got the monthly average in the whole Chinese coastal areas from concentration in every zones, B_t can be quantitatively compare the monthly changes of chlorophyll-a concentration in different zones). B_t / A_t can be used to indicate that the rate of change of chlorophyll-a concentration in the zones. $C_t \sin(2\frac{\pi}{D_t} X + E_t)$ can be used to describe the chlorophyll-a concentration changes of monthly mean of the cyclical(Where, C_t is the amplitude of variation, D_t is the change in period, the ideal period is 12 months, E_t is the initial phase).

Nt means the residual error between actual values and simulated values. Weatherhead.et al regards *Nt* has autocorrelation and can be described as below:

$$N_t = \Phi N_{t-1} + \varepsilon_t \quad (2)$$

In (2), Φ means the autocorrelation among residual errors. ε_t is the noise of autocorrelation. The autocorrelation among residual will influence the precision of chlorophyll-a concentration variation trend. Weatherhead.et al think the precision of Model simulated trend (σ_B) is function of autocorrelation (Φ), time span (T) and standard deviation of residual (σ_N), and can be approximate represents as:

$$\sigma_B \approx \frac{\sigma_N}{T^{3/2}} \sqrt{\frac{1+\Phi}{1-\Phi}}. \quad (3)$$

Computational formula of standard deviation is:

$$\sigma_N = \sqrt{\frac{\sum_{j=1}^{T} (Y_j - \bar{Y})}{T}}. \quad (4)$$

If trend B_t and precision σ_B meet the condition $\left|\frac{Bt}{\sigma_B}\right| > 2$, we can consider that month average of chlorophyll-a concentration variation trend (Bt) is notable when the

Críticas

used in the present paper is the Level-3 monthly average products of 9km-resolution from September 1997 to December 2010.

C. Pocedure

1) Data processing

The raw data was in the HDF format. Through the coordinate registration, mask, computing and other treatments, we got the raster files which are more suited for spatial analysis. Influencing by environment such as cloud, the remote sensing data have many outliers. Compared by test, the inverse distance weighted interpolation method (IDW) can eliminate the outliers best (The correlation of IDW interpolation results of test points and in situ data: R^2=0.807).

The raw data of chlorophyll-a with the HDF format were converted to raster through coordinates correction, mask, computing and other treatments, we got the raster files which are more suited for spatial analysis. Influencing by environment such as cloud, the remote sensing data have many outliers. Compared by test, the inverse distance weighted interpolation method (IDW) can eliminate the blank value best (The correlation of IDW interpolation results of test points and in situ data: R^2=0.807).

2) Cluster analysis

Spatial clustering is a kind of unsupervised data mining. According to a large number of location-related space information contained in data, it could discover the distribution of spatial objects. This paper analysis the research area used one of the spatial clustering methods based on grid density named ISODATA (Iterative Self-Organizing Data Analysis Technique).

In view of the overall situation of Chinese coast, we got the monthly average in the whole Chinese coastal areas from

concentration in every zones, B_t can be quantitatively compare the monthly changes of chlorophyll-a concentration in different zones). B_t/A_t can be used to indicate that the rate of change of chlorophyll-a concentration in the zones. $C_t \sin(2\frac{\pi}{D_t}X + E_t)$ can be used to describe the chlorophyll-a concentration changes of monthly mean of the cyclical(Where, C_t is the amplitude of variation, D_t is the change in period, the ideal period is 12 months, E_t is the initial phase).

Nt means the residual error between actual values and simulated values. Weatherhead.et al regards *Nt* has autocorrelation and can be described as below:

$$N_t = \Phi N_{t-1} + \varepsilon_t \quad (2)$$

In (2), Φ means the autocorrelation among residual errors. ε_t is the noise of autocorrelation. The autocorrelation among residual will influence the precision of chlorophyll-a concentration variation trend. Weatherhead.et al think the precision of Model simulated trend (σ_B) is function of autocorrelation (Φ), time span (T) and standard deviation of residual (σ_N), and can be approximate represents as:

$$\sigma_B \approx \frac{\sigma_N}{T^{3/2}}\sqrt{\frac{1+\Phi}{1-\Phi}}. \quad (3)$$

Computational formula of standard deviation is:

$$\sigma_N = \sqrt{\frac{\sum_{j=1}^{T}(Y_j - \bar{Y})}{T}}. \quad (4)$$

If trend B_t and precision σ_B meet the condition $\left|\frac{Bt}{\sigma_B}\right| > 2$, we can consider that month average of chlorophyll-a concentration variation trend (Bt) is notable when the

Variação mensal da concentração de clorofila-a por zona

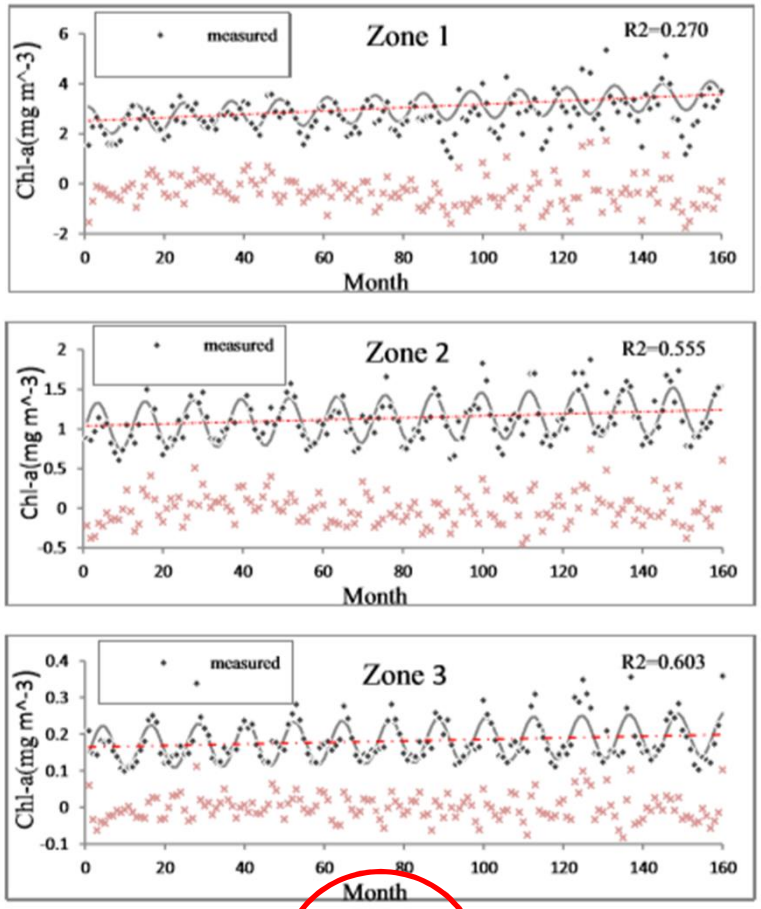


Figure 6. Fitting results of monthly average chlorophyll-a concentration in five zones.

2.517 mg m-3

1.028 mg m-3

0.162 mg m-3

Sugestão de técnica

Research Track Paper

SCAN: A Structural Clustering Algorithm for Networks

Xiaowei Xu
University of Arkansas at Little Rock
xwxu@ualr.edu

Nurcan Yuruk, Zhidan Feng
University of Arkansas at Little Rock
{nxyuruk, zxfeng@ualr.edu}

Thomas A. J. Schweiger
Acxiom Corporation
Tom.Schweiger@acxiom.com

ABSTRACT

Network clustering (or graph partitioning) is an important task for the discovery of underlying structures in networks. Many algorithms find clusters by maximizing the number of intra-cluster edges. While such algorithms find useful and interesting structures, they tend to fail to identify and isolate two kinds of vertices that play special roles – vertices that bridge clusters (hubs) and vertices that are marginally connected to clusters (outliers). Identifying hubs is useful for applications such as viral marketing and epidemiology since hubs are responsible for spreading ideas or disease. In contrast, outliers have little or no influence, and may be isolated as noise in the data. In this paper, we proposed a novel algorithm called SCAN (Structural Clustering Algorithm for Networks), which detects clusters, hubs and outliers in networks. It clusters vertices based on a structural similarity measure. The algorithm is fast and efficient, visiting each vertex only once. An empirical evaluation of the method using both synthetic and real datasets demonstrates superior performance over other methods such as the modularity-based algorithms.

Categories and Subject Descriptors

I.5.3 [**PATTERN RECOGNITION**]: Clustering – *Algorithms, Similarity measures.*

General Terms

Algorithms, Performance

Keywords

Network clustering, Graph partitioning, Community Structure, Hubs, Outliers

1. INTRODUCTION

Much data of current interest to the scientific community can be modeled as networks (or graphs). A network is sets of vertices, representing objects, connected together by edges, representing the relationship between objects. For example, a social network can be viewed as a graph where individuals are represented by vertices; and the friendship between individuals are edges [1]. Similarly, the world-wide web can be modeled as a graph, where web pages are represented as vertices that are connected by an edge when one pages contains a hyperlink to another [2] [3].

Network clustering (or graph partitioning) is a fundamental approach for detecting hidden structures in networks that, because computer science [4][5], physics [11], and bioinformatics [6]. Various methods have been developed. These methods tend to cluster networks such that there are a dense set of edges within every cluster and few edges between clusters. Modularity-based algorithms [6][11][12] and normalized cut [4][5] are successful examples. However, they do not distinguish the roles of the vertices in the networks. Some vertices are members of clusters; some vertices are hubs that bridge many clusters but don't belong to any, and some vertices are outliers that have only a weak association with a particular cluster. The situation is illustrated in Figure 1.

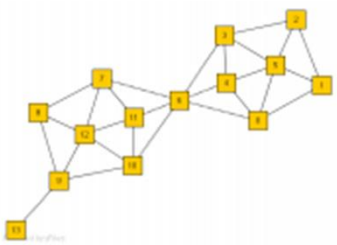


Figure 1. A Network with 2 Clusters, a Hub and an Outlier.

The existing methods such as modularity-based algorithm [12] will partition this example into two clusters: one consisting of vertices 0 to 6 and the other consisting of vertices 7 to 13. They do not isolate vertex 6, a hub whose membership in either cluster is disputable, or vertex 13, which has only a single connection to the network.

The identification and isolation of hubs is essential for many applications. As an example, the identification of hubs in the WWW improves the search for relevant authoritative web pages [7]. Furthermore, hubs are believed to play a crucial role in viral marketing [8] and epidemiology [9].

In this paper, we propose a new method for network clustering called SCAN (Structural Clustering Algorithm for Networks). The goal of our method is to find clusters, hubs, and outliers in large networks. To achieve this goal, we use the neighborhood of the vertices as clustering criteria instead of only their direct connections. Vertices are grouped into the clusters by how they

Método baseado em redes sociais.

Trabalha com a interação entre vizinhança.

Identifica conectores e outliers.

Vantagens

- É aplicado em dados de “larga escala”;

- Alta eficiência;

- Computacionalmente viável.

Desvantagem

Ignora a variação temporal.

Effectiveness

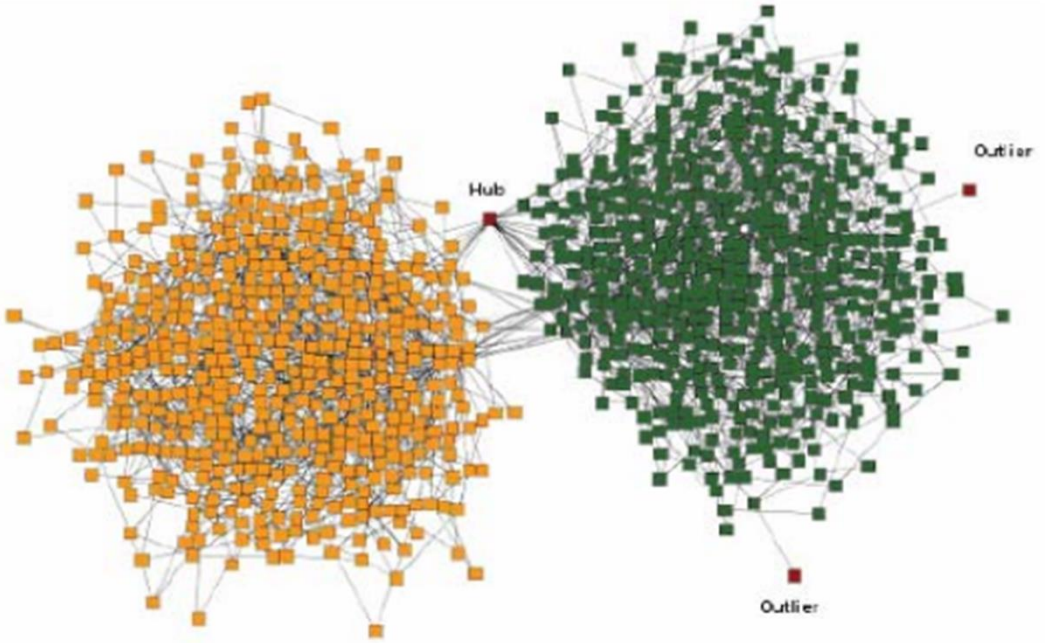


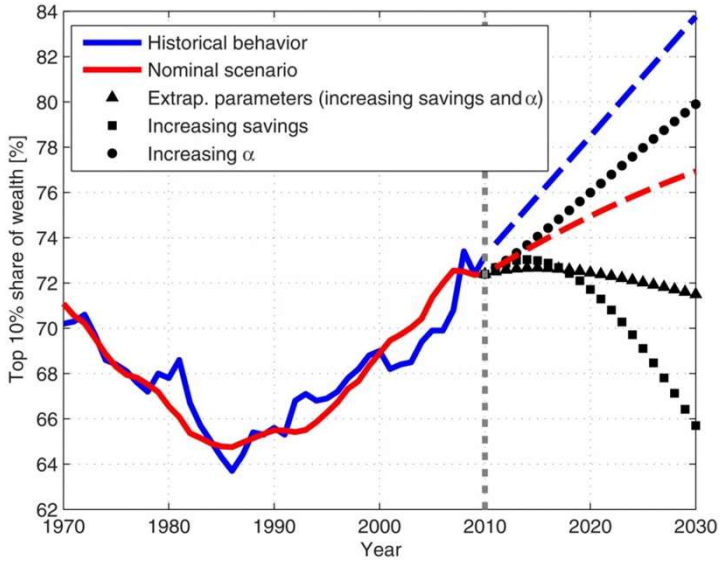
Figure 3. A Synthetic Graph with 1,000 Vertices

Census x-11

Técnica matemática desenvolvida na década de 30, para o estudo de padrões temporais (National Bureau of Economic Research in the US - Fredrick R. Macaulay).

1) Cálculo de média móvel
2) Retirada da tendência por componentes irregulares
3) Estimativa de cenários futuros

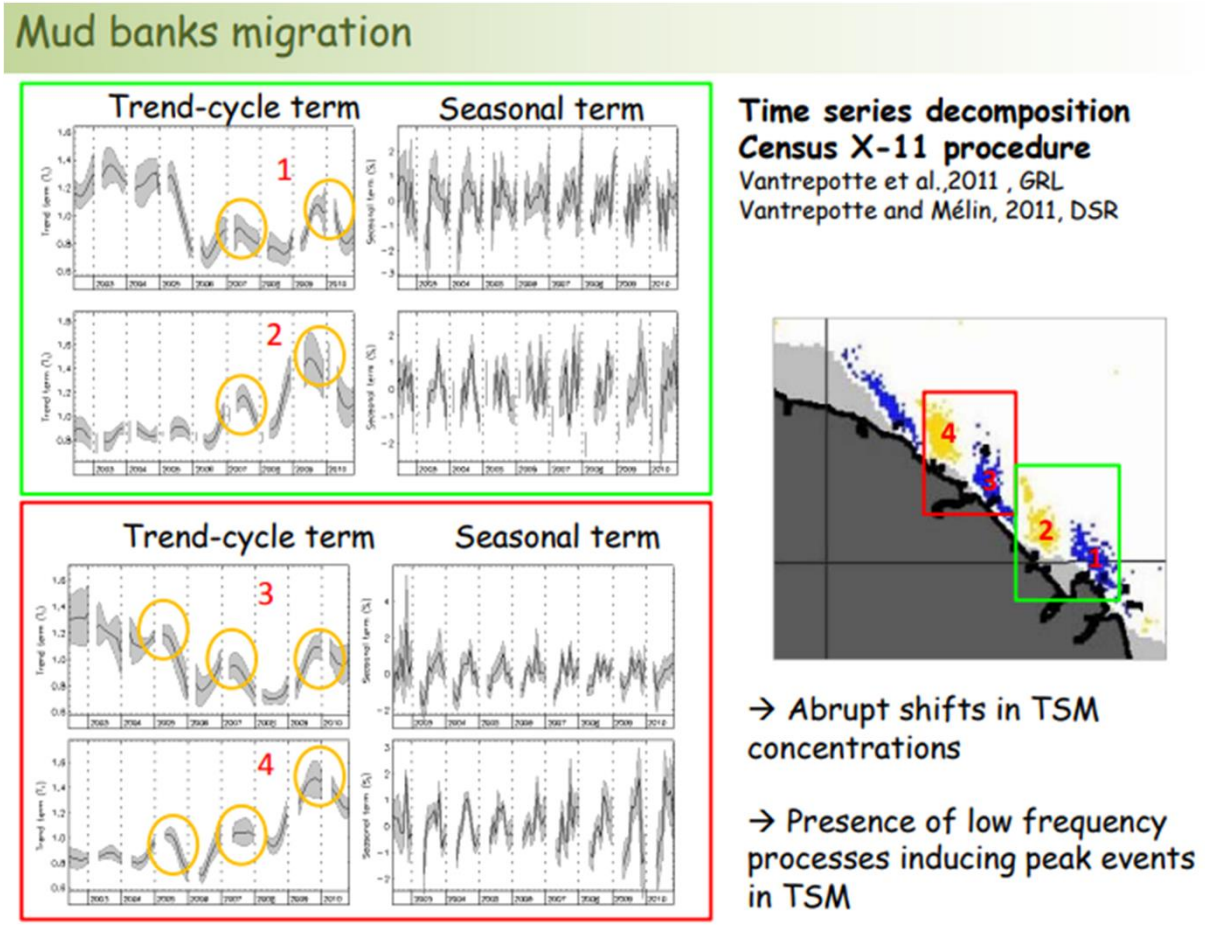
Trabalhos de Hubert Loisel e Vincent Vantrepotte



Fonte: Berman et al. (2015)

Modelo

Census x-11

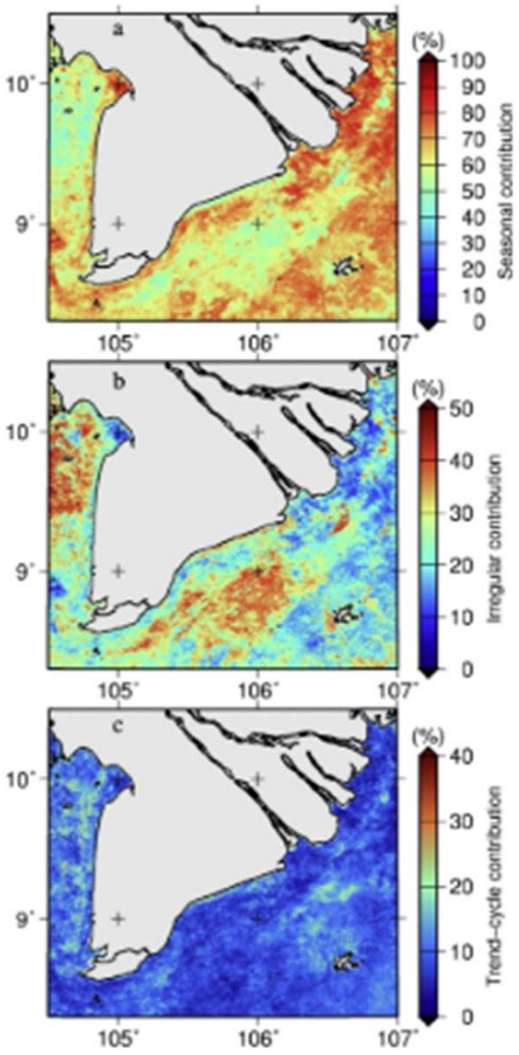


Time series decomposition
Census X-11 procedure
Vantrepotte et al.,2011 , GRL
Vantrepotte and Mélin, 2011, DSR

→ Abrupt shifts in TSM concentrations

→ Presence of low frequency processes inducing peak events in TSM

Gardel e Gratiot (2005)



Vantrepotte e Mélin (2011)

Referências Bibliográficas

Berman, Yonatan, Yoash Shapira, and Eshel Ben-Jacob. 2015. “Modeling the Origin and Possible Control of the Wealth Inequality Surge.” **PLOS ONE**, 10(6): e0130181.

Gardel, A. & Gratiot, N. A satellite image-based method for estimating rates of mud bank migration, French Guiana, **South America. J. Coast. Res**. 21, 720–728 (2005).

Vantrepotte, V., & Mélin, F. (2011). Inter-annual variations in the SeaWiFS global chloro-phyll a concentration (1997–2007). **Deep Sea Research** I.http://dx.doi.org/10.1016/j.dsr.2011.02.003.

Wang, J.; Jiang, H. Satellite remotely-sensed analysis of temporal-spatial variations of chlorophyll-a concentration in South China Sea, **19th International Conference on Geoinformatics**, 2011. DOI: 10.1109/GeoInformatics.2011.5980882

Weatherhead, E. C.; Reinsel, G, C, Tiao, G. Factors affecting the detection of trends: statistical considerations and applications to environmental data. **J Geophys Res**, vol. 103, 1998, pp. 17149-17161.

Xu, X.; Yuruk, N.; Geng, Z. and Schweiger, T. A. J. SCAN: A Structural Clustering Algorithm for Networks. **In Proc. KDD**, pages 824–833, 2007.

Obrigado!