Tạp chí Khoa học ĐHQGHN, Khoa học Tự nhiên và Công nghệ 28, Số 3S (2012) 71-77

Phân tích thành phần chính áp dụng vào tập số liệu mực nước biển các trạm dọc bờ Việt Nam

Phạm Văn Huấn*

Trường Đại học Khoa học Tự nhiên, ĐHQGHN, 334 Nguyễn Trãi, Hà Nội, Việt Nam

Nhận ngày 15 tháng 6 năm 2012

Tóm tắt. Triển khai ứng dụng phương pháp khai triển thành phần chính để phân loại các quá trình dao động mực nước biển trên toàn dải bờ Việt Nam. Đã xác định được một cách định lượng những nhóm trạm có tính đồng nhất về đặc điểm dao động mực nước biển làm cơ sở ứng dụng phương pháp khôi phục và dự báo hiệu quả mực nước biển cho mỗi trạm quan trắc thuộc các đoạn bờ biển khác nhau. Phân tích các ma trận chuyển tiếp cho thấy những đặc điểm chính quyết định sự khác nhau trong dao động với thời gian của mực nước biển tại nhóm trạm mực nước bao gồm tính chất thủy triều, ảnh hưởng thủy triều nước nông, diễn biến dòng nước sông ảnh hưởng tương ứng với hình thái vùng bờ và địa hình khu vực lân cận trạm được xét. Ví dụ khôi phục mực nước cho một trong ba nhóm trạm dọc bờ Việt Nam cho kết quả khá tốt, có triển vọng áp dung thực tế khôi phục và dự báo mực nước biển cho các vùng bờ khác nhau.

1. Mở đầu

Phương pháp phân tích thành phần chính (*principal component analysis*) (xem [1,2]) là công cụ toán thống kê vạn năng được áp dụng rất hiệu quả khi xử lý thông tin từ các ma trận số liệu quan trắc. Nếu có bảng giá trị của các yếu tố theo thời gian, người phân tích nhận ra được những nhóm yếu tố có cùng kiểu biến thiên thời gian khác với các nhóm yếu tố khác, tức nhận định được bản chất nguồn gốc của mỗi nhóm yếu tố. Nếu xét bảng giá trị của một yếu tố được quan trắc trong thời gian ở nhiều điểm không gian, người phân tích nhận ra các nhóm điểm có cùng kiểu biến thiên thời gian trong khi các nhóm điểm khác có cách biến thiên khác, từ đó đưa ra thông tin phân vùng khách quan. Nếu số liệu được sắp xếp thành bảng giá trị của các yếu tố tại một số điểm trong không gian, thì người ta nghiên cứu được quy mô ảnh hưởng của từng yếu tố và khả năng khái quát sự hiệp biến của các yếu tố.

Gần đây đã bắt đầu xuất hiện các nghiên cứu trong ngành khoa học trái đất có áp dụng phương pháp thành phần chính do nhu cầu phân tích một cách định lượng đối với số liệu quan trắc môi trường, kinh tế, xã hội [3].

Nhiệm vụ khôi phục số liệu hay dự tính mực nước ở các vùng bờ biển, nhất là đối với các đoạn bờ thưa trạm mực nước, rất hay được đặt ra trong thực tế. Theo truyền thống, bảng thủy triều dự báo mực nước cho các cảng phụ

* ĐT: 84-912116661
E-mail: pvhuan@viettel.vn

được thực hiện theo phương pháp so sánh, dựa vào quan hệ tương quan thực nghiệm giữa mực thủy triều trạm phụ – trạm chính gần và thuộc cùng một vùng biển có tính chất thủy triều giống nhau. Đó là cách giải quyết vấn đề phụ thuộc nhiều vào kinh nghiệm chủ quan của người thực hiện và độ dài của chuỗi quan trắc ở trạm phụ cần nghiên cứu.

Bài này áp dụng phương pháp phân tích thành phần chính cho phép sử dụng thông tin từ tập số liệu quan trắc mực nước biển trên vùng không gian lớn toàn dải bờ Việt Nam để rút ra nhận định về các quá trình chủ yếu quyết định sự biến thiên (dao động) của mực nước ở từng trạm, nhóm trạm, nó cho phép nhận ra những nhóm trạm có cùng kiểu dao động và quy mô dao động. Từ đó quyết định phương pháp dự báo mực nước, hay khôi phục mực nước cho trạm bất kỳ dựa vào giá trị của mực nước ở các trạm khác cùng kiểu.

Trong mục 1 sẽ giới thiệu tóm tắt về cơ sở phương pháp, thiên về trình bày tuần tự các bước tính toán thực tế của phương pháp giúp người mới áp dụng biết cách thực hiện. Mục 2 thông báo kết quả ứng dụng, phân tích ý nghĩa kết quả tính để chứng minh tính hiệu quả của phương pháp trong nghiên cứu mực nước biển.

2. Cơ sở phương pháp và thủ tục tính toán của phân tích thành phần chính

Về phương diện toán học, mô hình phương pháp các thành phần chính được phát biểu như sau: Giả sử có một tập dữ liệu gồm N quan trắc về M biến. Có nghĩa là có N vectơ quan trắc dạng $\mathbf{x} = \{x_1, x_2, ..., x_M\}$ làm thành ma trận dữ liệu X gồm N dòng và M cột. Tương ứng với ma trận dữ liệu này sẽ là một ma trận các giá trị của những thành phần chính **F** cũng gồm N dòng và M cột. Mỗi cột trong ma trận **F** là một thành phần chính, một vectơ dạng $\mathbf{f}_j = \{f_1, f_2, ..., f_M\}$ [1]. Số biến M và số thành phần chính bằng nhau. Số quan trắc N của dữ liệu xuất phát và số giá trị của từng thành phần chính bằng nhau. Khi đó công thức khai triển thành các thành phần chính có dạng

$$\mathbf{x} = \mathbf{F} \cdot \mathbf{A}^{\mathrm{T}}. \qquad (1)$$

Ở đây ma trận **A** là ma trận các hệ số liên hệ giữa các biến và các thành phần chính. Ma trận **A** chứa M dòng và M cột. Mỗi cột của ma trận **A** chứa các hệ số liên hệ của thành phần chính đang xét với tất cả các biến xuất phát. Ký hiệu $\mathbf{A}^{\mathrm{T}}$ trong công thức (1) chỉ ma trận chuyển vị của **A**.

Có thể viết lại công thức của phương pháp các thành phần chính đối với quan trắc thứ i của biến j trong ma trận dữ liệu như sau:

$$x_{i,j} = \sum_{k=1}^{M} a_{i,k} f_{k,j}. \qquad (2)$$

Để tìm thành phần chính trước hết cần xác định những hệ số liên hệ của từng biến j với từng thành phần k, những hệ số này tạo thành ma trận **A** – ma trận các hệ số khai triển (hay còn gọi là ma trận chuyển đổi).

Theo toán thống kê, thủ tục tìm ma trận A quy về việc tìm các vectơ riêng của ma trận tương quan R của các biến quan trắc.

Tìm các vectơ riêng bắt đầu bằng việc tìm các giá trị riêng của ma trận tương quan nhờ giải phương trình đặc trưng:

$$|\mathbf{R} - \Lambda\mathbf{I}| = 0. \qquad (3)$$

trong đó Λ – vectơ các số riêng; **I** – vectơ đơn vị, tức tìm M nghiệm của phương trình đặc trưng đối với định thức của ma trận tương quan:

$$\begin{bmatrix} 1-\lambda & r_{1,2} & ... & r_{1,M} \\ r_{2,1} & 1-\lambda & ... & r_{2,M} \\ ... & ... & ... & ... \\ r_{M,1} & r_{M,2} & ... & 1-\lambda \end{bmatrix} = 0 \qquad (4)$$

Bước thứ hai - giải hệ các phương trình tuyến tính để xác định ma trận các vectơ riêng:

$$\mathbf{A} \cdot (\mathbf{R} - \Lambda\mathbf{I}) = 0, \qquad (5)$$

tức M lần giải hệ phương trình đại số dạng

$$\begin{aligned} &a_{i,1}(1-\lambda_i) + a_{i,2}r_{1,2} + ... + a_{i,M}r_{1,M} = 0 \\ &a_{i,1}r_{2,1} + a_{i,2}(1-\lambda_i) + ... + a_{i,M}r_{2,M} = 0 \\ &..... .... \\ &a_{i,1}r_{M,1} + a_{i,2}r_{M,2} + ... + a_{i,M}(1-\lambda_i) = 0 \end{aligned}$$

Trong mỗi lần giải, ta đưa một giá trị riêng λ_i vào hệ phương trình trên đây và nhận được M nghiệm – đó chính là những trị số của vectơ riêng i làm thành một cột của ma trận **A**.

Sau khi tìm được các số riêng và các vectơ riêng A, nhiệm vụ đánh giá cuối cùng gồm:

Thứ nhất, đánh giá tầm quan trọng của từng thành phần chính. Đóng góp tương đối của thành phần chính thứ i vào phương sai chung của các biến theo công thức

$$d_i = \frac{\lambda_i}{\sum_{j=1}^{M}\lambda_j} \qquad (6)$$

Thứ hai, tính những giá trị của từng thành phần chính, tức các vectơ $\mathbf{f}_j$. Ở đây áp dụng phương pháp hồi quy tuyến tính. Thủ tục này gồm các bước:

a) Tính ma trận các hệ số hồi quy **B** trên cơ sở vectơ các giá trị riêng (Λ) và ma trận các vectơ riêng **A** theo công thức

$$\mathbf{B} = \Lambda^{1/2}\,\mathbf{C}, \text{ với } \mathbf{C} = (\mathbf{A}^{\mathrm{T}})^{-1}.$$

Viết cho từng phần tử, công thức này có dạng

$$b_{i,j} = \sqrt{\lambda_j}\, c_{i,j}. \qquad (7)$$

b) Khôi phục ma trận các thành phần chính theo công thức

$$\mathbf{F} = \mathbf{X} \cdot \mathbf{B},$$

hay viết cho từng phần tử

$$f_{i,j} = \sum_{k=1}^{M} x_{i,k}b_{k,j} \;\; (i = 1,..N;\; j = 1,..M), \qquad (8)$$

trong đó k – số hiệu của biến xuất phát.

Tùy mục đích ứng dụng phương pháp, ta có thể thực hiện tất cả các bước tính toán trên đây hoặc chỉ cần tính toán tới ma trận chuyển đổi đã có thể được nhiều thông tin để phân tích.

3. Phân tích mực nước ven bờ Việt Nam, khôi phục và phân loại dao động mực nước các vùng

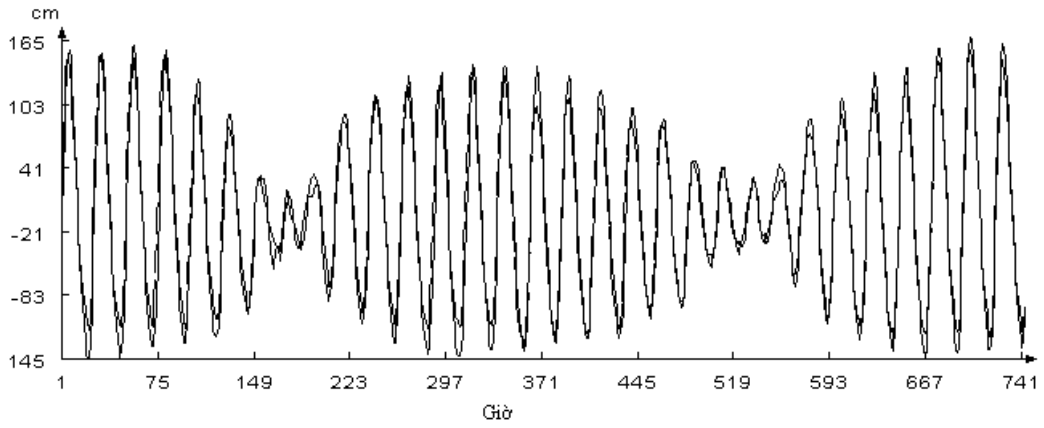
Sử dụng số liệu mực nước quan trắc từng giờ tại 15 trạm dọc bờ Việt Nam trong 5 năm (2002-2006). Bảng số liệu tạo thành ma trận 15 cột ứng với 15 trạm theo xu hướng phân bố trạm từ phía bắc tới phía nam và 43824 dòng, mỗi dòng ứng với một giờ quan trắc bắt đầu từ 0 giờ ngày 1/1/2002 đến 23 giờ ngày 31/12/2006.

Kết quả tính theo sơ đồ đã giới thiệu ở mục 1 được ghi tóm tắt thành bảng 1. Trong bảng này, dòng thứ hai ghi tỉ phần đóng góp của từng thành phần chính vào phương sai dao động chung của mực nước, dòng thứ ba ghi tỉ trọng tích lũy. Các cột giá trị của từng thành phần chính gồm các hệ số liên hệ của thành phần chính đang xét với các trạm mực nước. Các thành phần chính trong bảng 1 đã được sắp xếp theo thứ tự giảm dần của tỉ phần đóng góp vào dao động chung. Dưới đây sẽ phân tích bảng này để thấy ý nghĩa của phương pháp phân tích thành phần chính trong trường hợp cụ thể này.

1) Như đã nói, số thành phần chính nhận được bằng số trạm mực nước trong bảng số liệu xuất phát. Trong trường hợp cụ thể này, ta có 15 thành phần chính. Mỗi thành phần chính thể hiện một quá trình trừu tượng phản ánh những

đặc điểm chung nhất trong dao động của mực nước ở tất cả các trạm cũng như những nét đặc thù rất riêng của một số nhóm trạm. Nếu xét về phương diện đóng góp vào biến thiên chung của mực nước các trạm dọc bờ Việt Nam thì chỉ một số những thành phần chính – những quá trình đầu tiên được xem là quan trọng. Từ bảng 1, thấy rằng chỉ khoảng 5 quá trình đầu tiên đã góp hơn 95 % phương sai dao động mực nước các trạm dọc bờ Việt Nam. Khi áp dụng phương pháp vào khôi phục hay dự báo mực nước, ta hoàn toàn có thể chỉ cần tính tới những thành phần chính quan trọng này, bỏ qua các quá trình thứ yếu. Hình 1 là ví dụ ứng dụng dự báo. Hình này so sánh mực nước thực và dự báo tháng 1/2002 cho trạm Cửa Cấm theo nhóm trạm có cùng tính chất thủy triều (5 trạm ven bờ tây vịnh Bắc Bộ, từ Hòn Dấu tới Cửa Hội) bằng phương pháp hồi quy tuyến tính. Sai số quân phương dự báo mực nước giờ ở Cửa Cấm từ 2002 đến 2006 bằng 20,2 và hệ số tương quan giữa quan trắc và dự báo 0,96; với trạm Cửa Hội: hai tham số đánh giá tuần tự bằng 11,2 cm và 0,98; Phú An: 14,7 cm và 0.98. 2) Hãy chú ý, trong mỗi cột của ma trận chuyển đổi bao giờ cũng có một số hệ số liên hệ có giá trị lớn trội hơn so với các hệ số còn lại. Dấu hiệu này dùng để giải nghĩa mỗi thành phần chính diễn tả một quá trình thực tế nào đó. Ví dụ, xét thành phần chính 1, đây là thành phần chính quan trọng nhất, đóng góp hơn 49 % phương sai dao động mực nước. Các hệ số liên hệ của nó với các trạm Cửa Cấm, Hòn Dấu, Ba Lạt, Cẩm Nhượng, Hòn Ngư và Cửa Hội trội hơn hẳn, lớn hơn 0,3. Đây chính là quá trình dao động mực nước mang tính chất nhật triều đặc trưng của vùng ven bờ tây vịnh Bắc Bộ. Nếu vẽ đồ thị biến thiên của thành phần chính 1 theo thời gian (hình 2), nó sẽ có hình dạng giống như đường cong dao động mực nước tại các trạm mực nước với tính chất thủy triều toàn nhật (hãy so sánh hình 2 với hình 1).

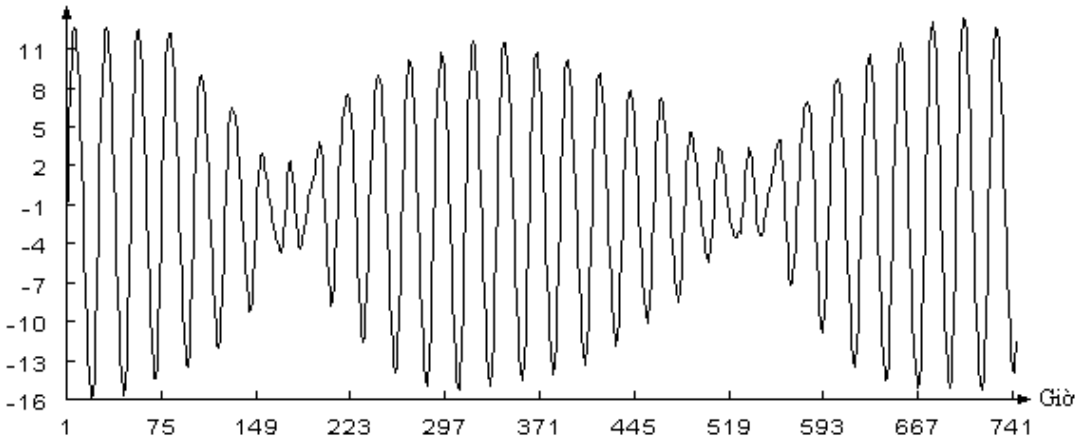
Quá trình quan trọng thứ hai liên quan tới thành phần chính 2 (hình 3) diễn tả dao động mực nước mang tính chất bán nhật triều không đều biên độ lớn điển hình của vùng ven bờ và trong sông đồng bằng sông Cửu Long. Hình dáng của đường cong dao động của thành phần này rất giống với đường cong mực nước thủy triều trạm Vũng Tàu.



Hình 1. Mực nước thực (liền nét) và dự báo (gạch nối) trạm Cửa Cấm.

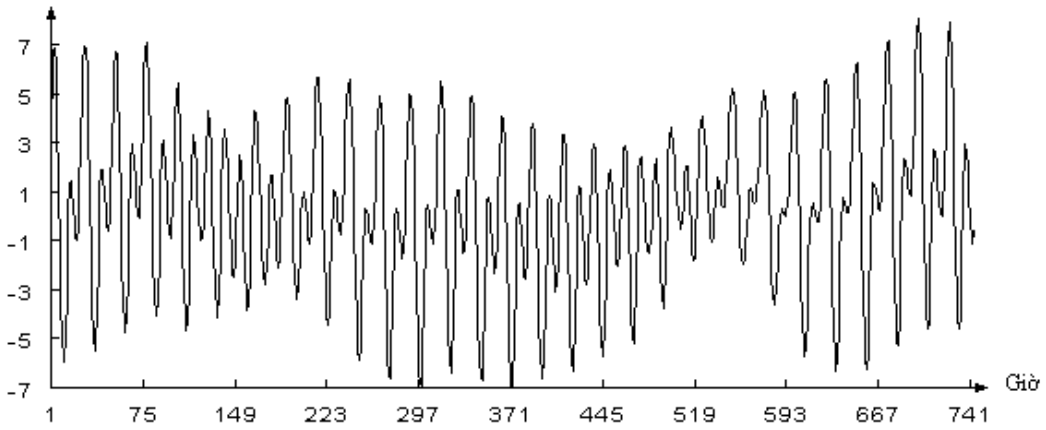
Bảng 1. Ma trận chuyển đổi

TPC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Tỉ phần (%)	49,36	24,22	15,74	3,78	2,02	1,75	1,13	0,70	0,41	0,28	0,17	0,15	012	0,11	0,06
Tích lũy (%)	49,36	73,57	89,31	93,09	95,12	96,87	97,99	98,69	99,11	99,39	99,56	99,71	99,83	99,94	100,0
Cửa Cấm	0,36	-0,01	-0,01	0,59	-0,16	0,34	0,24	0,08	-0,13	-0,03	-0,10	-0,10	-0,07	0,53	-0,04
Hòn Dấu	0,34	0,11	0,03	-0,09	-0,51	-0,45	-0,12	-0,39	0,31	0,05	-0,17	-0,03	0,25	0,20	-0,03
Ba Lạt	0,35	0,08	0,34	0,48	0,32	0,03	-0,33	-0,18	0,18	0,02	-0,08	-0,01	-0,02	-0,48	-0,01
Cẩm Nhượng	0,34	0,16	-0,36	0,04	-0,25	-0,05	0,45	0,15	-0,25	0,09	-0,07	0,13	0,08	-0,57	0,11
Hòn Ngư	0,34	0,13	-0,46	-0,11	0,60	-0,28	-0,01	-0,04	0,07	0,14	-0,05	0,20	-0,21	0,29	0,10
Cửa Hội	0,33	0,14	0,06	-0,48	-0,04	0,69	-0,02	-0,17	0,24	-0,14	-0,02	0,18	-0,03	-0,01	0,12
Cửa Việt	0,17	0,26	0,56	-0,17	0,04	-0,22	0,00	0,12	-0,50	-0,10	0,01	0,11	0,00	0,15	0,46
Cầu Lâu	-0,11	0,25	-0,06	-0,01	0,00	0,01	-0,02	0,44	0,40	0,01	-0,40	-0,47	-0,01	-0,01	0,43
Quy Nhơn	-0,30	0,12	0,15	0,13	0,18	0,00	0,63	-0,52	0,22	0,12	0,09	-0,03	0,03	0,00	0,29
Sơn Trà	-0,20	0,17	-0,43	0,22	-0,17	0,14	-0,44	-0,35	-0,25	-0,14	0,16	-0,04	-0,01	-0,01	0,48
Rạch Giá	0,21	0,31	-0,01	-0,05	0,00	-0,07	0,04	0,12	0,12	-0,06	0,78	-0,44	-0,01	0,00	-0,12
Vũng Tàu	-0,20	0,34	0,07	0,22	-0,23	0,00	-0,09	0,29	0,31	0,29	0,21	0,62	-0,16	0,06	0,04
Nhà Bè	-0,14	0,42	-0,08	0,10	0,21	-0,01	0,05	0,07	0,01	-0,54	-0,10	0,19	0,58	0,06	-0,25
Phú An	-0,12	0,41	0,02	-0,04	-0,13	-0,09	0,06	-0,18	-0,09	-0,31	-0,24	-0,09	-0,69	-0,07	-0,33
Tân An	-0,10	0,43	0,02	-0,10	0,07	0,20	-0,12	-0,12	-0,29	0,66	-0,17	-0,21	0,21	0,05	-0,28

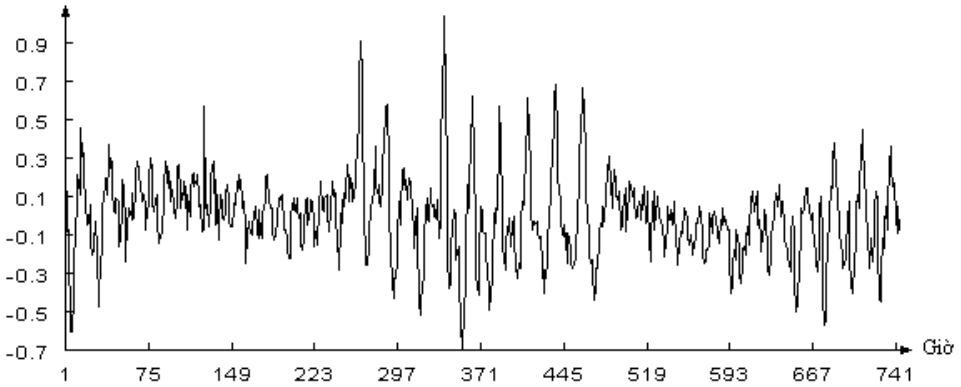


Hình 2. Thành phần chính 1 diễn tả quá trình dao động nhật triều thuần túy đặc trưng cho các trạm ven bờ tây vịnh Bắc Bộ

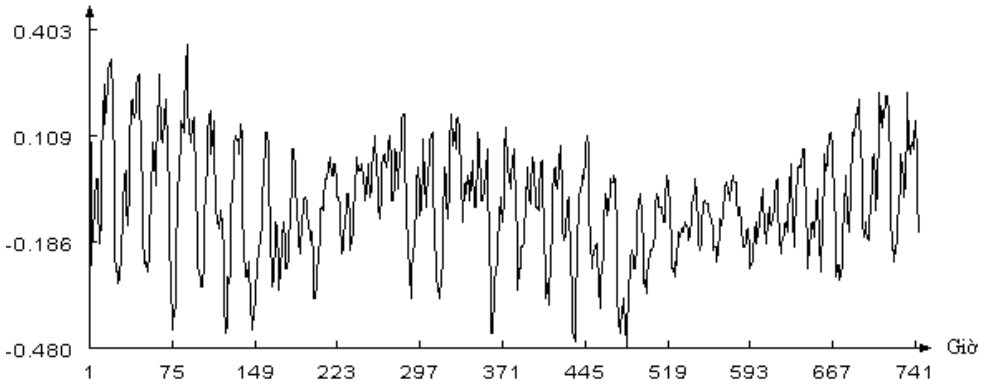
Các hình 4, 5 là đồ thị biến thiên thời gian của các thành phần chính 3 và 4. Các thành phần chính này có biên độ biến thiên nhỏ hơn rất nhiều so với các thành phần chính 1 và 2 đã xét ở trên, chúng diễn tả những quá trình dao động mực nước có ảnh hưởng của các sóng nước nông, phản ánh những nét riêng biệt quy mô nhỏ của các nhóm trạm mực nước do vị trí của trạm ở vùng bờ quyết định. Tương tự, ta có thể giải nghĩa tất cả các thành phần chính khác trong bảng 1 và gán cho từng thành phần một quá trình dao động phản ánh đặc thù riêng của một nhóm trạm.



Hình 3. Thành phần chính 2 diễn tả quá trình dao động bán nhật triều không đều đặc trưng cho các trạm ven bờ và trong sông đồng bằng sông Cửu Long



Hình 4. Thành phần chính 3 diễn tả quá trình dao động có yếu tố nước nông ảnh hưởng đặc trưng cho các trạm ven bờ cửa sông thuộc vịnh Bắc Bộ (Ba Lạt, Cẩm Nhượng, Hòn Ngư, Cửa Việt, Sơn Trà)



Hình 5. Thành phần chính 4 diễn tả dao động có yếu tố nước nông ảnh hưởng đặc trưng cho các trạm ven bờ cửa sông thuộc vịnh Bắc Bộ (Cửa Cấm, Ba Lạt, Cửa Hội)

4. Kết luận

Phân tích thành phần chính áp dụng vào tập số liệu mực nước cho phép nhận ra một cách khách quan và định lượng được những quá trình chính đóng góp vào dao động chung của mực nước trên toàn dải ven bờ nước ta cũng như những chi tiết khác biệt. Hai quá trình chính quyết định đặc điểm dao động mực nước là dao động thủy triều toàn nhật thuần túy và bán nhật không đều tương ứng ở hai cận phía bắc và phía nam của dải bờ biển. Các chi tiết tinh tế nhưng thứ yếu hơn thể hiện sự ảnh hưởng của dao động sóng nước nông ở một số nhóm tram.

Thí nghiệm dự báo mực nước theo các nhóm trạm cho kết quả rất tốt về độ chính xác. Cách dự báo này nên được sử dụng trong thực tế khôi phục số liệu mực nước nhằm các mục đích tính toán ứng dụng và dự tính thủy triều cho các vùng bờ hiếm số liệu quan trắc.

Tài liệu tham khảo

[1] Smirnov N. P., Vainovsky P. A., Titov Iu. E. *Chẩn đoán và dự báo thống kê các quá trình hải dương học*. Nxb Đại học Quốc gia Hà Nội, 2005

[2] В. Н. Малинин, П. П. Чернышков, С. М. Гордеева, Канарский аппвеллинг: круномасштабная изменчивость и прогноз температуры воды. Санкт-Петербург, Гидрометеоиздат, 2002, 154 ст.

[3] Pham Van Cu, Philippe Charette, Dinh Thi Dieu, Pham Ngoc Hai, Le Quang Toan, Application of the principal component analysis to explore the ralation between land use and solid waste generation in the Duy Tien district, Ha Nam province, Vietnam. *VNU Journal of Science, Earth sciences* 25 (2009) 65.

Principal component analyis applied to the sea level data along the shoreline of Vietnam

Pham Van Huan

VNU University of Science, 334 Nguyen Trai, Hanoi, Vietnam

This paper presents the procedures of application of the method of principal component analysis to the set of sea level data and shows the examples on the content interpretation of the transition matrix for the purpose of classification of the oscillation processes of sea level along shoreline of Vietnam. From the analysis of the transition matrix the groups of tide gauges of homogeneous level oscillation have been determined qualitatively and this served as a basis for the effective recovering or predicting sea levels for stations located in different fragments of the shoreline. The main processes affecting the differentiation in the time oscillation of sea level include the durnal and semi-durnal tidal properties, the influence of the shallow water tides and the river run-off in the vicinity of a considered station. The examples of recovering sea level for diferent groups of tide gauges of Vietnam shoreline have shown a good accuracy and promised a perspective application.