Phát hiện trị dị thường trong chuỗi trị đo vị trí điểm GNSS liên tục

- **Trần Đình Trọng** [1]
- **Đào Duy Toàn** [1]
- **Vũ Sơn Tùng** [2]
- **Lương Ngọc Dũng** [1]
- **Vũ Đình Chiều** [1]
- **Bùi Ngọc Sơn** [1]
- **Hà Thị Hằng** [1]

[1] Trường Đại học Xây dựng, Hà Nội, Việt Nam

[2] Xí nghiệp Tài nguyên Môi trường - Môi trường biển, Công ty Tài nguyên - Môi trường

(Bản thảo nhận ngày 28 tháng 06 năm 2016, hoàn chỉnh sửa chữa ngày 10 tháng 08 năm 2016)

TÓM TẮT

Số liệu tọa độ của các điểm lưới GNSS được sử dụng để xác định các tham số của hoạt động kiến tạo như vận tốc chuyển dịch, chuyển dịch theo mùa,... Việc đầu tiên của xử lý chuỗi tọa độ điểm lưới GNSS là loại bỏ các trị dị thường trong chuỗi. Thông thường, dị thường được phát hiện dựa trên kinh nghiệm trực quan của người xử lý. Do vậy, mất nhiều thời gian và có thể nhầm lẫn, nhất là đối với các trạm đo có dữ liệu lâu năm. Bài báo giới thiệu một số thuật toán, phương pháp loại bỏ trị dị thường và đánh giá ưu - nhược điểm của các phương pháp khi xử lý chuỗi trị đo vị trí GNSS liên tục.

Từ khóa: *mạng lưới GNSS liên tục, dị thường, chuỗi vị trí điểm GNSS, lọc trị dị thường*

1. GIỚI THIỆU

Hệ thống định vị vệ tinh toàn cầu (GNSS) hiện được sử dụng phổ biến để xác định sự chuyển dịch của bề mặt đất, độ chính xác đạt được cỡ ±1mm [1]. Xác định dịch chuyển theo thời gian của bề mặt đất thông qua phân tích chuỗi tọa độ các trạm GNSS liên tục là cách tiếp cận hiệu quả. Các tham số vật lý của hoạt động kiến tạo được xác định dựa trên chuỗi vị trí các điểm GNSS liên tục trong thời gian dài.

Các mạng lưới GNSS liên tục được xây dựng rộng rãi trên thế giới, đặc biệt trên những khu vực có nhiều hoạt động địa kiến tạo như Nhật (mạng lưới GEONET), Đài Loan (lưới quốc gia Đài Loan), Châu Âu (mạng lưới RGB), Mỹ (PBO),… Các mạng lưới này có đặc điểm là ngày càng nhiều dữ liệu lưu trữ và phạm vi lưới càng được chêm dày, mở rộng.

Thực tế, khi xử lý chuỗi vị trí điểm GNSS, chúng tôi nhận thấy, tất cả các chuỗi đều tồn tại các dị thường (outlier), chúng có giá trị lớn, không theo quy luật. Chúng có thể xuất hiện theo dạng đơn lẻ, cũng có thể xuất hiện theo

dạng nhóm,… Các dị thường này do nhiều nguyên nhân, trong đó phổ biến là do tín hiệu xấu, thiếu dữ liệu; do phần mềm bị lỗi tính toán, xử lý; do thời tiết (chủ yếu do băng tuyết đóng trên an-ten);… Chúng ảnh hưởng rất lớn đến kết quả tính toán, vận tốc và các tham số khác từ chuỗi. Việc phát hiện và loại bỏ chúng là điều cần thiết đầu tiên khi xử lý chuỗi.

Thông thường, dị thường được phát hiện dựa trên kinh nghiệm trực quan của người xử lý. Do vậy, mất nhiều thời gian và có thể nhầm lẫn dẫn đến mất dữ liệu hoặc không loại bỏ được dị thường, đặc biệt là đối với các trạm đo GNSS có dữ liệu lâu năm.

Ở Việt Nam, công tác xây dựng các mạng lưới địa động đã và đang được triển khai ở các tỉnh miền núi phía Bắc, miền Trung - Tây Nguyên và miền Nam. Mạng lưới này được xây dựng bao gồm các điểm trạm CORS có thu dữ liệu liên tục trong nhiều năm, giống như các mạng lưới địa động khác trên thế giới. Do đó, bài báo có triển vọng trong việc giúp ích cho công tác xử lý số liệu địa động tại Việt Nam.

2. CƠ SỞ LÝ THUYẾT

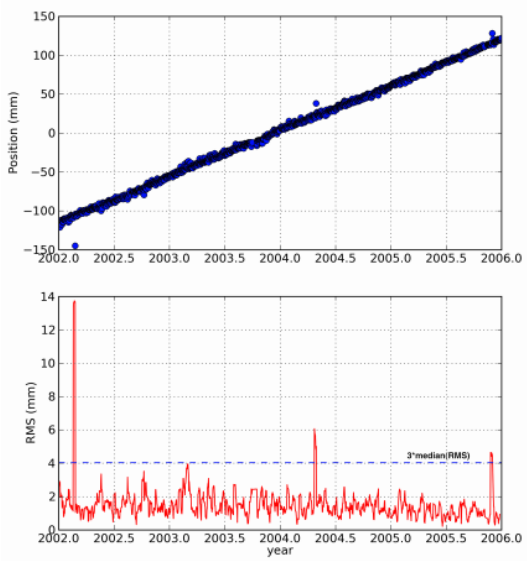
Thông thường, để phát hiện các giá trị dị thường trong một chuỗi dữ liệu nào đó, phương pháp kinh điển thường được sử dụng là các phương pháp kiểm tra thống kê, có thể kể tới như F-test, T-test,... Các phương pháp này dựa trên quy luật tồn tại của sai số thô trong dãy kết quả đo và sử dụng các luật phân phối để phát hiện chúng. Do đó, các phương pháp thống kê có nhược điểm là chỉ phát hiện được các dị thường nếu chúng chiếm số lượng ít trên tổng số trị đo trong chuỗi.

Đối với chuỗi tọa độ GNSS liên tục, chúng tôi nhận thấy, số lượng dị thường chiếm tỷ lệ lớn, đặc biệt ở một số trạm đo có điều kiện tự nhiên khắc nghiệt như vùng có băng tuyết (các điểm trên dãy Alps - lưới RENAG), vùng có lớp thực phủ biến đổi lớn theo mùa trong năm (lưới GEONET - Nhật Bản). Đối với các mạng lưới này, phương pháp thống kê toán học tỏ ra không hiệu quả; do đó, cần sử dụng các phương pháp khác hiệu quả hơn. Dưới đây, chúng tôi giới thiệu một số phương pháp và ứng dụng thực nghiệm để kiểm tra khả năng phát hiện trị dị thường trong chuỗi tọa độ GNSS, cụ thể với một vài điểm trong mạng lưới RENAG - Pháp.

2.1 Phương pháp hình học

2.1.1. Phương pháp thứ nhất

Phương pháp này dựa trên sự biến đổi của các độ lệch chuẩn của chuỗi nhỏ (cửa sổ) trong chuỗi lớn vị trí. Hiểu đơn giản là, nếu trong chuỗi nhỏ xuất hiện dị thường, thì độ lệch chuẩn của nó cũng là dị thường (hình 1). Từ đó xác định và loại bỏ tọa độ dị thường tương ứng với dị thường độ lệch chuẩn.



Hình 1. Mối liên hệ giữa trị dị thường và độ lệch chuẩn dị thường

Giả sử, có chuỗi tọa độ liên tục {x1, x2, …, xn}. Đầu tiên, chia dãy tọa độ vị trí điểm thành nhiều cửa sổ của s vị trí kề nhau. Cụ thể, cửa sổ thứ nhất bắt đầu từ tọa độ thứ nhất tới tọa độ thứ s, cửa sổ thứ hai bắt đầu từ tọa độ thứ hai tới tọa độ thứ s+1,... kiểu này được gọi là cửa sổ trượt.

Giá trị RMS (root mean square - độ lệch chuẩn) của cửa sổ thứ i được xác định như sau:

$$rms_i = \left(\sqrt{\frac{\sum_{j=1}^{n} (x_j - \overline{x_i})^2}{s-1}} \right)_i \quad (1)$$

Trong đó: x_j là trị đo thứ j; i là số thứ tự của cửa sổ; $\overline{x_i}$ là giá trị trung bình ở cửa sổ thứ i; s là kích thước cửa sổ.

Kích thước cửa sổ s có thể được lựa chọn tùy ý, điều này phụ thuộc vào kinh nghiệm của người sử dụng, thường chọn s = 7 [3]. Nếu một cửa sổ có dị thường, giá trị rms_i của cửa sổ đó cũng là một dị thường và việc phát hiện dị thường là hoàn toàn có thể thực hiện được (hình 1). Giá trị rms được coi là dị thường với xác xuất là 96% nếu nó có giá trị lớn hơn 3 lần số trung vị (median) của chuỗi rms đó:

$$rms_i > 3.\text{median}(rms) \quad (2)$$

Nhận xét: Việc lập trình tính toán tự động là hoàn toàn khả thi. Tuy nhiên, khi có nhiều trị dị thường tồn tại kề nhau (nhiều hơn hai) thì việc xử lý trở nên phức tạp hơn.

2.1.2. Phương pháp thứ hai

Phương pháp dựa trên sự chênh lệch giữa trị dị thường so với giá trị trung vị trong chuỗi nhỏ (cửa sổ). Nếu cửa sổ tồn tại dị thường thì chênh lệch của dị thường với trị trung vị của cửa sổ là rất lớn. Từ đó xác định và loại bỏ dị thường.

Tương tự như phương pháp thứ nhất, chia chuỗi dữ liệu tọa độ thành nhiều cửa sổ của s vị trí liền nhau (thường chọn $s \geq 5$), có thể tăng giảm sao cho hợp lý nhất. Tiến hành tìm kiếm trong từng cửa sổ để xác định các dị thường. Số lượng dị thường nhiều nhất trong một cửa sổ không vượt quá một giới hạn xác định [4] và được xác định theo công thức:

$$\text{outlier}_{max} \leq (s-1)/2 \quad (3)$$

(*a*) (*b*)

Hình 2. a) tọa độ của chuỗi ở cửa sổ thứ i; b) sự thay đổi giữa hai giá trị p_j và p_{j+1} trong cửa sổ i

(a) (b)

Hình 3. a) tọa độ của chuỗi ở cửa sổ thứ i, và b) sự thay đổi giữa hai giá trị p_j và p_{j+1} trong cửa sổ i

Trong đó: s kích cỡ của cửa sổ trượt, nếu s = 5 thì số outliermax không vượt quá 2, nếu s = 7 thì số outliermax không vượt quá 3. Để xác định và loại bỏ các dị thường, thường sử dụng giá trị giới hạn là 3σ; cách tính toán được thực hiện theo nguyên lý sau [4]:

$$x_j = \begin{cases} x_j ; \mid \rho(s) - x_j \mid \leq 3\sigma \\ - ; \mid \rho(s) - x_j \mid \geq 3\sigma \end{cases} \quad (4)$$

Trong đó: xj là trị đo thứ j; s là kích thước của cửa sổ; ρ(s) là trung vị của cửa sổ; σ là độ lệch chuẩn; (−) thể hiện giá trị dị thường đã bị loại bỏ.

Nhận xét: việc xác định trị dị thường và loại bỏ có thể dễ dàng được thực hiện theo hướng tự động hóa.

2.2 Phương pháp hồi quy

Phương pháp dựa trên cơ sở sử dụng hàm hồi quy mô tả chuỗi dữ liệu và xác định các dị thường không tuân theo quy luật của hàm mô tả, từ đó loại bỏ chúng khỏi chuỗi dữ liệu.

Đối với chuỗi số liệu tọa độ liên tục {x1, x2,... , xn}. Giả sử rằng các xi là giá trị rời rạc của hàm hồi quy đơn nguyên có dạng [2]:

$$Y_t = F(x_t, \theta) + \varepsilon_t \quad (5)$$

Trong đó: Y_t là đối tượng dự đoán; εt là sai số dự đoán; $f(x_t, \theta)$ là hàm biểu diễn sự biến đổi của Y_t theo biến x_t ở thời điểm t với các tham số mô hình θ.

Khi lượng số liệu nhiều hơn số tham số (k) của hàm hồi quy, ta sẽ giải phương hệ phương trình trên theo nguyên lý số bình phương nhỏ nhất và tìm được các tham số gần đúng của hàm F. Từ đó, xác định được độ lệch phân bố của các vị trí rời rạc x_i và giá trị xác suất của hàm hồi quy y_i tại thời điểm i là v_i. Giá trị này có thể được xác định như sau:

$$v_i = |y_i - x_i|; \; i = (1, 2, 3, \ldots, n) \quad (6)$$

Nếu giá trị $v_i > t.m_0$ thì giá trị x_i ở thời điểm đó là một dị thường. Trong đó:

$$m_0 = \pm\sqrt{\frac{[vv]}{n-k}}$$

t là giá trị xác suất, thường lấy t = 2 ÷ 3.

Nhận xét: Việc ứng dụng hàm hồi quy để loại bỏ dị thường phục thuộc vào lựa chọn dạng hàm hồi quy. Các mảng địa kiến tạo thường có chuyển động tịnh tiến, nên sử dụng hàm tuyến tính là phù hợp.

2.3 Phương pháp lọc Kalman

Phương pháp dựa trên sự chêch lệch giữa trị ước lượng và trị thực. Nếu giá trị chênh lệch vượt quá một giới hạn xác định thì trị thực được xem là dị thường và bị loại bỏ khỏi chuỗi số liệu tọa độ liên tục. Dưới đây, tác giả trình bày gắn gọn thuật toán lọc Kalman dạng đa thức.

Mô hình toán học của lọc Kalman gồm phương trình trạng thái và phương trình trị đo [1]:

$$x_k = \Phi_k x_{k-1} + G_k u_{k-1} + w_{k-1} \quad (7)$$

$$z_k = H_k x_k + v_k \quad (8)$$

Trong đó: x_k là vector trạng thái thực của hệ thống tại thời điểm tk; Φ_k là ma trận thay đổi trạng thái trong thời gian từ t_{k-1} tới t_k; G_k là ma trận hệ số điều khiển ngoại lực; u_k là vector kiểm soát; w_k là nhiễu động thái ở thời điểm t_{k-1}; z_k là vector trị đo của hệ thống ở thời điểm t_k; H_k là ma trận trị đo ở thời điểm t_k; v_k là nhiễu của trị đo ở thời điểm t_k;

Nếu w_k và v_k thỏa mãn đặc tính thống kê:

$E(w_k) = 0, \; E(v_k) = 0$

$Cov(w_k, v_{kj}) = Q_k \delta_{kj}$, $Cov(v_k, v_j) = R_k \delta_{kj}$, $Cov(w_k, v_j) = 0$

Trong đó: Q_k là ma trận phương sai nhiễu động thái; R_k là ma trận phương sai nhiễu trị đo; δ_{kj} là hàm số Kronecker, thì có thể dẫn tới công thức suy rộng dần lọc Kalman:

$$\delta kj = \begin{cases} 1, k = j \\ 0, k \neq j \end{cases}$$

Dự báo trạng thái:

$$\hat{x}_{k/k-1} = \Phi_k \hat{x}_{k-1} + G_k u_k + w_k \quad (9)$$

Dự báo ma trận hiệp phương sai trạng thái:

$$P_{k/k-1} = \Phi_k P_{k-1} \Phi_k^T + Q_k \quad (10)$$

Ước lượng trạng thái:

$$\hat{x}_k = \hat{x}_{k/k-1} + K_k (z_k - H_k \hat{x}_{k/k-1}) \quad (11)$$

Ước lượng ma trận hiệp phương sai trạng thái:

$$P_k = (I - K_k H_k) P_{k/k-1} \quad (12)$$

Trong đó, K_k là ma trận hiệu ích lọc:

$$K_k = P_{k/k-1} H_k^T (H_k P_{k/k-1} H_k^T + R_k)^{-1} \quad (13)$$

Điều kiện trạng thái ban đầu là:

$$\hat{x}_0 = E(x_0) = \mu_0, \hat{P}_0 = Var(x_0) \quad (14)$$

Từ (11) ta thấy, sau khi tiến hành đo hệ thống z_k ở thời điểm tk, thì có thể dùng trị đo này tiến hành tu chỉnh trị dự báo để ước lượng trạng thái (trị lọc) $\hat{x}_k$ của hệ thống ở thời điểm t_k, lặp lại như vậy, suy rộng dần dự báo và lọc.

Điều quan trọng là bậc lọc hoặc phương trình cơ bản (Φ_k) đưa vào phải phù hợp với sự thay đổi thực tế của đối tượng. Trong bài báo này, mục đích của tác giả là sử dụng lọc Kalman để tìm trị dị thường, cho nên chỉ sử dụng phương trình bậc đa thức bậc 0.

Nếu đa thức trị đo là một hằng số $x = a_0$, đạo hàm của nó $\dot{x} = 0$, không gian trạng thái mà nó mô tả bằng phương trình vi phân như sau:

$$\dot{x} = \Phi x \quad (15)$$

Từ (11) ta có phương trình Kalman dạng đa thức bậc 0:

$$\hat{x}_k = \hat{x}_{k/k-1} + K_k (z_k - \hat{x}_{k/k-1}) \quad (16)$$

Tại phương trình ước lượng trạng thái (11), trước khi cập nhật trị đo, nên tính toán chênh lệch giữa trị ước lượng và trị thực Δx.

Nếu $\Delta x = |z_k - \hat{x}_{k/k-1}| > t.\sigma$ thì z_k là dị thường, loại bỏ z_k khỏi chuỗi và coi $\hat{x}_k = \hat{x}_{k/k-1}$, tiếp tục quá trình lọc với trị đo tiếp theo. Trong đó, $\sigma^2 = R_k$; thường lấy $t = 2 \div 3$.

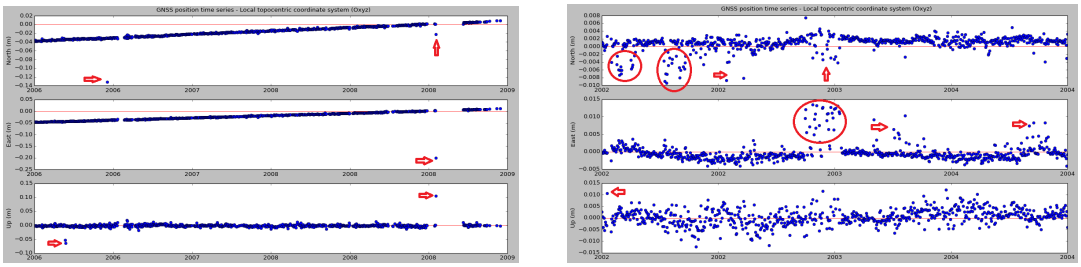
3. THỰC NGHIỆM

3.1 Giới thiệu

Thực nghiệm cho hai trạm CORS trong mạng lưới RENAG – Pháp, trong đó, điểm BANN có dữ liệu từ đầu năm 2006 đến hết năm 2008, và điểm MEDI có dữ liệu từ đầu 2002 tới hết năm 2003. Tọa độ các điểm nàu đã được tính trong hệ ITRF và được tính chuyển về hệ tọa độ địa diện chân trời (North, East, Up).

Ghi chú ký hiệu trong các hình của kết quả xử lý số liệu: Chấm màu đỏ là dị thường, chấm màu xanh là tọa độ tin cậy trong chuỗi.

Chuỗi tọa độ điểm được biểu diễn ở hình 4a, b theo 3 thành phần n, e, u. Đối với điểm BANN, số lượng dị thường không nhiều và đơn giản (hình 4a). Nguyên nhân là do lỗi thiếu dữ liệu của một thời điểm, dẫn tới việc tính toán sai tọa độ.



a) b)

Hình 4. Dị thường trong chuỗi tọa độ điểm a) BANN và b) MEDI

Trong chuỗi dữ liệu của điểm MEDI, có nhiều giá trị nghi ngờ là dị thường (hình 6b) dưới dạng nhóm và trải rộng trong chuỗi (vùng khoanh tròn đỏ). Nguyên nhân là do gần điểm MEDI có cây cao, làm cho tín hiệu bị ảnh hưởng xấu ở một vài thời điểm trong năm (khi lùm cây cao, không chặt bỏ).

3.2 Kết quả tính toán

3.2.1. Phương pháp hình học thứ nhất và thứ hai

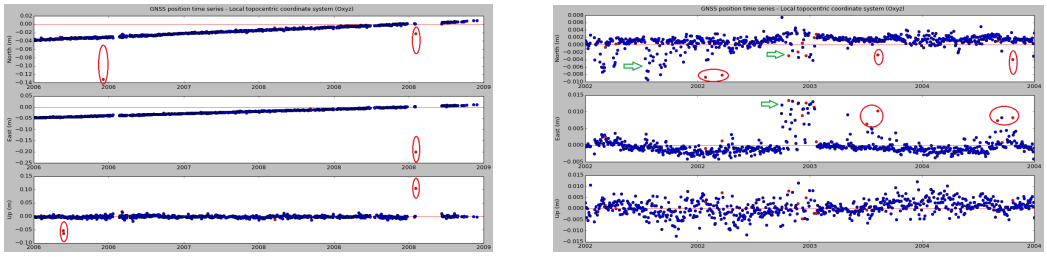
Các dị thường trong chuỗi dữ liệu của điểm BANN đều được phát hiện đúng và tự động loại bỏ (chấm màu đỏ) trong cả 2 phương pháp, hình 5a, 6a. Còn trong chuỗi của điểm MEDI, phương pháp thứ nhất chỉ được loại bỏ một số ít dị thường (chấm màu đỏ), hình 5b; phương pháp thứ hai đã loại bỏ gần hết sau ba lần tính lặp (chấm màu đỏ); tuy nhiên, cũng có nhiều giá trị tin cậy trong chuỗi cũng bị loại bỏ (hình 6b).

3.2.2. Phương pháp hồi quy tuyến tính

Các dị thường trong chuỗi dữ liệu của điểm BANN đều được phát hiện đúng và tự động loại bỏ, dữ liệu nằm trong vùng giữa hai đường biên màu xanh đều là giá trị đáng tin cậy, hình 7a. Đa số dị thường trong chuỗi dữ liệu của điểm MEDI đã được loại bỏ (vùng chấm đỏ được khoanh tròn), hình 7b; tuy nhiên, vẫn có một số giá trị nghi là dị thường còn tồn tại và phân bố tản mát trong chuỗi.

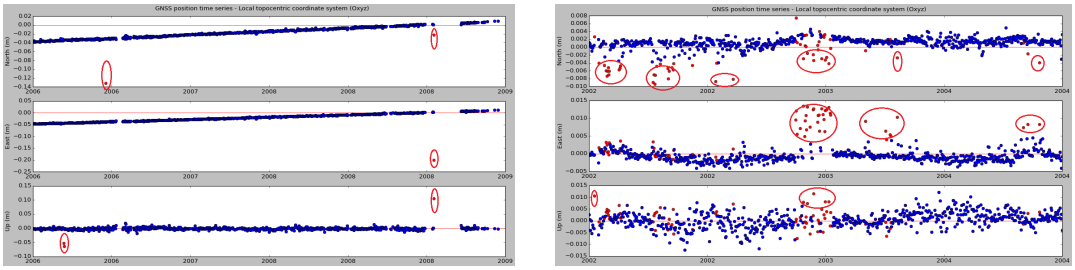
3.2.4. Phương pháp lọc Kalman bậc 0

Các dị thường trong chuỗi dữ liệu của điểm BANN đều được phát hiện đúng và tự động loại bỏ, hình 6a. Đường màu đỏ (hình 8a) là đường ước lượng được thực hiện bởi lọc Kalman đa thức bậc 0. Đa số dị thường trong chuỗi dữ liệu của điểm MEDI đã được phát hiện và loại bỏ (chấm màu đỏ khoanh tròn), hình 6b. Tuy nhiên, cũng thấy rằng có một số giá trị tin cậy cũng đã bị loại khỏi chuỗi sau khi lọc (hình 8b).



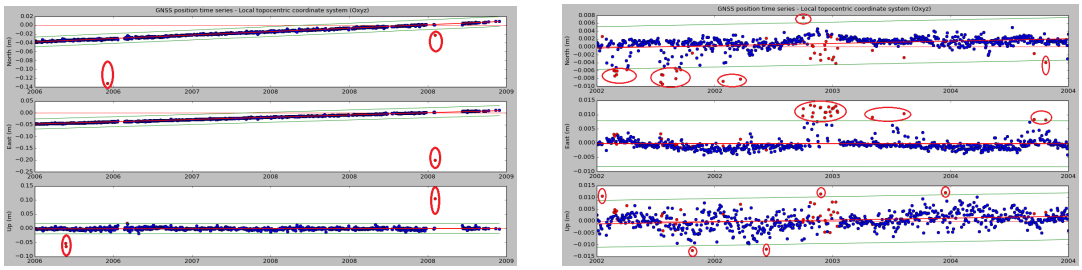
a) b)

Hình 5. Loại bỏ dị thường bằng phương pháp hình học thứ nhất của điểm a) BANN và b) MEDI



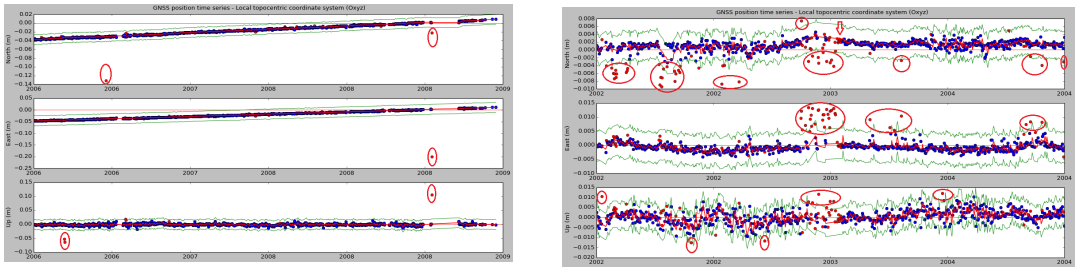
a) b)

Hình 6. a) Loại bỏ dị thường bằng phương pháp hình học thứ hai của điểm BANN, và b) MEDI



a) b)

Hình 7. Loại bỏ dị thường bằng phương pháp hồi quy của điểm a) BANN và b) MEDI



a) b)

Hình 8. Loại bỏ dị thường bằng lọc Kalman đa thức bậc 0 của điểm a) BANN và b) MEDI

Nhận xét chung: Với dãy số liệu của điểm BANN, chỉ có một vài dị thường, các phương pháp đều tự động tìm và tách dị thường khỏi chuỗi (các hình a). Với dãy số liệu của điểm MEDI, các phương pháp đã không phát hiện và loại bỏ hoàn toàn được các dị thường (các hình b). Tuy nhiên, nếu sử dụng phép lọc lặp nhiều lần thì kết quả thu được khả quan hơn (hình 8b).

4. KẾT LUẬN

Kết quả thực nghiệm cho thấy các phương pháp hình học nên được áp dụng để tìm và loại các dị thường độc lập hoặc nhóm dị thường đơn giản. Phương pháp hồi quy có thể loại bỏ hầu hết các dị thường xuất hiện trong chuỗi. Phương pháp Kalman cũng cho thấy hiệu quả nhất định. Song, lựa chọn bậc lọc và trị khởi đầu cần cẩn thận, khi thấy lọc không đúng thì cần điều chỉnh cho phù hợp hơn.

Đối với tập dữ liệu khảo sát trong bài báo, phương pháp lọc Kalman là hiệu quả nhất, sau đó tới phương pháp hồi quy, cuối cùng là các phương pháp hình học.

Rõ ràng là không có phương pháp tối ưu, tùy từng trường hợp cụ thể mà áp dụng phương pháp phù hợp. Với các chuỗi dữ liệu có dị thường phức tạp (như của điểm MEDI), không thể áp dụng một phương pháp loại bỏ tất cả các dị thường. Sau khi lọc tự động, nên áp dụng cách thức kiểm tra thủ công dựa trên kinh nghiệm của các chuyên gia.

Outlier detection in GNSS position time series

- **Tran Dinh Trong** [1]
- **Dao Duy Toan** [1]
- **Vu So Tung** [2]
- **Luong Ngoc Dung** [1]
- **Vu Dinh Chieu** [1]
- **Bui Ngoc Son** [1]
- **Ha Thi Hang** [1]

[1] National University of Civil Engineering, Hanoi, Vietnam

[2] Enterprise of Natural Resource and Environment - Marine Environment, Hanoi, Vietnam

ABSTRACT

The continuous GNSS stations are used to determine the displacement velocities, seasonal variation, amplitude of tectonic activities,… To accurately determine these factors, the first is to remove outliers in GNSS position time series. In general, filtering approaches are subjectively selected based on the experience and visual interpretation of experts. Therefore, the process may lead to a waste of time or confusion, especially for stations with long-term continuously recorded data. The purpose of paper is to introduce the applicability of several algorithms and methods of filtering outliers in GNSS position time series.

***Keywords**: Permanent GNSS networks, outliers, GNSS position time series, filtering outliers.*

TÀI LIỆU THAM KHẢO

[1]. Hill, EM., Davis JL., et, *Characterization of sitespecific GPS eror using a short - baseline network of braced monuments at Yucca Mountain, southern Nevada*, J geophys Res, 114 (B1), 1402 (2009).

[2]. Đào Duy Toàn, *Nghiên cứu phương pháp xử lý số liệu quan trắc chuyển dịch công trình khi đo bằng công nghệ hiện đại*, Luận văn Thạc sỹ, Thư viện Đại học Mỏ - Địa chất (2014).

[3]. Tran, D.T., *Analyse Rapide Et Robust Des Solutions GPS Pour La Technique*, PhD Thesis, National Library, (2013).

[4]. Stig Olsen, *On Automatic Data Processing and Well-Test Analysis in Read-time Reservoir Management Applications*, PhD Thesis, The University of Bergen, (2011).