

# NGHIÊN CỨU ỨNG DỤNG HÀM PHÂN PHỐI CHUẨN TRONG XỬ LÝ SỐ LIỆU QUAN TRẮC MÔI TRƯỜNG NUỚC BIỂN MIỀN TRUNG VIỆT NAM

**Nguyễn Ngọc Linh**

Trường Đại học Tài nguyên và Môi trường Hà Nội

## Tóm tắt

Ở Việt Nam, hệ thống quan trắc - cảnh báo môi trường, dịch bệnh thủy sản đã được thực hiện trong cả nước. Vì vậy, xử lý số liệu thực nghiệm thu được tại các điểm và khu vực biển giúp cho bộ số liệu quan trắc có độ chính xác cao và độ tin cậy mong muốn là rất cần thiết, và là biện pháp phát hiện, cảnh báo các chất ô nhiễm có hiệu quả nhất, kịp thời kiểm soát và ngăn ngừa ô nhiễm môi trường biển, từ đó có các biện pháp giảm thiểu các tác hại gây ra đối với hoạt động nuôi trồng biển. Bằng phương pháp thực nghiệm để thu thập số liệu quan trắc môi trường nước biển từ trạm quan trắc và phân tích môi trường biển miền Trung, kết hợp tính toán lý thuyết hàm phân phối chuẩn. Kết quả nghiên cứu của bài báo sẽ cung cấp thêm một phương pháp đơn giản, dễ sử dụng trong việc biến đổi, tính toán số liệu quan trắc môi trường nước biển, qua đó người sử dụng có thể tính toán, xử lý số liệu quan trắc bằng công cụ đơn giản là máy tính bỏ túi.

**Từ khóa:** Hàm phân phối chuẩn; Xử lý số liệu; Nước biển; Miền Trung Việt Nam.

## Abstract

***Application of normal distribution function in processing monitoring data of sea water in Central Vietnam***

In Vietnam, the system of monitoring and warning for environment and aquatic diseases has been implemented throughout the country. The processing of empirical data collected at marine areas to improve the accuracy and reliability of monitoring data sets is very necessary and is the most effective means of detecting pollutants to control and prevent marine pollution promptly and minimize the bad impacts to mariculture activities. By combining the empirical methods of on-site collecting seawater environmental monitoring data from the marine environment monitoring and analyzing station in the Central of Vietnam, with the theoretical calculations of the normal distribution function, this paper provides a simple and easy-to-use method of transforming and calculating seawater monitoring data for the users to calculate and process monitoring data using simply tools such as a pocket calculator.

**Keywords:** Normal distribution function; Processing data; Seawater; Central of Vietnam.

## 1. Đặt vấn đề

Biển và đại dương là nơi tiếp nhận phần lớn các chất thải từ lục địa. Công ước quốc tế về Luật biển năm 1982 đã chỉ ra 5 nguồn gây ô nhiễm biển chính là: Các hoạt động trên đất liền, thăm dò và khai

thác tài nguyên trên thềm lục địa và đáy đại dương; thả các chất độc hại ra biển; vận chuyển hàng hóa trên biển và ô nhiễm biển từ không khí.

Những năm qua, đồng hành với sự gia tăng diện tích và sản lượng nuôi trồng

thủy sản, những hoạt động kinh tế - xã hội trong cả nước không ngừng tăng trưởng, kéo theo hàng loạt các vấn đề về môi trường tác động đến khu vực ven biển, trong đó có hoạt động nuôi biển. Vì vậy, công việc quan trắc, phân tích các chất ô nhiễm trong nước biển và ở các khu nuôi trồng hải sản tập trung ven biển là rất cần thiết; và là biện pháp phát hiện, cảnh báo các chất ô nhiễm có hiệu quả nhất, kịp thời kiểm soát và ngăn ngừa ô nhiễm môi trường biển, từ đó có các biện pháp giảm thiểu các tác hại gây ra đối với hoạt động nuôi biển.

Quan trắc môi trường là quá trình theo dõi có hệ thống về thành phần môi trường, các yếu tố tác động lên môi trường nhằm cung cấp thông tin đánh giá hiện trạng, diễn biến chất lượng môi trường và các tác động xấu đối với môi trường (theo luật bảo vệ môi trường Việt Nam 2014). Số liệu quan trắc đo đạc ngoài hiện trường và phân tích trong phòng thí nghiệm được xử lý, tính toán nhằm biến các số liệu thô đã có thành cơ sở dữ liệu dễ truy cập và sử dụng khi cần thiết. Vì vậy việc ứng dụng các hàm thống kê nói chung và hàm phân phối chuẩn nói riêng trong xử lý số liệu thực nghiệm thu được sẽ giúp cho bộ số liệu quan trắc có tin cậy mong muốn.

### **2. Phương pháp nghiên cứu**

Dựa trên các kết quả nghiên cứu trước đây, thông qua việc tìm hiểu nhằm đánh giá các ưu nhược điểm của các phương pháp xử lý số liệu, tác giả đề xuất áp dụng hàm phân phối chuẩn để xử lý số liệu quan trắc môi trường nước biển miền Trung Việt Nam.

#### **2.1. Phương pháp thu thập số liệu**

- Điều tra thu thập các tài liệu, số liệu kết hợp với thu thập số liệu quan trắc từ biển bắn kết quả đo, phân tích tại hiện trường.

- Mẫu nước biển thu thập tại hiện trường gồm khoảng từ 35 đến 40 quan trắc, đo vào 2 đợt tháng 4 và tháng 10, tại 6 điểm đo (Đèo Ngang, Thuận An, Đà Nẵng, Dung Quất, Sa Huỳnh, Quy Nhơn)

- Các yếu tố được quan trắc vào thời điểm chân triều và đỉnh triều bao gồm:

+ Đo nhiệt độ nước, độ mặn nước biển ở 2 tầng mặt và đáy;

+ Mẫu phân tích chất dinh dưỡng (tổng Nitơ) ở tầng mặt và tầng đáy.

- Hàm lượng Nitơ được xác định bằng phương pháp trắc quang, so màu của máy DREL/2010 sau khi lấy mẫu, kèm theo hóa chất của hãng sản xuất [9].

- Số liệu quan trắc môi trường nước biển khu vực miền Trung các năm 2013, 2014, 2015 [9].

#### **2.2. Phương pháp xử lý số liệu**

Trên cơ sở số liệu thu thập được, chọn ra 3 chỉ tiêu để xử lý đó là: Nhiệt độ nước biển, tổng Nitơ trong nước biển và độ mặn nước biển.

Xây dựng công thức Toán học kết hợp các bài toán ước lượng, kiểm định trong thống kê đối với các đại lượng tuân theo luật phân phối chuẩn để phát hiện các giá trị bất thường, khử sai số quan trắc, xác định độ chính xác, số quan trắc cần thiết và so sánh hai dãy kết quả quan trắc đối với 3 chỉ tiêu: nhiệt độ nước biển, tổng Nitơ trong nước biển và độ mặn nước biển.

#### **2.3. Một số công thức lý thuyết.**

- Trung bình số học ( $\bar{x}$ ) là đại lượng dùng để chỉ giá trị đạt được khi chia tổng các kết quả quan trắc lặp lại cho số quan trắc lặp lại.

Giả sử có tập số liệu quan trắc lặp lại gồm  $n$  giá trị, ký hiệu:  $x_1, x_2, \dots, x_n$  thì giá trị trung bình số học của tập số liệu gồm  $n$  quan trắc lặp lại là:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n} \quad (1)$$

- Phương sai ( $\sigma^2$  và  $S^2$ ) là giá trị trung bình của tổng bình phương sự sai khác giữa các giá trị riêng rẽ trong tập số liệu so với giá trị trung bình.

+ Phương sai phản ánh mức độ phân tán của các giá trị quan sát chung quanh giá trị trung bình, do đó dùng để so sánh độ phân tán mẫu của các mẫu nghiên cứu.

+ Nếu tập số liệu lớn ( $N \rightarrow \infty$  thì phương sai mẫu tổng thể được ký hiệu là  $\sigma^2$  và được tính theo công thức

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N} \quad (2)$$

- Nếu tập số liệu nhỏ thì phương sai mẫu thống kê được ký hiệu là  $S^2$  và được tính theo công thức:

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (3)$$

### 3. Kết quả nghiên cứu

#### 3.1. Một số kết quả lý thuyết

Trong đo đạc thực nghiệm, dù có chính xác đến mức nào thì vẫn có sai lệch khác nhau. Ta gọi đại lượng đang xét là  $x$  và giá trị thực của đại lượng đó là  $\mu$ . Khi đó, giá trị  $|x - \mu|$  được gọi là sai số thực nghiệm. Sai số thực nghiệm thường không biết và giá trị thực cũng không biết một cách chính xác.

*Bài toán đặt ra là: Hãy ước lượng giá trị gần đúng của  $\mu$  với sai số bé nhất.* Để làm được điều đó cần phải xử lý số liệu thực nghiệm, cụ thể là phát hiện các giá trị bất thường và khử sai số thực nghiệm dựa vào phân phối chuẩn.

- Phát hiện các trị bất thường dựa vào phân bố chuẩn

Gọi  $x_i$  là các giá trị của mẫu dữ liệu. Theo lý thuyết phân bố chuẩn [7], 99% các giá trị  $x_i$  có phân bố chuẩn sẽ nằm trong khoảng  $(\bar{x} - 2,58.S; \bar{x} + 2,58.S)$  với  $\bar{x}$  là trung bình mẫu và  $S$  là độ lệch mẫu.

Các giá trị nằm ngoài khoảng  $(\bar{x} - 2,58.S; \bar{x} + 2,58.S)$  là các giá trị bất thường.

Nếu trong tập số liệu có rất nhiều số hoặc rất lớn hoặc rất nhỏ thì nên lấy giá trị trung vị thay cho trung bình và lược bỏ sơ bộ một số giá trị bất thường để tính độ lệch chuẩn của tập số liệu.

- Khử sai số quan trắc

Khi nhận thấy có một kết quả quan trắc  $x_0$  khác xa với kết quả quan trắc khác thì ta nghĩ ngay đến hiện tượng là đã mắc phải sai số thô. Khi đó, ta xem xét các quan trắc còn lại  $x_1, x_2, \dots, x_n$  như là tập hợp riêng còn  $x_0$  được xem như là quan trắc đột biến.

$$\text{Tính } \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Giả sử  $x$  là đại lượng ngẫu nhiên có luật phân phối chuẩn ( $\mu, \sigma^2$ ).

Với mức ý nghĩa  $\alpha$  ta tính:

$$u = \frac{|x_0 - \bar{x}|}{\sigma \sqrt{\frac{n+1}{n}}} \quad (4)$$

Tìm  $u_\alpha$  sao cho  $1 - 2\Phi(u_\alpha) = u_\alpha$  bằng cách tra bảng giá trị Laplace [8].

Thông thường tính ba cấp:  $u_{0,05} = 1,96$ ;  $u_{0,01} = 2,58$ ;  $u_{0,001} = 3,29$

Nếu  $u > u_\alpha$  thì bỏ giả thiết  $x_0 = \bar{x}$  và loại  $x_0$ ;

Nếu  $u < u_\alpha$  thì nhận lại  $x_0$  (nghĩa là  $x_0$  vẫn thuộc tập hợp dùng được);

Khi đó có nghĩa là ta đã loại  $x_0$  chứa sai số thô với mức ý nghĩa  $\alpha$  (hoặc độ tin

## Nghiên cứu

cậy  $1-\alpha$ ) hoặc là đã nhận lại  $x_0$  vì  $x_0$  không chừa sai số thô với mức ý nghĩa  $\alpha$  (hoặc độ tin cậy  $1-\alpha$ ).

Nếu  $\sigma$  chưa biết ( $n > 30$ ), công thức (3.1) được thay bởi:

$$u = \frac{|x_0 - \bar{x}|}{S \sqrt{\frac{n+1}{n}}} \quad (5)$$

- Độ chính xác của các quan trắc

Sau khi khử những quan trắc có sai số thô, ta còn lại các quan trắc chỉ có sai số ngẫu nhiên. Giả sử các sai số ngẫu nhiên tuân theo luật phân phối chuẩn. Vậy thì độ chính xác của các quan trắc như thế nào? Nghĩa là, độ chính xác của các dụng cụ đo đặc thế nào? Khi đó ta cần ước lượng phương sai  $\sigma^2$  của luật phân phối này (luật phân phối của các  $x_i$ )

Với một loạt gồm  $n$  quan trắc, ta có:

$$\sigma^2 \approx S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (6)$$

- Số lượng quan trắc cần thiết

Ta đã chủ trương không quan trắc tổng thể được toàn bộ môi trường nước biển mà chỉ quan trắc một mẫu ngẫu nhiên. Nhưng nếu kích thước mẫu n nhỏ thì quan trắc nhanh và tính toán đơn giản song tính tiêu biểu của mẫu đối với tổng thể kém. Ngược lại, nếu kích thước mẫu n lớn thì đặc trưng của mẫu đối với tổng thể tốt nhưng thực nghiệm lâu và tính toán phức tạp hơn. Chính vì vậy, trước mỗi lần thực nghiệm với yêu cầu cho trước (đảm bảo độ tin cậy  $\gamma$  cho trước, sai số cho phép  $\varepsilon$  không vượt quá  $\varepsilon_0$ ) ta phải xác định kích thước mẫu tối thiểu n bằng bao nhiêu là thích hợp nhất.

Giả sử muốn ước lượng giá trị trung bình  $\mu$  của tổng thể với sai số không quá  $\varepsilon$  cho trước với độ tin cậy  $\gamma$

thì kích thước mẫu n nhỏ nhất thỏa mãn bất đẳng thức:

$$n \geq \left( \frac{S_u \alpha}{\varepsilon} \right)^2 \quad (7)$$

Trong đó:

$$u_\alpha \text{ sao cho} \\ \Phi(u_\alpha) = \frac{\gamma}{2} \quad (\alpha = 1 - \gamma)$$

bằng cách tra bảng giá trị Laplace.

$S$  được xác định theo (1.4)

Thông thường tính ba cấp:  $u_{0,05} = 1,96$ ;  $u_{0,01} = 2,58$ ;  $u_{0,001} = 3,29$

- So sánh hai giá trị trung bình

Trong cuộc sống hàng ngày cũng như trong công tác quan trắc chúng ta luôn phải làm phép so sánh: So sánh chất lượng của hai loại sản phẩm, của các loại dịch vụ, so sánh hai cơ hội đầu tư,... Và trong khuôn khổ đề tài nghiên cứu này là so sánh hai dãy kết quả quan trắc môi trường nước biển.

Giả sử ta có hai mẫu ngẫu nhiên độc lập là  $\{x_1, x_2, \dots, x_n\}$  và  $\{y_1, y_2, \dots, y_m\}$

Ta muốn so sánh hai giá trị trung bình tổng thể  $\mu_1, \mu_2$  dựa vào hai mẫu độc lập trên. Ta kiểm định cặp giả thiết:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases}$$

Ta chọn thống kê:

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} \quad (8)$$

Với mức ý nghĩa  $\alpha$  nếu  $T > u_\alpha$  thì bác bỏ  $H_0$

Trong đó:  $u_\alpha$  sao cho  $\Phi(u_\alpha) = \frac{1-\alpha}{2}$  bằng cách tra bảng giá trị Laplace.

Thông thường tính ba cấp:  $u_{0,05} = 1,96$ ;  $u_{0,01} = 2,58$ ;  $u_{0,001} = 3,29$

### 3.2. Nhiệt độ nước biển khu vực miền Trung Việt Nam năm 2015

Bảng 1. Nhiệt độ nước biển khu vực miền Trung năm 2015

Đợt đo	Giá trị quan trắc (đơn vị: °C)					
	Đèo Ngang	Thuận An	Đà Nẵng	Dung Quất	Sa Huỳnh	Quy Nhơn
Tháng 4	22,3 23,2	27,3 24,5	26,5 22,9	24,8 22,6	23,2 24,9	24,2 26,8
	21,7 19,5	24,7 23,7	21,5 23,6	22,8 24,9	23,6 24,7	24,3 26,4
	23,8 23,4	26,7 23,5	22,4 25,5	22,8 25,8	23,4 26,2	25,2 28,4 25,6
Tháng 10	28,6 28,2	28,6 28,5	28,6 28,7	28,4 28,3	27,7 27,4	28,2 27,9
	28,8 28,4	28,2 28,3	29,1 28,6	28,1 27,8	27,6 27,4	27,4 28,9
	28,9 29,1	27,9 28,7	29,3 28,1	27,5 27,8	27,2 27,6	29,4 28,8 29,2

- Nhiệt độ quan trắc tháng 4/2015

Sử dụng công thức (1) và (3) ta tính được:  $x = 24,2$ ;  $S=1,77$

Theo lý thuyết phân bố chuẩn, 99% các giá trị  $x_i$  có phân bố chuẩn sẽ nằm trong khoảng  $(\bar{x} - 2,58.S; \bar{x} + 2,58.S) \approx (19,63; 28,77)$

Trong Bảng 1, bộ số liệu tháng 4 có giá trị quan trắc  $x_0 = 19,5 \notin (19,63; 28,77)$

Do đó, giá trị quan trắc  $x_0 = 19,5$  trong dãy số liệu thu được là giá trị bất thường.

Vậy giá trị bất thường  $x_0 = 19,5$  có mắc phải sai số thô hay không? Để trả lời câu hỏi đó ta xem xét các số liệu thu được còn lại (gồm 36 giá trị) là tập hợp riêng còn giá trị  $x_0 = 19,5$  được xem như là quan trắc đột biến.

Ta có  $x' = 29,4$ ;  $S \approx 1,65$  vì  $\sigma$  chưa biết và số giá trị quan trắc là  $36 > 30$  nên theo (3.2) ta có:

$$u = \frac{|x_0 - x'|}{S \sqrt{\frac{n+1}{n}}} = \frac{|19,5 - 29,4|}{1,65 \sqrt{\frac{36+1}{36}}} \approx 2,93$$

Với mức ý nghĩa  $\alpha = 1\%$  ta có  $u_{\alpha} = 2,58 \Rightarrow u = 2,93 > u_{\alpha} = 2,58$  do đó ra loại bỏ giá trị  $x_0 = 19,5$  nghĩa là  $x_0 = 19,5$  chưa sai số thô với mức ý nghĩa  $\alpha = 1\%$

Nhận xét: Với mức ý nghĩa  $\alpha = 1\%$  giá trị bất thường  $x_0 = 19,5$  chưa sai số thô, nguyên nhân do vi phạm các điều kiện

cơ bản của quan trắc hoặc do sơ suất của nhân viên kỹ thuật. Khi đó việc xử lý phải bỏ đi hoặc quan trắc lại.

- Nhiệt độ quan trắc tháng 10/2015

Sử dụng công thức (1) và (3) ta tính được:  $x \approx 28,3$ ;  $S \approx 0,597$

Theo lý thuyết phân bố chuẩn, 99% các giá trị  $x_i$  có phân bố chuẩn sẽ nằm trong khoảng  $(\bar{x} - 2,58.S; \bar{x} + 2,58.S) \approx (26,76; 29,84)$

Trong Bảng 1, bộ số liệu tháng 10 không có giá trị quan trắc nào nằm ngoài khoảng  $(26,76; 29,84)$  do đó trong dãy số liệu thu được tháng 10/2015 không có giá trị bất thường, nghĩa là không có giá trị nào chứa sai số thô. Như vậy, từ kinh nghiệm rút ra trong quá trình xử lý số liệu tháng 4/2015, các điều kiện cơ bản của quá trình quan trắc đã được kỹ thuật viên làm theo đúng yêu cầu và không có sơ suất trong quá trình quan trắc.

Câu hỏi đặt ra là: Khi các giá trị quan trắc nhiệt độ tháng 10/2015 không chứa sai số thô, các sai số còn lại đều là sai số ngẫu nhiên tuân theo luật phân phối chuẩn, vậy độ chính xác của dụng cụ đo đặc thế nào?

Sử dụng công thức (6) ta lần lượt tính được độ chính xác của bảng số liệu quan trắc tháng 4/2015 và tháng 10/2015 lần lượt là  $S_1 = 1,77$ ;  $S_2 = 0,597$ . Số liệu này

## Nghiên cứu

hoàn toàn phù hợp vì trong bảng số liệu tháng 4/2015 vẫn còn chừa sai số thô, trong khi bảng số liệu tháng 10/2015 đã không mắc phải sai số thô. Hay nói cách khác,

### 3.3. Tổng Nitơ trong nước biển khu vực miền Trung năm 2015.

Bảng 2. Tổng Nitơ trong nước biển khu vực miền Trung năm 2015 [9]

Đợt đo	Giá trị quan trắc (đơn vị: mg/l)									
	Đèo Ngang	Thuận An	Đà Nẵng	Dung Quất	Sa Huỳnh	Quy Nhơn				
Tháng 4	0,1 0,1 0,15	0,2 0,2 0,3	0,2 0,16 0,4	0,1 0,25 0,6 0,54	0,3 0,35 0,62 0,78	0,37 0,25 0,36 0,54	0,43 0,8 0,75 0,78	0,48 0,15 0,68 0,28	0,3 0,53 0,28	0,38 0,78 0,5 0,74 0,7 0,2
	0,5 0,8 0,6 0,3	1,0 0,2 0,5 0,7	0,2 0,55 0,8 0,8	0,9 0,8 0,7 0,7	0,6 0,4 0,4 0,7	0,58 0,3 0,6 0,6	0,8 0,7 0,7 0,7	0,6 0,55 0,7 0,7	0,7 0,5 0,8 0,8	0,9 0,6 0,95 0,4 0,73 0,35 0,7 0,8

#### Tổng Nitơ quan trắc tháng 4/2015

Sử dụng công thức (1) và (3) ta tính được:  $x = 0,415$ ;  $S_1 = 0,22$

Theo lý thuyết phân bố chuẩn, 99% các giá trị  $x_i$  có phân bố chuẩn sẽ nằm trong khoảng  $(\bar{x} - 2,58.S_1; \bar{x} + 2,58.S_1) \approx (-0,153; 0,9826)$

Trong Bảng 2, tất cả các giá trị trong bộ số liệu quan trắc tháng 4 đều thuộc khoảng  $(-0,153; 0,9826)$ . Do đó, dãy số liệu thu được không có giá trị bất thường, hay không mắc phải sai số thô.

#### Tổng Nitơ quan trắc tháng 10/2015

Sử dụng công thức (1) và (3) ta tính được:  $\bar{y} = 0,61$ ;  $S_2 = 0,21$

Theo lý thuyết phân bố chuẩn, 99% các giá trị  $x_i$  có phân bố chuẩn sẽ nằm trong khoảng  $(\bar{y} - 2,58.S_2; \bar{y} + 2,58.S_2) \approx (0,068; 1,1518)$

Trong Bảng 2, tất cả các giá trị trong bộ số liệu quan trắc tháng 4 đều thuộc khoảng  $(0,068; 1,1518)$ . Do đó, dãy số liệu thu được không có giá trị bất thường, hay không mắc phải sai số thô.

độ chính xác của dụng cụ đo trong tháng 10/2015 tốt hơn tháng 4/2015 do trước khi quan trắc, cán bộ quan trắc đã chỉnh máy theo đúng hướng dẫn của nhà sản xuất.

### 3.3. Tổng Nitơ trong nước biển khu vực miền Trung năm 2015.

Bảng 2. Tổng Nitơ trong nước biển khu vực miền Trung năm 2015 [9]

Như vậy bộ số liệu quan trắc về tổng Nitơ trong nước biển khu vực miền Trung năm 2015 không mắc phải sai số thô, nghĩa là trong quá trình qua trắc cán bộ đo đã thực hiện đúng kỹ thuật và không vi phạm các điều kiện cơ bản.

Tuy nhiên, ta nhận thấy trong hai mẫu tháng 4 và tháng 10, tổng Nitơ trung bình tháng 4 thấp hơn so với tháng 10, hay có sự chênh lệch rõ rệt theo mùa. Từ đó giả thiết đặt ra là: Tổng Nitơ trung bình  $\mu_2$  tháng 10 cao hơn tổng Nitơ trung bình  $\mu_1$  tháng 4.

Để trả lời câu hỏi đó, ta kiểm định cặp giả thiết:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_2 > \mu_1 \end{cases}$$

Theo (3.5), ta có:

$$T = \frac{\bar{y} - x}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} = \frac{0,61 - 0,415}{\sqrt{\frac{0,21^2}{40} + \frac{0,22^2}{38}}} \approx 4$$

Với mức ý nghĩa  $\alpha = 0,01 \Rightarrow u_{0,01} = 2,58$  (tra bảng phân phối chuẩn)

Ta có  $T = 4 > u_{0,01} = 2,58$  do đó bác bỏ

giả thiết  $H_0$  chấp nhận giả thiết  $H_1$  nghĩa là giả thiết tổng Nito trung bình tháng 10 cao hơn tổng Nito trung bình tháng 4. Điều này hoàn toàn phù hợp vì nồng độ Nito vào mùa mưa (tháng 10) cao hơn mùa khô (tháng 4).

Trong Bảng 2, đối với số liệu quan trắc tháng 4/2015 ta nhận thấy rằng sai số quan trắc  $\varepsilon = 2,58.S_1 = 2,58.0,22 \approx 0,568$  là cao so với các giá trị quan trắc, hơn nữa số liệu quan trắc không mắc phải sai số thô.

Vậy để sai số quan trắc không vượt

### 3.4. Độ mặn nước biển khu vực miền Trung năm 2015

Bảng 3. Độ mặn nước biển khu vực miền Trung năm 2015 [9]

Đợt đo	Giá trị quan trắc (đơn vị: %)					
	Đèo Ngang	Thuận An	Đà Nẵng	Dung Quất	Sa Huỳnh	Quy Nhơn
Tháng 4	32,8	32,8	24,9	32,8	34,0	31,5
	33,2	33,2	28,7	33,7	33,8	31,7
	33,3	33,3	29,8	32,9	34,2	32,4
	33,6	33,6	33,1	33,7	32,1	31,4
	21,1	21,1	30,2	33,3	33,8	31,7
	27,8	27,8	32,3	34,0	33,6	33,5
	32,6	33,4	32,3	33,1	32,5	32,9
	32,8					
Tháng 10	32,4	23,2	29,9	34,0	21,2	23,7
	32,6	23,2	33,7	33,6	25,4	24,9
	32,7	32,7	31,9	34,1	23,4	20,2
	33,0	25,6	33,5	25,6	21,6	20,1
	30,7	33,2	32,9	24,3	25,1	32,8
	30,5	26,9	33,7	23,1	25,2	33,6
	17,9					

Độ mặn quan trắc tháng 4/2015

Sử dụng công thức (1) và (3) ta tính được:  $x = 32$ ;  $S = 2,65$

Theo lý thuyết phân bố chuẩn, 99% các giá trị  $x_i$  có phân bố chuẩn sẽ nằm trong khoảng  $(\bar{x} - 2,58.S; \bar{x} + 2,58.S) \approx (25,2; 38,8)$

Trong Bảng 3, bộ số liệu tháng 4 có 2 giá trị quan trắc  $x_1 = 21,1$ ;  $x_2 = 24,9 \notin (19,63; 28,77)$

Do đó giá trị quan trắc  $x_1 = 21,1$ ;

quá  $0,08 \text{ mg/l}$  ta cần quan trắc thêm bao nhiêu giá trị nữa?

Theo (3.4) thì số giá trị quan trắc phải thỏa mãn:

$$n \geq \left( \frac{S_1 \cdot u_\alpha}{\varepsilon} \right)^2 = \left( \frac{0,22 \cdot 2,58}{0,08} \right)^2 \approx 50,3$$

Vậy số giá trị cần quan trắc ít nhất là 51 giá trị, nghĩa là ta cần quan trắc bổ sung thêm  $51 - 38 = 13$  giá trị quan trắc nữa.

$x_2 = 24,9 \notin (19,63; 28,77)$  trong dãy số liệu thu được là giá trị bất thường.

Vậy các giá trị bất thường  $x_1 = 21,1$ ;  $x_2 = 24,9$  có mắc phải sai số thô hay không? Để trả lời câu hỏi đó ta xem xét các số liệu thu được còn lại (gồm 35 giá trị) là tập hợp riêng còn các giá trị  $x_1 = 21,1$ ;  $x_2 = 24,9$  được xem như là quan trắc đột biến.

Ta có  $x' = 32,5$ ;  $S \approx 1,47$  vì  $\sigma$  chưa biết và số giá trị quan trắc là  $35 > 30$  nên theo

(3.2) ta có:

## Nghiên cứu

$$u_1 = \frac{|\bar{x}_1 - \bar{x}'|}{S\sqrt{\frac{n+1}{n}}} = \frac{|21,1 - 32,5|}{1,47\sqrt{\frac{35+1}{35}}} \approx 7,65$$

$$\text{và } u_2 = \frac{|\bar{x}_2 - \bar{x}'|}{S\sqrt{\frac{n+1}{n}}} = \frac{|24,9 - 32,5|}{1,47\sqrt{\frac{35+1}{35}}} \approx 5,09$$

Với mức ý nghĩa  $\alpha=1\%$  ta có:  $u_{\alpha}=2,58 \Rightarrow u_1=7,65 > u_{\alpha}=2,58$ ;  $u_2=5,09 > u_{\alpha}=2,58$  do đó ra loại bỏ giá trị  $x_1=21,1$ ;  $x_2=24,9$  nghĩa là  $x_1=21,1$ ;  $x_2=24,9$  chưa sai số thô với mức ý nghĩa  $\alpha=1\%$

**Nhận xét:** với mức ý nghĩa  $\alpha=1\%$  các giá trị bất thường  $x_1=21,1$ ;  $x_2=24,9$  chưa sai số thô, nguyên nhân do vi phạm các điều kiện cơ bản của quan trắc hoặc do số suất của nhân viên kỹ thuật. Khi đó việc xử lý phải bỏ đi hoặc quan trắc lại.

Độ mặn quan trắc tháng 10/2015

Trong Bảng 3, bộ số liệu tháng 10 ta nhận thấy có một quan trắc đột biến so với các giá trị quan trắc còn lại là  $x_0=17,9$  còn 36 quan trắc còn lại có giá trị trung bình và độ lệch lần lượt là:  $\bar{x} \approx 28,45$ ;  $S \approx 4,84$ . Câu hỏi đặt ra là giá trị  $x_0=17,9$  có chưa sai số thô hay không?

Tính theo (3.2) ta có:

$$u = \frac{|\bar{x}_0 - \bar{x}|}{S\sqrt{\frac{n+1}{n}}} = \frac{|17,9 - 28,45|}{4,84\sqrt{\frac{36+1}{36}}} \approx 2,15$$

Với mức ý nghĩa  $\alpha=1\%$  ta có  $u_{\alpha}=2,58 \Rightarrow u=2,15 < u_{\alpha}=2,58$  do đó giá trị  $x_0=17,9$  không chưa sai số thô nghĩa là nhận lại giá trị  $x_0=17,9$  vào bộ số liệu quan trắc độ mặn tháng 10/2015.

Khi đó sử dụng công thức (1) và (3) ta tính được:  $\bar{x} = 28,16$ ;  $S = 5,08$

Theo lý thuyết phân bố chuẩn, 99% các giá trị  $x_i$  có phân bố chuẩn sẽ nằm trong khoảng:

$$\left( \bar{x} - 2,58 \cdot S; \bar{x} + 2,58 \cdot S \right) \approx (15,05; 41,27) \\ \Rightarrow x_0 = 17,9 \in (15,05; 41,27)$$

Nghĩa là giá trị  $x_0=17,9$  không phải là giá trị bất thường cân loại bỏ, phù hợp với kết luận giá trị  $x_0=17,9$  không chứa sai số thô.

## 4. Kết luận và kiến nghị

### 4.1. Kết luận

Hiện nay, có ba bộ chương trình chuyên dụng phục vụ cho xử lý và phân tích số liệu thống kê rất thông dụng trên thế giới, đó là SAS, SPSS và STATA. SAS là bộ chương trình mà nhiều người sử dụng có trình độ cao ưa thích bởi sức mạnh và khả năng lập trình của nó. Do SAS là một bộ chương trình mạnh như vậy nên khó học nhất. Để sử dụng SAS, ta phải viết chương trình để thao tác dữ liệu và thực hiện các phân tích dữ liệu của mình. Nếu chương trình mắc lỗi, cái khó là phải biết tìm lỗi ở đâu và cách sửa thế nào. SPSS cũng có một ngôn ngữ cú pháp có thể học bằng cách dán cú pháp lệnh vào cửa sổ cú pháp từ một lệnh vừa chọn và thực hiện, nhưng nói chung khá phức tạp và không trực giác. STATA sử dụng các lệnh trực tiếp, có thể vào mỗi lệnh ở một thời điểm để thực hiện. Tuy nhiên, cả 3 loại phần mềm này đều có chung một hạn chế là người sử dụng phải đầu tư một khoản kinh phí lớn để mua.

Với kết quả của bài báo người sử dụng có thể tính toán, xử lý số liệu quan trắc bằng máy tính bỏ túi, và đặc biệt không cần mua phần mềm. Tác giả đã thu được một số kết quả chính sau:

+ Áp dụng được hàm phân phối chuẩn để loại bỏ giá trị bất thường trong số liệu quan trắc.

+ Áp dụng được hàm phân phối chuẩn để phát hiện và loại bỏ sai số thô mắc phải trong số liệu quan trắc.

+ Áp dụng được hàm phân phối chuẩn để so sánh được giá trị trung bình giữa các bộ số liệu quan trắc độc lập.

+ Phương pháp sử dụng hàm phân phối chuẩn để xử lý số liệu thực nghiệm

được nêu trong đề tài cũng được áp dụng vào xử lý số liệu thực nghiệm trong các lĩnh vực khác như kinh tế, nông - lâm - ngư nghiệp, điều tra xã hội học, kinh tế, y tế, sinh học, ...

#### **4.2. Kiến nghị**

+ Trước khi đo, cán bộ làm công tác thực nghiệm cần thực hiện đúng các điều kiện cơ bản của quan trắc, cũng như cần phải chỉnh máy theo đúng hướng dẫn của nhà sản xuất để tránh mắc phải sai số thô.

+ Tiến hành lặp lại quá trình quan trắc nhiều hơn.

#### **TÀI LIỆU THAM KHẢO**

[1]. Barnett, V. (Ed.) (2004). *Environmental Statistics. Methods and Applications* John Wiley & Sons, Ltd.

[2]. PGS.TS. Nguyễn Kỳ Phùng (2014). *Phương pháp xử lý số liệu thống kê trong môi trường*. NXB ĐHQG TPHCM.

[3]. Chu Đức, Hoàng Chí Thành (2006). *Tính toán trong các hệ sinh thái*. NXB Giao thông vận tải.

[4]. PGS.TS. Tạ Thị Thảo (2010). *Giáo trình thống kê trong hóa phân tích*. NXB ĐHQGHN.

[5]. TS. Ché Đình Lý (2014). *Thống kê và xử lý dữ liệu môi trường*. NXB ĐHQG TPHCM.

[6]. Chu Đức (2001). *Mô hình toán các hệ thống sinh thái*. NXB ĐHQG HN.

[7]. Nguyễn Cao Văn, Trần Thái Ninh (1996). *Lý thuyết xác suất và thống kê toán*. NXB Khoa học kỹ thuật.

[8]. Đặng Hùng Thắng (2008). *Thống kê và ứng dụng*. NXB Giáo dục.

[9]. Viện Hàn lâm Khoa học và Công nghệ Việt Nam (2017). *Báo cáo tổng hợp kết quả quan trắc môi trường vùng biển ven bờ miền Trung*.

[10]. Tuân, Trần Văn (2009). [http://www.ykhoanet.com/baigiang/lamsangthongke/lstk\\_uoctinhcomau.pdf](http://www.ykhoanet.com/baigiang/lamsangthongke/lstk_uoctinhcomau.pdf)

[11]. Luật bảo vệ môi trường (2014).

BBT nhận bài: 02/10/2019; Phản biện xong: 28/10/2019

#### **ỨNG DỤNG PHƯƠNG PHÁP HỒI QUY BỘI TÍNH TOÁN...**

(Tiếp theo trang 44)

#### **5. Kết luận**

Công tác dự báo mực nước lớn nhất bằng phương pháp hồi quy bội tại trạm Tân Châu trên lưu vực sông Cửu Long có ý nghĩa rất lớn đến phát triển kinh tế - xã hội của khu vực Nam Bộ đặc biệt trong thời kỳ mưa lũ. Trong những năm gần đây lượng dòng chảy phía thượng nguồn sông Mê Công ngày một suy giảm do có sự tham gia hoạt động của các công trình lấy nước, hồ chứa. Trong nghiên cứu này phương pháp hồi qui bội, phục vụ dự báo mực nước lớn nhất tại trạm Tân Châu trên sông Cửu Long có một ý nghĩa quan trọng nhằm rút ngắn thời gian tính toán và nâng cao chất lượng dự báo. Kết quả tính toán cho thấy chất lượng dự báo mực nước tại trạm Tân Châu trên sông Cửu Long đã tăng lên đáng kể với chất lượng dự báo

đều đạt trên 80%.

#### **TÀI LIỆU THAM KHẢO**

[1]. Nguyễn Lan Châu (2001). *Xây dựng các phương án nhận định hạn dài đỉnh lũ năm các sông chính ở Việt Nam*. Đề tài nghiên cứu khoa học cấp Tổng cục.

[2]. Hà Văn Khối (2006). *Giáo trình Quy hoạch và Quản lý Tài nguyên nước*. Trường Đại học Thủy Lợi, Hà Nội.

[3]. Nguyễn Văn Tuân, Đoàn Quyết Trung, Bùi Văn Đức (2003). *Dự báo Thủy văn*. NXB Đại học Quốc Gia.

[4]. Trần Thanh Xuân (2007). *Đặc điểm thủy văn và nguồn nước sông Việt Nam*. NXB Nông nghiệp, Hà Nội.

[5]. <https://fsppm.fuv.edu.vn>

BBT nhận bài: 30/10/2019; Phản biện xong: 06/11/2019