**Dự đoán mực nước Thủy văn bằng Time series analysis**

Contents

[A. Tổng quan 1](#_Toc103553156)

[B. Model 8](#_Toc103553157)

[1. Linear regression 8](#_Toc103553158)

[2. SARIMA 9](#_Toc103553159)

[a) Chạy SARIMA cho data ban\_don 12](#_Toc103553160)

[B) Chạy SARIMA cho data an\_thuan 13](#_Toc103553161)

[C. Kết luận 18](#_Toc103553162)

# A. Tổng quan

Hydrology (Thủy văn) là một môn khoa học liên ngành gồm toán, vật lý, cơ học, tin học và hóa học để nghiên cứu quy luật và sự vận động của nước trên bề mặt, lòng đấy, hay sông hồ. Các ứng dụng chính của thủy văn: Dự báo hạn hán, lũ lụt, phòng chống thiên tai, xây dựng đê điều cầu đường, công trình thủy, quy hoạch tài nguyên nước.

Diagram

Description automatically generated

Hình 1: Vòng tuần hoàn nước

Cũng tương tự như các ngành khoa học khác, thủy văn cũng có mô hình dự báo. Việc dự báo chính xác mực nước ở các hệ thống thủy lợi là một trong những yêu cầu cấp bách nhằm trợ giúp việc giảm thiểu các rủi ro do lũ gây ra và có ý nghĩa quan trọng trong việc xây dựng phương án phòng, chống lũ.

Một số mô hình số về thủy lực và thủy văn truyền thống đang được sử dụng để dự báo ở Việt Nam cũng như trên thế giới. Các mô hình này yêu cầu một số lượng lớn các dữ liệu đầu vào như: đặc điểm lưu vực, quan hệ mưa - dòng chảy, địa hình, dự báo lượng mưa, quan hệ lưu lượng - mực nước theo thời gian tại một số vị trí.

Diagram, engineering drawing

Description automatically generated

Hình 2: Mô hình thủy văn

Các mô hình loại này dược xây dựng để mô phỏng đặc tính vật lý và quá trình của dòng chảy. Do khả năng mô phỏng có độ chính xác cao các quá trình vật lý và phân tích do nhạy cảm một cách toàn diện, hơn nữa các mô hình này rất hữu ích cho các nhà khoa học trong việc giải thích được toàn bộ quá trình ẩn đẳng sau, do đó các mô hình loại này được áp dụng khá rộng rãi ở nhiều khu vực trên thế giới Tuy nhiên, việc sử dụng các mô hình này thường yêu cầu một lượng dữ liệu chi tiết về đặc tính của lưu vực như các số liệu thông tin địa lý, mưa, dòng chảy, địa chất… Bên cạnh nó việc hiệu chỉnh và kiểm định mô hình cũng rất phức tạp và đòi hỏi nhiều thời gian, kinh nghiệm và kiến thức của người chạy mô hình đối với từng lưu vực cụ thể. Vì vậy, khả năng áp dụng loại mô hình này ở nhiều khu vực và trong các bài toán dự báo thời đoạn ngắn vẫn còn bị hạn chế. Những hạn chế của các mô hình truyền thống nêu trên đã khuyến khích sự phát triển của các mô hình dựa vào số liệu (data-driven models), mà phổ biến nhất gần đây có thể kể đến là phương pháp máy học (Machine Learning - ML). Các mô hình ML là công cụ tiềm năng trong việc dự báo dòng chảy do các mô hình này có thể được xây dựng dựa nhanh chóng, dễ dàng, không đòi hỏi phải có sự hiểu biết về các quá trình vật lý ẩn đằng sau. Ngoài ra, lượng dữ liệu yêu cầu tối thiểu, cùng với khả năng tính toán, hiệu chỉnh và kiểm định nhanh hơn so với các mô hình vật lý truyền thống, và cách sử dụng ít phức tạp hơn là những ưu điểm lớn mà các mô hình dựa vào số liệu mang lại.

***Do đó, nghiên cứu này được thực hiện nhằm mục đích áp dụng và so sánh khả năng dự báo của các mô hình Machine Learning (SARIMA) ở các trường hợp tính toán khác nhau với các điều kiện số liệu khác nhau qua đó tìm ra được mô hình phù hợp cho nghiên cứu đặc điểm thủy văn và công tác dự báo phòng chống thiên tai.***

Do hệ thống sông của thủy văn có 2 kiểu vùng sông là: sông ảnh hưởng triều bởi thủy triều và vùng ngọt ko ảnh hưởng thủy triều nên ta có 2 bộ dữ liệu gồm: ban\_don (Triều) và an\_thuan (Ngọt)

An thuận và Bản Đôn là tên 2 trạm đo thủy văn của nước ta thuộc 2 vùng địa lý tương ứng là Nam Bộ và Tây nguyên. 2 Trạm được đặt tại 2 con sông lần lượt là sông Hàm Luông (Bến Tre) và Sê Rê Pôk (Đăk Lăk).

|  |  |
| --- | --- |
| An thuận – an\_thuan    **87648 rows × 2 columns** | Bản Đôn – ban\_don    **1096 rows × 5 columns** |

Hai bộ dữ liệu trên được lấy tại [Trung tâm Quan trắc Khí tượng thủy văn quốc gia](http://vnmha.gov.vn/chi-dao-dieu-hanh-103/trung-tam-quan-trac-khi-tuong-thuy-van-nghien-cuu-danh-gia-thu-nghiem-cong-nghe-thiet-bi-quan-trac-moi-9041.html).

Bộ dữ liệu gồm 4 cột tương ứng với 4 biến X H Q R là:

* H: Mực nước
* X: Lượng mưa
* Q: Lưu lượng
* R: Lưu lượng chất lơ lửng

1. Data an\_thuan

Bộ data an\_thuân chỉ gồm 1 biến H kéo dài 10 năm từ năm 2010 đến 2019

Bộ dữ liệu ghi lại mực nước từng giờ của trạm đo vì vậy khi visualize data bộ dữ liệu quá dày nên ta sẽ lấy bộ dữ liệu ghi lại mực nước trong trung bình từng ngày trong tháng để train mô hình.

|  |  |
| --- | --- |
| An thuận – an\_thuan | An thuận – an\_thuan2 |
|  |  |

Chart

Description automatically generated

- Bộ dữ liệu có 3 năm từ 2016 – 2018: ta chi bộ dữ liệu thành 2 phần train (2016, 2017) và test (2018).

2. Data ban\_don

* Dữ liệu đầy đủ không bị mất dữ liệu
* Dữ liệu có 2 kiểu là int64 và float64;
* Biến X dữ liệu object là do trong đo đạc thủy văn những ngày không mưa thường được kí hiệu bằng đấu gạch ngang -;
* Date là dữ liệu ngày ở dạng object.

|  |  |
| --- | --- |
| - Ta chuyển bộ dữ liệu về dạng chuỗi thời gian (áp dụng cho cả bộ data an\_thuan) | Text  Description automatically generated |

Biến X dữ liệu object là do trong đo đạc thủy văn những ngày không mưa thường được kí hiệu bằng đấu gạch ngang ( - ). Do đó ta sẽ chuyển đấu ( - ) thành giá trị 0 là những ngày ko mưa để phù hợp với mô hình.

Table

Description automatically generated

- Trong bộ data ban\_don H là mực nước được tính là (cm) quá lớn so với Q (m3/s) và R (g/m3)

--> Ta quy đổi H từ cm sang m để phù hợp khi đưa vào mô hình tính toán.

**Text, table

Description automatically generated**

**Chart, histogram

Description automatically generated**

- Ở đường quá trình mực nước giờ ta thấy dữ liệu được diễn biến theo mùa. Mùa lũ (mùa có mực nước lớn) thường diễn ra vào tháng từ tháng 10 đến tháng 12.

Ta xét sự tương qua của dữ liệu

* H với Q có tương quan tốt nhất;
* Tiếp đến là H vs R và R vs Q;
* Ta sẽ loại biến X khi chạy mô hình độ tương quan thấp.

Chart, treemap chart

Description automatically generated

H quan hệ với Q là 1 đường cong Chart, scatter chart

Description automatically generated

# B. Model

## 1. Linear regression

Mô hình hồi quy tuyến tính (Linear Regression-LR) [6] được chia thành 2 loại: hồi quy đơn biến (là mô hình hồi quy với một biến) và hồi quy đa biến (là mô hình hồi quy với nhiều biến). Xét mô hình hồi quy tổng quát để giải bài toán diễn đoán thủy lực, thủy văn xác định mức độ hài lòng của các hộ dân dùng dịch vụ nước tưới tiêu, thông thường được viết như sau:

Y= E(X,β )

Trong đó

* X: Các biến dự báo/độc lập (predictor/independent variables);
* Y: Các biến đáp ứng/phụ thuộc (responses/dependent variables);
* β : Các hệ số hồi quy (regression coefficients).

X dùng đề giải thích sự biến đổi của các đáp ứng Y. Y dùng để mô tả các hiện tượng (phenomenon) được quan tâm giải thích. Quan hệ giữa Y và X được diễn tả bởi sự phụ thuộc hàm của Y đối với X. β mô tả sự ảnh hưởng của X đối với Y.

A picture containing chart

Description automatically generated

Chạy mô hình Linear regression cho bộ data ban\_don

|  |  |
| --- | --- |
|  |  |

Mô hình ko chính xác, sai số lớn chuyển sang mô hình sARIMA

## 2. SARIMA

*Tổng quan mô hình SARIMA*

**AR(p)**

AR(p): autoregressive model mô hình tự động hồi quy, tham số p là một số nguyên xác nhận có bao nhiêu chuỗi bị trễ sẽ được sử dụng để dự báo các khoảng thời gian sắp tới, ví dụ:

Mực nước (H) trung bình của ngày hôm qua có mối tương quan cao với mực nước của ngày hôm nay, vì vậy chúng tôi sẽ sử dụng thông số AR (1) để dự báo mực nước trong tương lai.

AR (p) là: ****

* μ là số hạng không đổi, (constant term)
* p là khoảng thời gian được sử dụng trong hồi quy (the periods)
* θ là tham số phù hợp với dữ liệu. (the parameter fitted)

I (d) là phần khác biệt (he differencing part), tham số d cho biết có bao nhiêu thứ tự phân biệt sẽ được sử dụng, nó cố gắng làm cho chuỗi là tĩnh.

**MA (q)**

MA (q) là (**moving average**) mô hình trung bình động, q là số lỗi dự báo trễ trong phương trình dự đoán, ví dụ: Nó giả định rằng các lỗi trong quá khứ sẽ tương tự trong các sự kiện trong tương lai.

MA (p) là: 

* μ là số hạng không đổi,
* q là khoảng thời gian được sử dụng trên số hạng e
* Θ là tham số phù hợp với các lỗi
* 

**SARIMA**

p, d, q được viết hoa để khác với các tham số ARIMA.

**SAR(P)** (the seasonal autoregression of the series): chuỗi tự động hồi quy theo mùa

SAR (P) là: 

Trong đó :

* P là số lượng các số hạng tự động hồi quy sẽ được thêm vào (thường không quá 1 số hạng),
* s là số khoảng thời gian trước được sử dụng làm cơ sở
* θ là tham số phù hợp với dữ liệu.

Trong hiện tượng tự nhiên số liệu 12 tháng mang đến thông tin đóng góp cho giai đoạn hiện tại.

Ta sẽ đặt P = 1 (SAR (1)) sẽ thêm bội số Yt-s vào dự báo cho yt

**I (D)** sự khác biệt theo mùa

D không bao giờ được lớn hơn 1 và d + D không bao giờ được lớn hơn 2. Ngoài ra, nếu d + D = 2, thì số hạng không đổi sẽ bị loại bỏ.

**SMA (Q)**

Ta đặt Q = 1 (SMA (1)) sẽ thêm bội số lỗi vào dự báo cho yt

S Đó là khoảng thời gian theo mùa mà bạn sẽ tính toán các điều kiện P, D, Q. Nếu có mối tương quan theo mùa trong 52 tuần thì đây là con số được sử dụng cho tham số 'S'

**Trend**

* 'n' khi không có xu hướng được sử dụng (mặc định).
* ‘C’ chỉ ra một hằng số (tức là thành phần bậc 0 của đa thức xu hướng)
* ‘T’ biểu thị xu hướng tuyến tính theo thời gian
* ‘Ct’ vừa là xu hướng vừa là hằng số.

### a) Chạy SARIMA cho data ban\_don

Chart, histogram

Description automatically generated

* Có xu hướng tuy nhiên xu hướng tăng ko đáng kể. Trong sông mực nước tăng phụ thuộc nhiều vào nhiệt độ hằng năm cùng với tình hình mua lũ và hoạt động sử dụng nước của con người do đó ta sẽ giả định bỏ qua trend và k\_diff đặt thành 1.
* Bộ data có sự khác biệt theo mùa, vì một số cổ tức hàng tháng rõ ràng là nhiều hơn các tháng khác.
* Do đó, k\_seasonal\_diff = True, seasonal\_periods = 12 vì có chu kỳ 12 tháng một lần.

Vẽ Biểu đồ ACF và PACF để xác định Mô hình SARIMA

Chart, histogram

Description automatically generated

* Theo biểu đồ ACF và PACF ở trên, xác định AR (1) cho mô hình không theo mùa.
* MA (1) cho mô hình theo mùa. Từ ACF, Xác nhận data có chu kỳ 12 tháng một lần.

🡪 Vùng sông ảnh hưởng thủy triều mực nước lên xuống bất thường trong năm do ảnh hưởng thủy triều nên mô hình dự đoán số liệu mực nước H với sai số lớn 🡪 mô hình chưa đạt.

### B) Chạy SARIMA cho data an\_thuan

Data an\_thuan2 là dữ liệu dạng rút gọn (mực nước trung bình ngày của data an\_thuan)

Graphical user interface

Description automatically generated with low confidence

- Ta giữ nguyên đơn vị H là cm ko quy đổi ra m

- Bộ dữ liệu có 10 năm từ năm 2010 – 2019: train (2010 đến 2016) và test (2017 đến 2019).

Chart

Description automatically generated

- Bộ dữ liệu ko có trend và diến biến theo mùa vụ

Histogram

Description automatically generated

Chart, line chart

Description automatically generated

- Mực nước trong năm thấp nhất vào khoảng tháng 6 đến tháng 8 là các tháng mùa cạn ít mưa, lượng nước bốc hơi nhiều.

- Mực nước trong năm cao nhất vào khoảng tháng 11 đến tháng 2 là các tháng mùa lũ mưa nhiều.

Xu hướng mực nước trong 10 năm ko tang giảm đồng đều. Do mực nước là yếu tố tự nhiên phụ thuộc rất nhiều vào môi trường và sinh vật sống xung quanh.Chart, line chart

Description automatically generated

- Kiểm tra tính ổn định:

* Một cách để kiểm tra xem chuỗi có đứng yên hay không là sử dụng hàm adfuller, nếu P-Value thấp hơn 5% (số thông thường được sử dụng cho loại nghiên cứu này) thì chuỗi là tĩnh thì ta có thể chạy mô hình.
* Nếu chuỗi không cố định ta transform data để sử dụng logarit tự nhiên, giảm phát, chênh lệch, v.v.
* Các phương pháp kiểm tra tính ổn định trong bài bao gồm:
* Hàm tự tương quan (ACF) (The autocorrelation function) - Dạng đồ thị: Cho biết mối tương quan giữa mực nước hiện tại so với các phiên bản trễ (lagged) của chính nó.
* Tự tương quan một phần (PACF) (The partial autocorrelation):Nó cho thấy mối tương quan mực nước hiện tại so với phiên bản bị trễ (lagged) loại trừ ảnh hưởng của độ trễ trước đó, ví dụ, nó cho thấy ảnh hưởng hiệu quả của độ trễ 3 đối với mực nước hiện tại không bao gồm ảnh hưởng của độ trễ 1 và 2.

Histogram

Description automatically generated

- Như chúng ta có thể thấy, chuỗi thời gian có một xu hướng giảm rồi tăng và điều này nguyên nhân do nguồn nước qua từng năm có thể bị thất thoát do nhiều nguyên nhân như xả đê xả lũ, sử dụng nguồn nước hay do nhiệt độ cao làm nguồn nước bị bốc hơi và thất thoát nguồn nước. Hoặc do rừng ở lưu vực bị chặt phá dẫn đến về mùa cạn thiếu lượng nước ngầm dự trữ để đổ vào sông

- Để tạo dự báo chuỗi thời gian, chuỗi phải cố định (giá trị trung bình không đổi, phương sai và tự tương quan).

Graphical user interface, histogram

Description automatically generated

- Ta có thể thấy sư tương quan âm là do trong các tháng mùa lũ mực nước tăng cao, các tháng mùa cạn mực nước giảm thấp

- Ta chọn mô hình sARIMA ((1,0,0),(0,1,1,12),'t')) theo trend vì có kết quả tốt nhất cải thiện 3.84%.

Chart

Description automatically generated

Chart, line chart

Description automatically generated

# C. Kết luận

- Project bước đầu đã thử nghiệm thành công mô hình SARIMA trong dự báo mực nước sông thủy văn của trạm ngọt vùng sông không ảnh hưởng thủy triều (Bộ dữ liệu an\_thuan2 trong năm 2019 tương ứng với dự báo hạn hán ở mức dài hạn. Các mô hình dự báo cho kết quả khá tốt vào mùa khô.

- Tuy nhiên khi ở bộ dữ liệu ban\_don mô hình lại không dự báo chính xác một cách đồng nhất dòng chảy lũ. Lý do của hiện tượng này là các mô hình không được huấn luyện tập trung vào dự báo dòng chảy lũ mà ưu tiên vào quá trình dòng chảy

- Ngoài ra, việc lựa chọn dữ liệu đầu vào phù hợp là yếu tố rất quan trọng quyết định nên hiệu quả dự báo của mô hình. Trong đó, dòng chảy trong quá khứ đóng là một trong những biến đầu vào quan trọng. Bên cạnh đó, số lượng dữ liệu đầu vào (số lượng features) cần phải đủ lớn để hỗ trợ cho mô hình ML trong việc khái quát hóa được mối quan hệ giữa biến đầu vào và dòng chảy đầu ra.

- Như vậy, bên cạnh các phương pháp truyền thống, các mô hình Machine Learning như Linear Regression hay SARIMA sẽ cung cấp một công cụ mới, hiệu quả để bổ trợ cho công tác dự báo phòng chống thiên tai. Các kết quả của nghiên cứu này làm tiền đề cho các nghiên cứu về việc sử dụng các mô hình học máy trong diễn đoán thủy lực, thủy văn của các hệ thống thủy lợi. Tuy nhiên, việc ứng dụng trong tương lai phụ thuộc rất lớn vào điều kiện và chất lượng số liệu ở địa điểm áp dụng.

***Xin chân thành cảm ơn !!!***