Dự đoán sản lượng thủy sản bằng các mô hình máy học

Nguyễn Trương Minh Văn, Nguyễn Hoàng Quý, Hồ Anh Khôi, Đỗ Trọng Hợp

Trường Đại học Công nghệ Thông tin - Đại học Quốc gia Thành phố Hồ Chí Minh
{20522146, 20521815, 20521477}@gm.uit.edu.vn, hopdt@uit.edu.vn

Tóm tắt nội dung. Trong bài báo này chúng tôi sẽ tiến hành so sánh, phân tích các mô hình hồi quy, mô hình chuỗi thời gian. Sử dụng các tiêu chuẩn của thống kê để lựa chọn mô hình tốt nhất.

Keywords: Hồi quy, chuỗi thời gian, VAR, ARIMA, ARIMAX, LSTM

1 Giới thiệu

Nông nghiệp nói chung và ngành thủy sản nói riêng giữ vai trò quan trọng trong sự phát triển của nền kinh tế Việt Nam với quy mô ngày càng mở rộng. Thương hiệu thủy sản Việt Nam không chỉ được khẳng định trong nước mà còn được đón nhận bởi nhiều quốc gia trên thế giới. Nhận thấy rằng ngành thủy sản chính là một trong những ngành kinh tế mũi nhọn của nước ta. Chính vì vậy, chúng tôi đã xây dựng mô hình dự đoán sản lượng thủy sản qua từng năm dựa vào một số thuộc tính như là diện tích nuôi trồng, số tàu khai thác biển, tổng sản lượng thu được để cho ra mô hình dự đoán áp dụng các phương pháp máy học tiêu biểu là ARIMAX, VAR, LSTM, Hồi quy tuyến tính.

2 Bộ dữ liệu

Bộ dữ liệu về sản lượng thủy sản Việt Nam được tổng hợp ở Tổng cục thống kê và một số nguồn khác:

Biến số	Mô tả
Nam	Năm
ChiSoPhatTrien	Chỉ số phát triển, được tính bằng sản lượng năm nay chia sản lượng năm trước * 100.
DienTichNuoiTrong	Diện tích nuôi trồng
SoTauKhaiThac	Số tàu khai thác
TongSanLuong	Tổng sản lượng

Bảng 1: Mô tả các biến số

Bộ dữ liệu được thu thập hàng năm từ năm 1990 đến năm 2020, có tổng cộng 25 điểm dữ liệu.

3 Các mô hình sử dụng dự đoán sản lượng thủy sản

3.1 Mô hình hồi quy tuyến tính

Mô hình hồi quy tuyến tính đưa ra dự đoán bằng cách tính tổng trọng số các đặc trưng đầu vào, sau đó cộng tổng này với một hằng số gọi là hệ số điều chỉnh (bias term, hoặc còn được gọi là hệ số chặn – intercept term)

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

Trong đó:

- $\hat{y}$ là giá trị dự đoán
- n là số lượng đặc trưng
- x_i là đặc trưng thứ i
- θ_j là tham số thứ j của mô hình (bao gồm cả hệ số điều chỉnh θ_0 và các trọng số đặc trưng $\theta_1, \theta_2, ..., \theta_n$).

3.2 Mô hình chuỗi thời gian

a) Vector Autoregression (VAR)

VAR là một mô hình chuỗi thời gian đa biến có thể được sử dụng để dự đoán nhiều hơn một biến số. Nó được sử dụng trong các tình huống trong đó các biến có sự phụ thuộc lẫn nhau. Trong mô hình VAR, mỗi biến được mô hình hóa dưới dạng kết hợp tuyến tính giữa các quan sát trong quá khứ của chính nó và các biến khác. Do đó nó có thể được mô hình hóa như một phương trình, trong đó mỗi biến nhận một phương trình có thể được biểu diễn dưới dạng vector. Giả sử chúng ta có vectơ dữ liệu chuỗi thời gian Y_t, khi đó mô hình VAR với k biến và độ trễ p có thể được biểu diễn toán học trong công thức (1), trong đó Y_t, β_0 là k × 1 vector cột và β_0, β_1, β_2, …, β_p là k × k ma trận hệ số.

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \varepsilon_t \quad (1)$$

Nếu chuỗi thời gian không có tính dừng, chúng ta phải biến đổi thành các chuỗi thời gian dừng bằng phương pháp Differencing trước khi huấn luyện mô hình, sau đó đảo ngược giá trị dự đoán thành giá trị dự đoán thực tế.

b) ARIMAX

Phần này chúng tôi sẽ giới thiệu về nền tảng toán học của các mô hình được sử dụng trong bài báo. Trước khi tìm hiểu về mô hình ARIMAX, ta cần phải tìm hiểu về mô hình Auto-Regressive (AR), Moving Average (MA), ARIMA.

Mô hình Autoregressive (AR) phụ thuộc vào độ trễ của nó. Trong một mô hình AR với độ trễ p, giá trị của khoảng thời gian t trong chuỗi thời gian được tính theo công

thức (2) trong đó Y_t là giá trị quan sát trong khoảng thời gian t, α là hằng số, β là tham số hồi quy, ε là sai số dự báo.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \varepsilon_t \quad (2)$$

Trong mô hình ARIMA, Moving Average (MA) đóng vai trò như là một công cụ truy xuất các thông tin về sai số dự báo ở các mốc thời điểm trong quá khứ sẽ ảnh hưởng như thế nào đến sai số dự báo ở mốc thời gian trong tương lai. Công thức tổng quát mô hình MA được biểu diễn như công thức (3) trong đó Y_t là giá trị quan sát trong khoảng thời gian t, α là hằng số, ϕ là hệ số hồi quy.

$$Y_t = \alpha + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \ldots + \phi_q \varepsilon_{t-q} \quad (3)$$

Mô hình Auto-Regressive Moving Average (ARMA) là mô hình cơ bản để phân tích chuỗi thời gian dừng. ARMA là mô hình hợp nhất của AR và MA. Trong mô hình ARMA có p độ trễ và q sai số phần dư, giá trị khoảng thời gian t trong chuỗi thời gian được tính theo công thức (4) trong đó X_t là giá trị dự đoán trong thời gian t, ϕ là hằng số nhân với biến trễ, θ là hằng số nhân với sai số phần dư.

$$Xt = \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (4)$$

Auto-Regressive Integrated Moving Average (ARIMA) là một mô hình dựa trên độ trễ của nó và các lỗi dự báo trễ. Trong mô hình ARIMA có p độ trễ và q sai số phần dư, giá trị của khoảng thời gian t trong một chuỗi thời gian được tính bằng Công thức (5) trong đó Y_t là giá trị quan sát trong khoảng thời gian t, α là hằng số, β là hằng số nhân các biến trễ và ϕ là các hằng số nhân các sai số phần dư.

$$Y_t = \alpha + \beta_1 Y_{t-1} + ... + \beta_p Y_{t-p} \varepsilon_t + \phi_1 \varepsilon_{t-1} + ... + \phi_q \varepsilon_{t-q} \quad (5)$$

Mô hình ARIMAX là một dạng mở rộng của mô hình ARIMA. Mô hình cũng dựa trên giả định về mối quan hệ tuyến tính giữa giá trị và phương sai trong quá khứ với giá trị hiện tại và sử dụng phương trình hồi qui tuyến tính được suy ra từ mối quan hệ trong quá khứ nhằm dự báo tương lai. Mô hình sẽ có thêm một vài biến độc lập khác và cũng được xem như một mô hình hồi qui động (hoặc một số tài liệu tiếng việt gọi là mô hình hồi qui động thái). Về bản chất ARIMAX tương ứng với một mô hình hồi qui đa biến nhưng chiếm lợi thế trong dự báo nhờ xem xét đến yếu tố tự tương quan được biểu diễn trong phần dư của mô hình. Nhờ đó cải thiện độ chính xác.

c) LSTM

Mạng trí nhớ ngắn hạn định hướng dài hạn (Long Short-Term Memory) còn được viết tắt là LSTM là một mạng thần kinh hồi quy (RNN) nhân tạo được sử dụng trong lĩnh vực học sâu. Không giống như các mạng thần kinh truyền thẳng (FNN) tiêu chuẩn, LSTM có chứa các kết nối phản hồi. Mạng không chỉ xử lí các điểm dữ liệu đơn lẻ, mà còn xử lí toàn bộ chuỗi dữ liệu. LSTM rất phù hợp đối với bài toán chuỗi thời gian.

Mô hình LSTM thông thường gồm một tế bào (cell), một cổng vào (input gate), một cổng ra (output gate) và một cổng quên (forget cell). Tế bào ghi nhớ các giá trị trong các khoảng thời gian bất ý và ba cổng sẽ điều chỉnh luồng thông tin ra/vào tế bào.

3.3 Tiêu chuẩn đánh giá mô hình

a) AIC

$$AIC = e^{2k/n}\frac{SSE}{n}$$

Trong đó k là số biến ước lượng (bao gồm cả hệ số chặn), n là số mẫu quan sát, $SSE = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ với $\hat{y}_i$ là giá trị dự báo, y_i là giá trị thực tế

b) RMSE

Căn bậc hai Trung binh Bình phương Sai số (Root Mean Square Error – RMSE) là một phép đo điển hình thường được sử dụng để đánh giá độ chính xác của các dự đoán thu bởi mô hình.

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

Trong đó $\hat{y}_i$ là giá trị dự báo, y_i là giá trị thực tế

c) MAE

Nhìn chung, RMSE khá phổ biến nhưng trong vài trường hợp nếu muôn sử dụng một phép đo khác ta có thể cân nhắc sử dụng trung bình sai số tuyệt đối (mean absolute error – MAE)

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}$$

d) MAPE

MAPE là giá trị trung bình của phần trăm sai số tuyệt đối phổ biến vì nó không phụ thuộc vào quy mô và dễ diễn giải. Gọi y_i và $\hat{y}_i$ lần lượt biểu thị giá trị thực và giá trị dự báo tại điểm dữ liệu và n là số mẫu quan sát. MAPE được tính bằng công thức

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

4 Phương pháp thực hiện

4.1 Phân tích, tiền xử lý dữ liệu

a) Xử lý giá trị thiếu

Biến	Số lượng giá trị bị thiếu	Tỉ lệ
ChiSoPhatTrien	0	0%
DienTichNuoiTrong	0	0%
SoTauKhaiThacBien	6	23%
TongSanLuong	0	0%

Bảng 2. Tỉ lệ số giá trị thiếu

Trong đề tài này, nhóm xử lí dữ liệu thiếu bằng phương pháp nội suy tuyến tính. Phương pháp nội suy tuyến tính được sử dụng trong tình huống muốn điền giá trị thiếu nằm trong khoảng giá trị của dữ liệu hiện có. Công thức nội suy tuyến tính:

$$y = y_1 + \frac{(x - x_1)(y_2 - y_1)}{x_2 - x_1}$$

Trong đó:

+ $(x_1, y_1), (x_2, y_2)$ tương ứng với điểm bắt đầu và kết thúc.
+ y là giá trị cần nội suy tại thời điểm x.

b) Kiểm định ADF

Xác định chuỗi dữ liệu dừng là một bước rất quan trọng trong việc xử lí dữ liệu chuỗi thời gian. Một chuỗi thời gian dừng nếu trung bình và phương sai của nó không đổi qua thời gian và giá trị hiệp phương sai giữa hai giai đoạn chỉ phụ thuộc vào khoảng cách giữa hai giai đoạn ấy chứ không phụ thuộc vào thời gian thực sự tại đó hiệp phương sai được tính. Nếu một chuỗi không dừng thì chúng ta không thể khái quát hóa kết quả phân tích cho các giai đoạn khác. Đối với mục đích dự báo, chuỗi không dừng sẽ không có giá trị thực tiễn, có thể gây ra hiện tượng hồi quy giả.

Vì vậy trong bài toán này, nhóm sử dụng kiểm định ADF để kiểm tra xem chuỗi thời gian có tính dừng hay không.

Với giả thuyết:

- H0: Chuỗi dữ liệu không dừng
- H1: Chuỗi dữ liệu dừng

Kết quả kiểm định ADF trước khi lấy sai phân với mức ý nghĩa 5%:

	ADF	P-value	
ChiSoPhatTrien	-2.8187	0.056	Chuỗi không dừng
DienTichNuoiTrong	-2.6104	0.091	Chuỗi không dừng
SoTauKhaiThac	-0.806	0.817	Chuỗi không dừng
TongSanLuong	3.631	1	Chuỗi không dừng

Bảng 3. Kết quả kiểm định ADF trước khi lấy sai phân với mức ý nghĩa 5%

Ta thấy 4 chuỗi trên đều không dừng. Tiến hành biến đổi chuỗi dừng bằng cách lấy sai phân bậc 2:

$$I(2) = \Delta^2(x_t) = \Delta(\Delta(x_t))$$

Với $\Delta(x_t) = x_t - x_{t-1}$

Kết quả kiểm định ADF khi lấy sai phân bậc 2 với mức ý nghĩa 5%:

	ADF	P-value	
ChiSoPhatTrien	-5.585	1.37e-6	Chuỗi dừng
DienTichNuoiTrong	-2.88	0.047	Chuỗi dừng
SoTauKhaiThac	-5.35	4.16e-6	Chuỗi dừng
TongSanLuong	-5.45	2.636e-6	Chuỗi dừng

Bảng 4. kết quả kiểm định ADF sau khi lấy sai phân bậc 2

Vậy sau khi lấy sai phân bậc 2 thì ta thấy cả 4 chuỗi đều là chuỗi dừng. Vậy ta sẽ dùng sai phân bậc 2 cho bài toán này.

c) Kiểm định đồng liên kết

Đồng liên kết là một thuộc tính thống kê tập hợp của các biến chuỗi thời gian. Đồng liên kết đã trở thành một đặc tính quan trọng trong phân tích chuỗi thời gian đương đại, chuỗi thời gian có xu hướng hoặc ngẫu nhiên. Kiểm tra đồng liên kết được sử dụng để xác định xem có mối tương quan giữa một số chuỗi thời gian trong dài hạn hay không.

Đặt giả thuyết:

- H0: số đồng liên kết = n
- H1: số đồng liên kết > n
- Với n là số đồng liên kết đang xét

Kết quả kiểm định tại mức ý nghĩa 5%:

	Test Statistic	Critical Value	Nhận định
Không có đồng liên kết	174.32	78.87	Có ít nhất 1 đồng liên kết
Có ít nhất 1 đồng liên kết	92.39	55.43	Có ít nhất 2 đồng liên kết
Có ít nhất 2 đồng liên kết	38.15	31.52	Có ít nhất 3 đồng liên kết
Có ít nhất 3 đồng liên kết	12.1	17.95	Có ít nhất 3 đồng liên kết

Bảng 5. Kết quả kiểm định đồng liên kết

Kết quả trên cho thấy các biến có đồng liên kết và mối quan hệ đồng liên kết này đã loại bỏ hiện tượng hồi quy không xác thực, đồng thời xác nhận có một mối quan hệ nhân quả Granger.

d) Kiểm định nhân quả Granger

Kiểm định nhân quả Granger được tiến hành để nghiên cứu chiều hướng tác động giữa tác biến, ví dụ: X tác động một chiều lên Y, Y tác động một chiều lên X hay cả hai đều tác động qua lại lẫn nhau. Phương pháp Granger dựa trên nguyên tắc nếu biến X gây ra biến Y thì một tỉ lệ giá trị của Y tại thời điểm nghiên cứu phải được giải

thích bằng các giá trị quá khứ của X. Nói một cách ngắn gọn, kiểm định nhân quả Granger dùng để xác minh tính hữu dụng trong việc dự đoán của một biến lên các biến khác.

Đặt giả thuyết:

- H0: Không có tác động nhân quả
- H1: Có tác động nhân quả

Kết quả kiểm định với mức ý nghĩa 5%:

	P-value
ChiSoPhatTrien tác động nhân quả lên TongSanLuong	0.1159
DienTichNuoiTrong tác động nhân quả lên TongSanLuong	0.007
SoTauKhaiThacBien tác động nhân quả lên TongSanLuong	0.0403

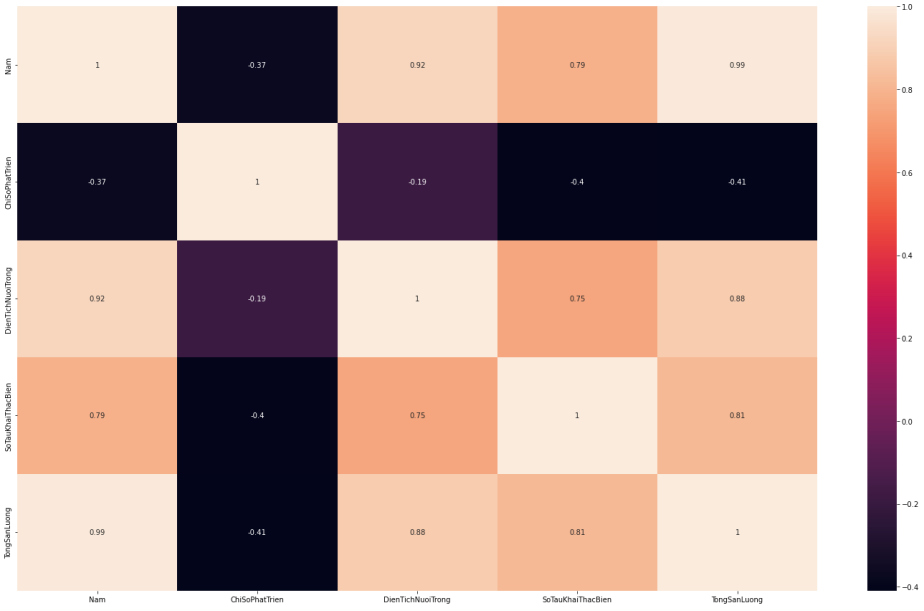
Bảng 6. Kết quả kiểm định nhân quả Granger

Dựa vào kết quả trên ta có kết luận:

- ChiSoPhatTrien không có tác động nhân quả lên TongSanLuong vì 0.1159 > 0.05
- DienTichNuoiTrong tác động nhân quả lên TongSanLuong 0.007 < 0.05
- SoTauKhaiThacBien tác động nhân quả lên TongSanLuong 0.0409 < 0.05

4.2 Thử nghiệm mô hình

a) Thử nghiệm mô hình hồi quy tuyến tính



Hình 1. Ma trận tương quan

Dựa vào ma trận tương quan, ta thấy yếu tố Năm có mối tương quan mạnh mẽ với yếu tố TongSanLuong. Do đó nhóm sẽ chia thành 3 tập dữ liệu khác nhau để xét ảnh hưởng của yếu tố Nam lên TongSanLuong:

- D1: Dữ liệu bao gồm các thuộc tính: ChiSoPhatTrien, DienTichNuoiTrong, SoTauKhaiThacBien và thuộc tính mục tiêu là TongSanLuong
- D2: Dữ liệu bao gồm các thuộc tính: Nam và thuộc tính mục tiêu là Tong-SanLuong
- D3: Dữ liệu bao gồm các thuộc tính: Nam, ChiSoPhatTrien, DienTichNuoi-Trong, SoTauKhaiThacBien và thuộc tính mục tiêu là TongSanLuong.

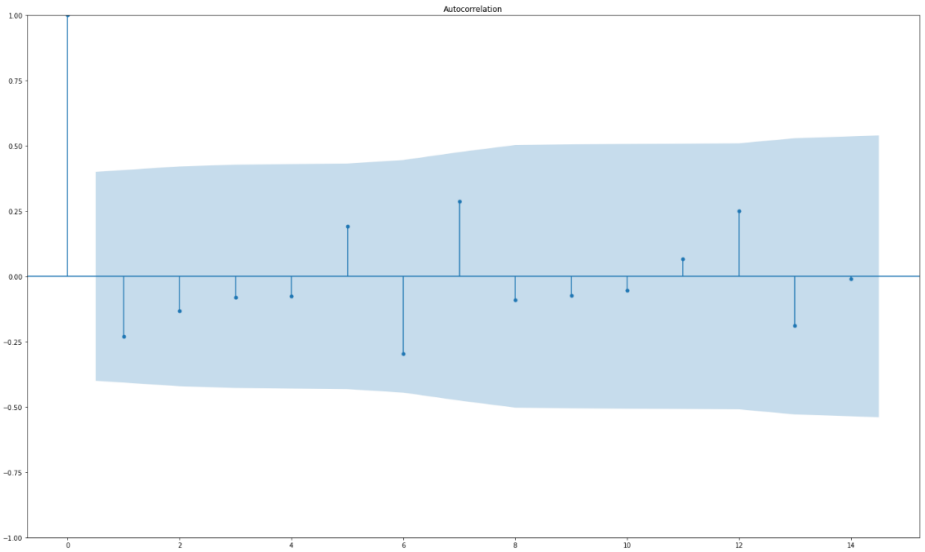
Hàm hồi quy tương ứng với từng tập dữ liệu:

+ D1: $y = -7.25x_1 + 4.0154x_2 + 0.0199x_3 - 1181.648$
+ D2: $y = 247.553x_4 - 492716.353$
+ D3: $y = 237.76x_4 - 0.729x_1 - 0.534x_2 + 0.014x_3 - 473859.036$

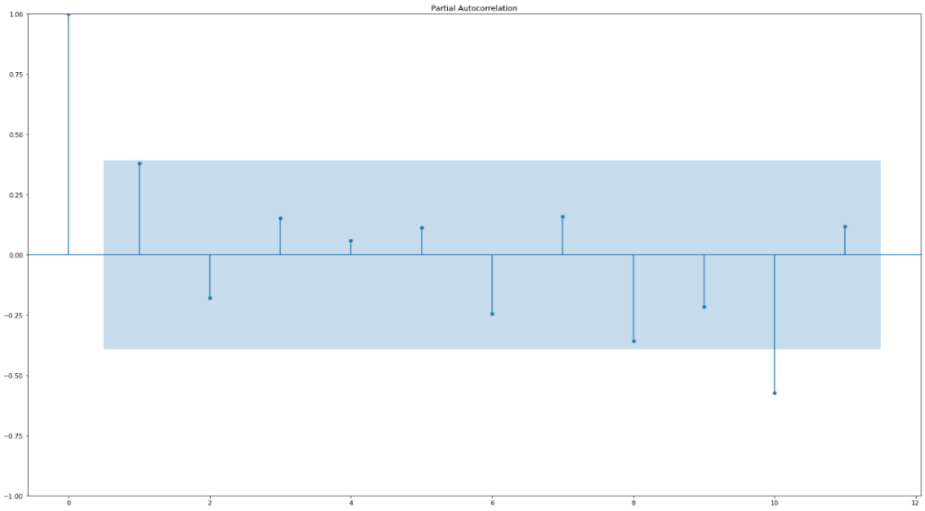
Trong đó:

+ x_1 tương ứng ChiSoPhatTrien
+ x_2 tương ứng DienTichNuoiTrong
+ x_3 tương ứng SoTauKhaiThacBien
+ x_4 tương ứng Nam

b) Thử nghiệm mô hình chuỗi thời gian đơn biến



Hình 2. Hàm ACF sai phân bậc 2 của TongSanLuong



Hình 3. Hàm PACF sai phân bậc 2 của TongSanLuong

Qua việc kiểm định các chuỗi thời gian, biến đổi chuỗi dừng, đồ thị ACF và PACF ta có các mô hình sau:

Mô hình	AIC
ARIMA(1,0,1)	231.314
ARIMA(1,0,2)	231.983
ARIMA(2,0,1)	232.848
ARIMA(2,0,2)	233.954

Bảng 7. Chỉ số AIC của các mô hình ARIMA

So sánh AIC của các mô hình trên ta thấy mô hình có AIC nhỏ nhất là ARIMA(1,0,1). Vậy đây là mô hình đơn biến phù hợp nhất để dự đoán Tổng sản lượng thủy sản.

c) Mô hình chuỗi thời gian đa biến

Vì ChiSoPhatTrien không tác động nhân quả lên TongSanLuong nên dữ liệu có thể chia làm 2 tập như sau:

+ T1: ChiSoPhatTrien, DienTichNuoiTrong, SoTauKhaiThacBien và thuộc tính mục tiêu là TongSanLuong

+ T2: DienTichNuoiTrong, SoTauKhaiThacBien và thuộc tính mục tiêu là Tong-SanLuong

Đối với mô hình chuỗi thời gian đa biến, nhóm sử dụng 3 mô hình là VAR, ARIMAX, LSTM

Mô hình Var chỉ có tham số cần được tinh chỉnh là độ trễ mô hình (p). Tham số mô hình Var được nhóm lựa chọn sẽ tóm tắt trong bảng sau:

Feature Set	Order(p)	AIC
T1	2	29.85
T2	3	23.8231

Bảng 8. Chỉ số AIC của các mô hình VAR

Mô hình ARIMAX được nhóm sử dụng là Auto Arima, được cung cấp bởi thư viện pmdarima. Các biến ngoại suy được cung cấp để mô hình đưa ra dự báo. Các tham số mô hình ARIMAX được nhóm lựa chọn sẽ tóm tắt trong bảng sau:

Feature Set	Model	AIC
T1	ARIMAX(0,0,1)	199.819
T2	ARIMAX(0,0,2)	227.824

Bảng 9. Chỉ số AIC của các mô hình ARIMAX

Mô hình LSTM được nhóm sử dụng bao gồm 1 lớp đầu vào, 2 lớp ẩn gồm 1 lớp LSTM với 1 lớp nơ ron và 1 lớp đầu ra. Kỹ thuật dừng sớm được sử dụng với mức ngưỡng 5 bước. Mô hình sẽ dừng huấn luyện sau 5 bước nếu lỗi huấn luyện và kiểm định đều tăng lên. Điều này giúp mô hình giảm thiểu quá khớp so với việc giảm kích thước epochs.

5 Kết quả

5.1 Mô hình hồi quy

Kết quả mô hình hồi quy tuyến tính cho 3 tập dữ liệu D1, D2, D3 được trình bày trong bảng sau:

	RMSE	MAE	R2
D1	2265.974	2097.977	-4.224
D2	728.239	665.178	0.46
D3	**651.229**	**553.69**	**0.568**

Bảng 10. Kết quả mô hình hồi quy tuyến tính

5.2 Mô hình chuỗi thời gian đơn biến

Kết quả mô hình chuỗi thời gian đơn biến ARIMA(1,0,1) được trình bày trong bảng sau:

	RMSE	MAE	MAPE
ARIMA(1,0,1)	481.85	459.42	0.058

Bảng 11. Kết quả mô hình ARIMA(1,0,1)

5.3 Mô hình VAR

Kết quả mô hình chuỗi thời gian đa biến VAR cho 2 tập dữ liệu T1, T2 được trình bày trong bảng sau:

	Model	RMSE	MAE	MAPE
T1	Var(1)	**99.835**	**97.555**	**0.012**
T2	Var(3)	152.25	136.38	0.018

Bảng 12. Kết quả mô hình VAR

5.4 Mô hình ARIMAX

Kết quả mô hình chuỗi thời gian đa biến ARIMAX cho 2 tập dữ liệu T1, T2 được trình bày trong bảng sau:

	Model	RMSE	MAE	MAPE
T1	ARIMAX(0,0,1)	503.028	478.289	0.06
T2	ARIMAX(0,0,2)	480.146	458.214	0.058

Bảng 13. Kết quả mô hình ARIMAX

5.5 Mô hình LSTM

Kết quả mô hình học sâu LSTM cho 2 tập dữ liệu T1, T2 được trình bày trong bảng sau:

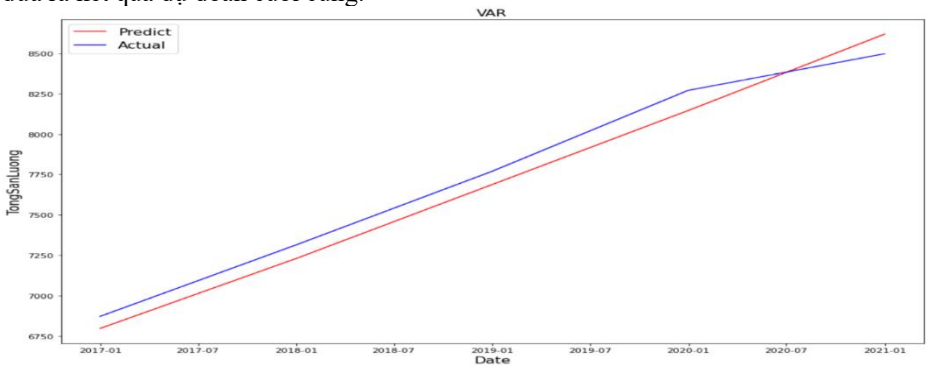
	RMSE	MAE	MAPE
T1	409.35	399.28	0.051
T2	396.7	382.23	0.049

Bảng 14. Kết quả mô hình LSTM

5.6 Dự đoán sản lượng thủy sản từ 2016-2020

Qua kết quả của các mô hình trên, ta thấy mô hình hồi quy tuyến tính có kết quả tốt nhất khi sử dụng tập dữ liệu D3 với các chỉ số RMSE, MAE, R2 nhỏ nhất lần lượt là 651.229, 553.69, 0.568. Tập dữ liệu D3 là tập dữ liệu sử dụng tất cả các yếu tố, trong đó yếu tố Nam là yếu tố rất quan trọng trong mô hình hồi quy.

Nhưng khi áp dụng các mô hình chuỗi thời gian và mô hình học sâu thì hồi quy tuyến tính thông thường có kết quả khá thấp. Mô hình VAR đạt được kết quả tốt nhất với tập dữ liệu T1 có chỉ số RMSE, MAE, MAPE lần lượt là 99.835, 97.555, 0.012. Bên cạnh đó, đối với 2 mô hình ARIMAX và LSTM thì kết quả của tập dữ liệu T2 có kết quả tốt hơn tập dữ liệu T1. Vậy ta sẽ sử dụng mô hình VAR(1) với tập dữ liệu T1 để đưa ra kết quả dự đoán cuối cùng.



Hình 4. Biểu đồ dự đoán từ năm 2016-2020 của mô hình VAR.

6 Kết luận

Trong đề tài này, nhóm đã đưa ra giải pháp dự đoán tổng sản lượng thủy sản thông qua sự so sánh giữa mô hình hồi quy tuyến tính thông thường và các mô hình chuỗi thời gian. Từ kết quả của mô hình hồi quy tuyến tính, ta thấy yếu tố Nam có tác động mạnh mẽ lên độ chính xác của mô hình. Dựa vào điều này, nhóm đã thử áp dụng các mô hình chuỗi thời gian. Nhìn chung, việc sử dụng các mô hình chuỗi thời gian cho kết quả khả quan hơn. Bên cạnh đó, nhóm cũng đã kiểm định các giả thuyết kiểm tra sự phù hợp của các chuỗi thời gian để đưa vào mô hình sử dụng. Mô hình chuỗi thời gian VAR đạt độ chính xác cao nhất với các chỉ số RMSE, MAE, MAPE lần lượt là 99.835, 97.555, 0.012. Trong tương lai, nhóm sẽ tìm và bổ sung thêm dữ liệu mới kết hợp với các mô hình chuỗi thời gian và học sâu mạnh mẽ khác để có độ chính xác cao hơn giúp góp phần đưa ra những dự đoán về tổng sản lượng thủy sản để nước ta có thể đưa ra các chính sách kinh tế, chiến lược phù hợp trong tương lai.

7 Tài liệu

1. A Comparison of ARIMAX, VAR and LSTM on Multivariate Short-Term Traffic Volume Forecasting
2. Stationarity and differencing, Forecasting – Otextshttps
3. Vector Autoregression (VAR) – Comprehensive Guide with Examples in Python
4. Introduction to Granger Causality
5. Cointegration: Definition, Examples, Tests
6. Multivariate Time Series Forecasting with LSTMs in Keras
7. How To Do Multivariate Time Series Forecasting Using LSTM
8. Understanding Time Series Modelling with Auto ARIMAX
9. ARIMAX models — PyFlux 0.4.7 documentation - Read the Docs