Cấu trúc của mô hình dự đoán ARIMA gồm 4 thành phần chính:

* Thành phần nhận dạng mô hình (Model identification): ước lượng và xấp xỉ cấu trúc cho mối liên hệ trong dữ liệu đã thu thập được. Thành phần này nhận dạng được các mẫu một cách tự động bằng cách phân tích hệ số tự tương quan và hệ số tự tương quan bán phần theo các độ trễ khác nhau. Do đó sẽ tránh được việc người dự đoán phải dựa vào các đồ thị dữ liệu để xác định mô hình.
* Thành phần lấy hiệu/tổng hợp (Recursive differencer/integrator): dùng để chuyển đổi các chuỗi dữ liệu không tĩnh hoặc/và có tính mùa sang chuỗi dữ liệu tỉnh trước khi áp dụng các giải thuật để ước lượng các tham số cho mô hình. Bởi vì mô hình ARIMA chỉ được kết hợp với chuỗi dữ liệu tĩnh, các thành phần bị loại bỏ khỏi chuỗi dữ liệu trong quá trình chuyển đổi phải được tích hợp lại vào dữ liệu khi dự đoán.
* Thành phần ước lượng tham số (Parameter estimator): ước lượng tham số của mô hình bằng giải thuật tối thiểu sai số đệ quy. Giải thuật này sử dụng đệ quy để giảm chi phí tính toán và kĩ thuật tối thiểu lỗi để tối thiểu bình phương các ước lượng lỗi. Giải thuật này có hai giả định là: mô hình phải tuyến tính với các tham số và chuỗi lỗi phải có phân phối tự nhiên với giá trị trung bình là 0 và phương sai không đổi. Nếu hai giả định này bị vi phạm thì mô hình ước lượng sẽ không chính xác.
* Thành phần dự đoán (N-step ahead predictor): tạo giá trị dự đoán cho các giai đoạn tiếp theo dựa vào mô hình đã ước lượng được.

MODEL STRUCTURE IDENTIFIER

RECURSIVE DIFFERENCER/ INTEGRATER

PARAMETER ESTIMATOR

N-STEP AHEAD PREDICTOR

Stationary Series

Non-stationary and/or Seasonal series

Forecasts for Original Series

Differenced Series

Estimated Parameters

Estimated Parameters

Forecasts for Original Series

Forecasts for Differenced Series

y(t)

y(t)

(t+n)

(t+n)

1. Thành phần nhận dạng mô hình

Nhận dạng mô hình là khâu đầu tiên trong quá trình mô hình hóa ARIMA. Chúng ta cần phải xác định các tham số p, d, q và P, Q, D, S của mô hình ARIMA. Những giá trị trị này được suy ra bằng cách trích rút những hệ số tương quan quan trọng trong chuỗi dữ liệu quan sát được.

Kĩ thuật truyền thống dùng để nhận dạng mô hình là quan sát đồ thị và trích rút các hệ số từ đồ thị. Tuy nhiên, phương pháp này có nhược điểm lớn là phụ thuộc vào độ hiểu biết của người dùng phần mềm. Do đó cần phải tìm ra phương pháp khác, sử dụng các công thức toán học để trích rút ra các hệ số.

* 1. Phân tích hệ số tự tương quan
     1. Hệ số tự tương quan

với ,

Dấu và độ lớn của hệ số tự tương quan xác định hướng và mức độ liên quan giữa các cặp giá trị quan sát. Ví dụ như hệ số tự tương quan cho biết rằng khi một giá trị tại thời điểm t tăng thì giá trị tại thời điểm t+k sẽ giảm và ngược lại, còn độ lớn của hệ số tương quan cho biết mức độ thay đổi tại thời điểm t+k phụ thuộc nhiều hay ít tại thời điểm t.

* + 1. Hệ số tự tương quan riêng phần

Bởi vì hệ số tương quan dùng để tính độ tương quan giữa và chịu ảnh hưởng bởi các giá trị . Do đó, ta cần phân tích thêm hệ số tự tương quan riêng phần . được ước lượng bằng phương pháp tối thiểu lỗi thông qua công thức:

Tuy nhiên phương pháp này rất tốn chi phí tính toán cho mỗi hệ số . Durbin đã đưa ra công thức xấp xỉ hệ số hồi quy thông qua hệ số tương quan và các hệ số đã tính trước đó.

* + 1. Tính chất mô hình trong ACFs và PACFs

Mô hình ARIMA sử dụng tính chất của ACFs và PACFs để kết hợp một một chuỗi dữ liệu với một mô hình thích hợp: mô hình tự hồi quy AR(p), mô hình trung bình di động MA(q), mô hình kết hợp tự hồi quy và trug bình di động ARMA(p,q). Hầu hết các hàm tương quan được chia ra làm 3 loại: giảm gián đoạn (abrupt decay), giảm theo hình sin hoặc hàm mũ, giảm rất chậm. Tùy theo tính chất của ACFs và PACFs mà chuỗi dữ liệu được kết hợp với 1 mô hình khác nhau

* Mô hình tự hồi quy AR(p): quá trình có ACFs với các độ trễ giảm chậm về không theo hình sin hoặc theo hàm mũ, trong khi đó PACFs gián đoạn về không sau một vài độ trễ.
* Mô hình trung bình di động MA(q): quá trình có tính chất ngược lại với mô hình AR(p), ACFs gián đoạn về không sau một vài độ trễ, trong khi đó PACFs với các độ trễ giảm về không theo hình sin hoặc theo hàm mũ.
* Mô hình kết hợp tự hồi quy và trung bình di động ARMA(p,q): quá trình có ACFs và PACFs giảm về không theo hình sin hoặc theo hàm mũ.
* Nếu ACFs giảm chậm thì chuỗi dữ liệu không tĩnh, có tính xu hướng cần phải được chuyển đổi trước khi ước lượng.

|  |  |  |
| --- | --- | --- |
| Process | ACF | PACF |
| AR(p) | Exponential or sinusoidal decay to zero | Spikes cut off to zero after lag p |
| MA(q) | Spikes cut off to zero after lag q | Exponential or sinusoidal decay to zero |
| ARMA(p,q) | Exponential or sinusoidal decay to zero | Exponential or sinusoidal decay to zero |

* + 1. Kĩ thuật nhận dạng mô hình

Quá trình nhận dạng mô hình ARIMA bao gồm 4 giai đoạn chính

* Xác định bậc lấy hiệu d để loại bỏ tính xu hướng trong chuỗi dữ liệu. Nếu ACF của chuỗi dữ liệu mang tính chất “giảm chậm” thì chuỗi dữ liệu được lấy hiệu cho đến khi ACF không mang tính chất “giảm chậm”. Sau khi lấy hiệu d lần, chuỗi dữ liệu đã mất tính xu hướng. Tuy nhiên nếu chuỗi dữ liệu có tính mùa, thì ACF rất lớn tại các cận của mùa. Khoảng cách giữa các cận chính là chiều dài mùa S.
* Loại bỏ tính mùa: chuỗi dữ liệu được lấy hiệu D lần với khoảng lấy hiệu là S để loại bỏ tính mùa.
* Xác định mô hình cho thành phần mùa: trong chuỗi dữ liệu được biến đổi, ta phân tích ACF và PACF tại các cận mùa để xác định P, Q.
* Xác định mô hình cho thành phần thường: trong chuỗi dữ liệu được biến đổi, ta phân tích ACF và PACF trong khoảng thời gian 1 mùa để xác định p, q.

Kết hợp 4 giai đoạn trên, ta được mô hình ARIMA

* 1. Tự động hóa phân tích hệ số tương quan

Dựa vào tính chất của ACFs và PACFs trong mỗi mô hình ARIMA, ta có thể xây dựng phương pháp để tự động nhận dạng ra mô hình:

* Nhận dạng các hệ số tương quan có ý nghĩa trong ACFs và PACFs sử dụng khoảng tin cậy thống kê.
* Trích xuất mô hình từ những hệ số tương quan đó
  + 1. Nhận dạng hệ số tương quan có ý nghĩa

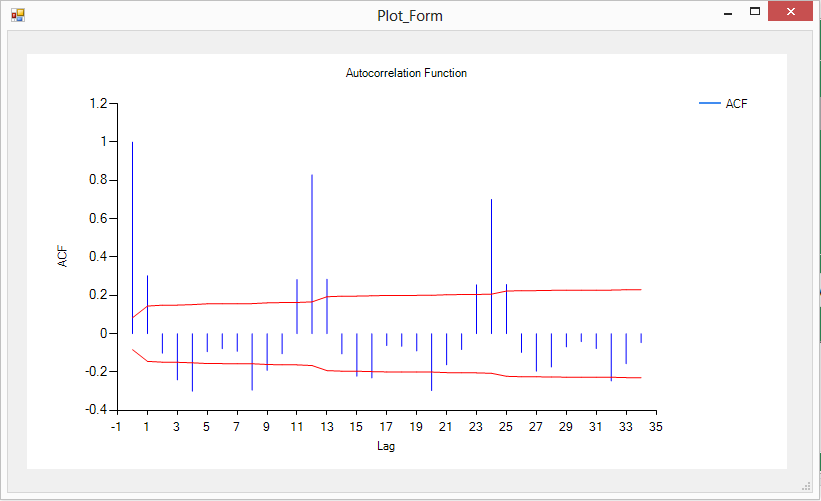
Trong các quá trình, các giá trị gần nhau sẽ liên kết với nhau hơn những giá trị xa nhau. Do đó, trong các quá trình tĩnh hệ số tự tương quan và tự tương quan riêng phần giảm về không tại những độ trễ rất lớn. Barlett đã đưa ra công thức để xấp xỉ độ lệch chuẩn của ACF và PACFs:

Từ đó, ta có thể tự động chọn những hệ số tương quan có ý nghĩa bằng cách chọn những giá trị vượt qua khoản tin cậy.

* + 1. Tự động xác định chu kì mùa

Đối với những quá trình có tình mùa, các giá trị quan sát không chỉ phụ thuộc vào những giá trị lân cận mà còn phụ thuộc vào những giá trị cách nó một bội số của chiều dài mùa. Do vậy để xác định chu kì mùa, ta cần:

* Xác định các hệ số tương quan tối ưu: các hệ số này cao hơn các hệ số lân cận và là giá trị cao nhất cục bộ.
* Tính khoảng cách giữa các hệ số tương quan tối ưu: sự sai khác về độ trễ giữa các hệ số liên tiếp
* Nhận dạng khoảng cách được lặp lại nhiều nhất, nếu tần suất của khoảng cách này lớn thì đó chính là chu kì mùa.



* + 1. Tự động phân tích hệ số tự tương quan

Hầu hết các ACFs và PACFs đều có mô hình thuộc một trong ba dạng sau:

* Giảm gián đoạn trong các quá trình không tĩnh
* Giảm theo hình sin hoặc hàm mũ trong quá trình tĩnh
* Giảm nhanh trong quá trình tĩnh
* Kiểm tra độ trễ

Nếu các hệ số tương quan chuẩn, không bị nhiễu và đơn điệu giảm thì ta có thể xác định loại mô hình dựa vào vị trí của độ trễ có ý nghĩa cuối cùng.

If( hight frequencies are monotonically decreasing)

{

If ( hight frequencies stay within MIN\_LAGS) then

Pattern = abrupt cut-off

Else if (hight frequencies extend beyond MAX\_LAGS) then

Pattern = slow decay

Else Pattern = exponential decay

}

trong đó MIN\_LAGS thường được chọn là 2, MAX\_LAGS là 10 dựa vào 5 mô hình phổ biến nhất: AR(1), AR(2), MA(1), MA(2), ARMA(1,1).

Tuy nhiên, các hệ số nhiễu thường xuyên xuất hiện trong ACFs và PACFs vì chịu ảnh hưởng của sai số mẫu hoặc các hệ số lớn lân cận. Do đó, ta sẽ dựa vào tỉ lệ thay đổi trung bình để xác định mô hình suy giảm trong ACFs và PACFs.

* Kiểm tra tỉ lệ thay đổi trung bình

Do những nhược điểm trên, nên phương pháp kiểm tra độ trễ thường được thay thế bằng phương pháp kiểm tra tỉ lệ thay đổi trung bình của độ lớn các hệ số tương quan có ý nghĩa:

với M là số hệ số tương quan có ý nghĩa

Phương pháp xác định mô hình trong hệ số tương quan:

If( > 65%) then pattern = abrupt cutoff

Else if ( < 10%) then pattern =slow decay

Else pattern = exponential decay

* + 1. Tự động xác định mô hình

Sau khi xác định tính chất của hệ số tương quan ACF và PACF, ta xác định mô hình ARIMA phù hợp theo giải thuật sau:

if (high frequencies exist in ACF or PACF)

{

if (ACF pattern = abrupt cut-off at lag q) then

Model = pure MA(q)

else if (PACF pattern = abrupt cut-off at lag p) then

Model = pure AR(p)

else if (both ACF and PACF pattern = exponentila decay) then

Model = ARMA(p,q)

}

else

{

p = q =0;  
}

1. Thành phần lấy hiệu/tổng hợp

Trong thực tế, các chuỗi dữ liệu rất phức tạp và thường mang tính xu hướng hoặc tính mùa. Do đó, các chuỗi dữ liệu cần được biến đổi thành các quá trình tĩnh trước khi đưa vào thành phần nhận dạng mô hình. Lấy hiệu là kĩ thuật được sử dụng phổ biến nhất.

* 1. Khái niệm cơ bản về lấy hiệu

Lấy hiệu là kĩ thuật biến đổi nhằm làm tĩnh giá trị trung bình của quá trình không tĩnh và mang tính nùa. Ở đây có hai loại lấy hiệu tùy thuộc vào khoảng lấy hiệu.

* Lấy hiệu thông thường: kĩ thuật này dùng cho quá trình không tĩnh và không có tính mùa. Khoảng lấy hiệu là một, tức là những giá trị liên tiếp nhau bị trừ để tạo ra chuỗi dữ liệu mới: . Ví dụ với chuỗi dữ liệu không tĩnh sau: 70.3, 100.5, 103.2, 160.7, 189.5, 220.2, 250.4

100.5 130.2 160.7 189.5 220.2 250.4

- 70.3 100.5 130.2 160.7 189.5 220.2

30.2 29.7 30.5 28.8 30.7 30.2

Sau khi lấy hiệu trên chuỗi dữ liệu, ta thu được một chuỗi dữ liệu tĩnh mới dao động quanh giá trị trung bình 30.

* Lấy hiệu theo mùa: kĩ thuật này dùng cho chuỗi dữ liệu có tính mùa. Khoảng lấy hiệu là một giá trị lớn hơn một S, chiều dài của mùa. Mỗi giá trị trong chuỗi dữ liệu chuyển đổi chính là sự sai khác giữa các giá trị cách nhau S giai đoạn trong chuỗi dữ liệu cũ: . Ví dụ với chuỗi dữ liệu có tình mùa sau: 3.2, 6.3, 9.5, 12.1, 50.0, 6.3, 9.7, 12.3, 15.1, 53.1, 9.4, 12.5, 15.5, 18.2, 56.4

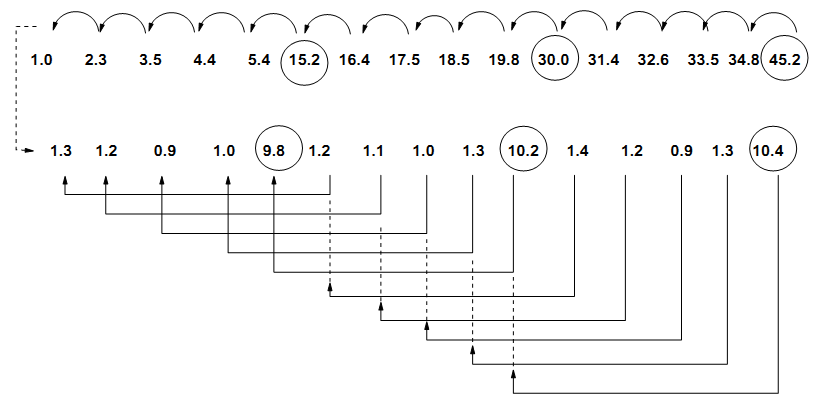
6.3 9.7 12.3 15.1 53.1 9.4 12.5 15.5 18.2 56.4

- 3.2 6.3 9.5 12.1 50.0 6.3 9.7 12.3 15.1 53.1

3.1 3.4 2.8 3.0 3.1 3.1 2.8 3.2 3.1 3.3

Sau khi lấy hiệu với khoảng lấy hiệu là 5, chiều dài của mùa thì ta được một chuỗi dữ liệu tĩnh dao động xung quanh giá trị trung bình 3.

Với một số chuỗi dữ liệu chỉ với một lần lấy hiệu ta có thể lấy được chuỗi dữ liệu tĩnh. Nhưng đối với một số chuỗi dữ liệu, ta phải lấy liệu nhiều lần và kết hợp cả hai phương pháp mới có được chuỗi dữ liệu tĩnh.



Lấy hiệu thông thường

Lấy hiệu theo mùa S=5

1. Thành phần ước lượng tham số

Trong phần này chúng tôi sẽ trình bày cấu trúc và thành phần của mô hình ARMA(p,q) đối với dữ liệu tĩnh. Chúng tôi sẽ phát triển mô hình từ hình thức đơn giản rồi mở rộng, cải tiến mô hình.

* 1. Ước lượng thông số cho mô hình AR(p)

Trong lý thuyết thống kê hệ thống tuyến tính, phương pháp bình phương tối thiểu truyền thống được dùng để ước lượng tham số cho chuỗi dữ liệu tỉnh chỉ với thành phần tự hồi quy. Với mô hình như sau:

Các bước để ước lượng tham số cho mô hình

* Khởi tạo giá trị mặc định cho các hệ số và tập hợp các giá trị quan sát trong quá khứ . Giá trị dự đoán của mô hình tại thời điểm kế tiếp (t) được tính dựa trên tổ hợp tuyến tính của p giá trị quan sát.
* Tại thời điểm kế tiếp t, sự khác biệt giữa giá trị quan sát và giá trị được dự đoán từ mô hình. Sự khác biệt gọi là lỗi trước dự đoán (priori prediction eror), chính là thành phần nhiễu không dự đoán được.
* Tính toán lỗi cho p+1 giai đoạn sẽ được công thức tương ứng với việc ước lượng p+1 tham số chưa biết .

trong đó:

* là véc tơ giá trị quan sát, nó bao gồm giá trị hiện tại y(t) và p giá trị trước đó.
* là ma trận giá trị quan sát quá khứ, mỗi hàng của ma trận bao gồm hằng số 1 và p giá trị quan sát trước đó .
* là véc tơ tham số, nó bao gồm các tham số cần được ước lượng a0, a1,…,ap.
* là véc tơ lỗi dự đoán, nó bao gồm các giá trị khác biệt giữa véc tơ giá trị quan sát hiện tại và véc tơ giá trị ước lượng .
* Sử dụng phương pháp tối thiểu bình phương để ước lượng tham số cho mô hình

Việc tối thiểu lỗi có thể giải quyết bằng cách gán các đạo hàm riêng phần tương ứng mỗi tham số bằng 0, cuối cùng ta được công thức:

* 1. Ước lượng hồi quy tham số cho mô hình AR(p)

Phương pháp bình phương tối thiểu truyền thống rất đơn giản, tuy nhiên nó có các nhược điểm sau:

* Nghịch đảo ma trận cần phải tính toán cho mỗi vòng ước lượng tham số. Chi phí cho việc tính toán này rất lớn, đặc biệt khi ma trận có kích thước lớn.
* Toàn bộ dữ liệu quan sát được sử dụng cùng lúc khi ước lượng tham số. Do đó với dữ liệu lớn, ta sẽ gặp phải vấn đề lưu trữ và xử lý dữ liệu lớn, chi phí sẽ lớn.

Do đó để đạt hiệu quả cho quá trình ước lượng, ta cần kết hợp giải thuật hồi quy kết hợp với phương pháp bình phương tối thiểu:

* Đệ quy nâng cao khả năng sử dụng các tham số được ước lượng ở giai đoạn trước, hạn chế việc lưu trữ và xử lý trên toàn bộ dữ liệu. Sử dụng đệ quy, ta có thể sử dụng và điều chỉnh các tham số đã được ước lượng.
* Tránh nghịch đảo ma trận: áp dụng bổ đề nghịch đảo ma trận kết hợp với đệ quy, cho phép dẫn suất trực tiếp ma trận nghịch đảo từ lần ước lượng ở giai đoạn trước.

Tuy nhiên mô hình này cần được mở rộng để áp dụng cho mô hình ARMA(p,q).

* 1. Ước lượng mở rộng hồi quy tham số cho mô hình AR(p)

Đối với các chuỗi dữ liệu có chứa nhiễu thì mô hình AR(p) không thể ước lượng chính xác. Do đó phương pháp bình phương tối thiểu cần được mở rộng (ELS) để áp dụng cho mô hình MA(q), ARMA(p,q). ELS khai thác kĩ thuật tự khởi động (bootstrapping) để tính toán ước lượng cho thành phần nhiễu e(t-1), e(t-2).

* + 1. Lỗi sau dự đoán

Trong chuỗi dữ liệu, nhiễu đại diện cho các thành phần của dữ liệu không thể giải thích được bởi cấu trúc chính của chương trình. ELS dùng lỗi sau dự đoán như là ước lượng cho lỗi. Khác với lỗi trước dự đoán, lỗi sau dự đoán sử dụng các tham số gần nhất thay vì các tham số được ước lượng trước đó.

Lỗi sau dự đoán được tính dựa trên sự khác biệt giữa giá trị quan sả hiện tại và giá trị dự đoán được.

là chuyển vị của véc tơ giá trị quan sát tại giai đoạn t-1, nó bao gồm giá trị quan sát và giá trị nhiễu ước lượng được với hằng số 1 tương ứng với tham số

* + 1. Hiện thực mô hình

Quá trình ước lượng tham số gồm 3 giai đoạn

SET INITIAL CONDITIONS

ESTIMATE PARAME

TERS

PREPARE FOR NEXT ESTIMATION

Compute\_AprioriPredictionError Compute\_GainFactor 

Set\_DefaultInitialConditions

Compute\_AposteriorPredictionError Update\_ArpartOfObservationVector Update\_MapartOfObservationVector Update\_RinverseMatrix

* Giai đoạn 1: khởi tạo điều kiện cho quá trình hồi quy. Véc tơ tham số được khởi tạo bằng 0, . Nghịch đảo của ma trận phướng sai với giá trị .
* Giai đoạn 2: giá trị mới của véc tơ tham số được ước lượng dựa vào lỗi trước dự đoán và hệ số khuyếch đại.

Hệ số khuyếch đại được tính hồi quy dựa vào giá trị cũ, véc tơ quan sát và nghịch đảo ma trận phương sai .

Gain factor =

* Giai đoạn 3: cập nhật giá trị cho các biến liên quan để chuẩn bị cho vòng tiếp theo.

Giá trị lỗi sau dự đoán được tính toán dựa vào giá trị quan sát hiện tại và véc tơ tham số mới nhất

Nghịch đảo ma trận phương sai được tính dựa vào giá trị cũ và hệ số khuếch đại

Véc tơ giá trị quan sát và lỗi được cập nhật lại.

Giai đoạn 2 và 3 được lặp lại cho mỗi mẫu dữ liệu quan sát, qua mỗi lần ước lượng như vậy véc tơ tham số sẽ được cập nhật chính xác hơn.

* 1. Mở rộng mô hình ARIMA cho dữ liệu có tính màu

Để ước lượng mô hình cho chuỗi dữ liệu có tính mùa, ta bổ sung vào mô hình ARIMA các tham số P, D, Q, S. Do đó mô hình

Tương tự như với ước lượng cho mô hình AR(p), ta được các tham số cho mô hình ARIMA(p,d,q)(P,D,Q)S.

1. Thành phần dự đoán

Sau khi xác định được mô hình dự đoán và ước lượng tham số cho mô hình, ta sẽ dùng kết quả đó để dự đoán cho cái thời đoạn tiếp theo.

* 1. Dự đoán 1 thời đoạn cho chuỗi dữ liệu tĩnh

Theo phương pháp tối thiểu bình phương lỗi, giá trị dự đoán của mô hình chính là giá trị của mô hình tại thời điểm tiếp theo (t+1)

Bời vì tất cả các tham số trong mô hình ngoại trừ , thành phần lỗi Gausian được giả thuyết là phân phối tự nhiên , chúng ta có thể ước lượng nó với giá trị kỳ vọng 0.

Do đó

* 1. Dự đoán nhiều thời đoạn cho chuỗi dữ liệu tĩnh

Sau khi dự đoán cho 1 giai đoạn, ta có thể tự khởi động quá trình dự đoán n giai đoạn dựa vào các giá trị đã tính toán được, các

* 1. Dự đoán nhiều thời đoạn cho chuỗi dữ liệu không tĩnh hoặc có tính mùa
* Tính toán n giai đoạn dự đoạn cho chuỗi dữ liệu đã được chuyển đổi
* Khôi phục lại chuỗi dữ liệu gốc: thực hiện ngược lại quá trình lấy hiệu đã thực hiện khi khử đi tính không tĩnh và tính mùa của chuỗi dữ liệu gốc.