ĐẠI HỌC QUỐC GIA TP HCM

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Đặng Kiên Cường

PHÂN TÍCH DỮ LIỆU CHUỖI THỜI GIAN TRONG CÁC BÀI TOÁN ĐÁNH GIÁ VÀ DỰ BÁO

(HỆ THỐNG HỖ TRỢ HỌC TẬP THÍCH NGHI
DỰA TRÊN ONTOLOGY CỦA MÔ HÌNH NGƯỜI HỌC)

Chuyên ngành: Khoa học máy tính
Mã số: 62.48.01.01

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

TP HỒ CHÍ MINH - NĂM 2020

Công trình được hoàn thành tại: Trường Đại học Công nghệ thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh

Người hướng dẫn khoa học:

1. TS. Trần Tích Phước
2. TS. Dương Tôn Đảm

Phản biện 1: ....
....

Phản biện 2: ....
....

Phản biện 3: ....
....

Luận án sẽ/đã được bảo vệ trước
Hội đồng chấm luận án cấp Trường tại:
....
....
vào lúc   giờ   ngày   tháng   năm

Có thể tìm hiểu luận án tại:

- Thư viện Quốc gia Việt Nam
- Thư viện Trường Đại học Công nghệ Thông tin

DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ LIÊN QUAN ĐẾN ĐỀ TÀI LUẬN ÁN

1. **Dang Kien Cuong**, Duong Ton Dam, Duong Ton Thai Duong, and Nguyen Kim Loi, Nguyen Son Vo, and Ayse Kortun, *"Extreme Value Distributions In Hydrological Analysis In The Mekong Delta: Case Study In Ca Mau, An Giang Provinces"*, **EAI Endorsed Transactions on Industrial Networks and Intelligent Systems Journal**, ISSN: 2410-0218. Vol 6, June 2019, http://dx.doi.org/10.4108/eai.13-6-2019.159122.

2. **Dang Kien Cuong**, Duong Ton Dam, Duong Ton Thai Duong, and Du Thuan Ngo, *"Applications of Bootstrap in Analyze General Extreme Value Distributions"*, **Journal of Mechanics Engineering and Automation**, ISSN: 2159-5275 Vol. 9, No. 7, 2019.

3. **Dang Kien Cuong**, Duong Ton Dam, Duong Ton Thai Duong, Du Thuan Ngo, *"Solutions to the jump-diffusion linear stochastic differential equations"*, **Science And Technology Development Journal**, Vol 3 No 2. 2019. Page: 115-119. DOI: https://doi.org/10.32508/stdjns.v3i2.663.

4. **Dang Kien Cuong**, Duong Ton Dam, and Duong Ton Thai Duong, *"Extreme value distributions in hydrological analysis of some areas in the Mekong Delta"*, Proceedings of the Second **Vietnam international Applied Mathematics** Conference (VIAMC 2017), Information and Communications Publishing House, ISBN: 978-604-80-0608-2.

MỞ ĐẦU

1. Tính cấp thiết của luận án

Trong thời kỳ phát triển công nghệ số, nhu cầu sử dụng dữ liệu ngày càng lớn, đặc biệt là liên quan công nghệ và xã hội. Với kinh tế tri thức, dữ liệu có vai trò quan trọng nhằm hỗ trợ ra quyết định, thực hiện dự báo cũng như đánh giá cho tương lai.

Dự báo trong thống kê là việc tiên đoán những giá trị của tương lai dựa vào số liệu của quá khứ, hiện tại và mô hình toán học phù hợp được thiết lập. Dự báo được đánh giá là bước quan trọng đầu tiên, không thể thiếu cho việc hoạch định các chính sách phát triển kinh tế xã hội phù hợp, cho các chiến lược kinh doanh hiệu quả của các tổ chức kinh tế, xã hội, các cấp chính quyền của các quốc gia. Chính vì thế, dự báo đã và đang được các nhà khoa học, đặc biệt là các nhà thống kê, máy học, khoa học máy tính quan tâm. Tuy nhiên, cho đến nay, việc dự báo và đánh giá dựa trên dữ liệu vẫn là bài toán chưa có lời giải cuối cùng.

Trong dự báo, đánh giá bằng phương pháp thống kê, hai mô hình chính đang được sử dụng rộng rãi là hồi quy và chuỗi thời gian. Khi xây dựng mô hình hồi quy phải giả sử nhiều điều kiện mà trong thực tế các dữ liệu thường không thỏa vì vậy kết quả dự báo có hạn chế. Trong khi đó, điều kiện quan trọng để xây dựng mô hình chuỗi thời gian là tính dừng của dữ liệu. Điều này có thể khắc phục được qua phương pháp lấy sai phân, nên mô hình chuỗi thời gian thường phù hợp với nhiều số liệu thực tế, được đánh giá có ưu điểm hơn mô hình hồi quy. Trong thực tế có rất nhiều số liệu về phát triển kinh tế xã hội được lưu trữ dưới dạng chuỗi thời gian.

Vấn đề dữ liệu chuỗi thời gian, trong việc quản lý, dự báo thiên tai (Khí tượng, Thủy văn), liên quan đến dữ liệu lớn (trên 30 năm), bài toán dữ liệu thiếu, khuyết trong quá trình quan trắc. Trong những năm gần đây vấn đề thiên tai xảy ra với cường độ và tần suất lớn, và trong quản lý khí tượng thủy văn chưa có các nghiên cứu liên quan để giải quyết vấn đề trên.

Luận án đã và đang giải quyết các bài toán về vấn đề khí tượng thủy văn, nghiên cứu đánh giá và dự báo dựa trên dữ liệu chuỗi thời gian, tìm quy luật và đặc tính của tập dữ liệu bằng các phương pháp toán học.

2. Mục tiêu của luận án

Đối tượng nghiên cứu: Tập dữ liệu là các giá trị ghi nhận được về những hiện tượng ngẫu nhiên trong thực tế theo dòng thời gian (chuỗi thời gian), với dạng chung của dữ liệu phụ thuộc là giá trị của hàm ngẫu nhiên $\mathrm{f}(\mathrm{t},\omega)$. Hàm ngẫu nhiên $\mathrm{f}(\mathrm{t},\omega)$: $R \times \Omega \to R$ là dạng tổng quát của 2 dạng hàm:

Hàm số thực trong toán giải tích $\mathrm{f}(\mathrm{x})\colon \mathrm{R} \to \mathrm{R}$

Biến ngẫu nhiên trong xác suất $\mathrm{f}(\omega)\colon \Omega \to \mathrm{R}$

Luận án nghiên cứu dữ liệu chuỗi thời gian, dựa trên dữ liệu khí hậu thủy văn khu vực Đồng bằng Sông Cửu Long giai đoạn từ 1975 đến 2016, so sánh với dữ liệu toàn cầu; nghiên cứu thực hiện khảo sát mô hình chuỗi thời gian mờ, cải tiến việc chọn các tham số, mối quan hệ mờ và vấn đề tính toán để nâng cao hiệu quả trong áp dụng; áp dụng mô hình cải tiến trong một số dự báo liên quan đến Đồng bằng Sông Cửu long, để xuất thuật toán, tiêu chuẩn đánh giá và vấn đề tính toán của mô hình chuỗi thời gian.

Công cụ xử lý: Với dữ liệu thực tế, công cụ toán để xử lý phải phù hợp và mở rộng nhiều so với các công cụ kinh điển (trong giải tích ngẫu nhiên có nhiều hàm không đâu có đạo hàm và vi phân) tích phân cũng được hiểu theo một nghĩa khác (tích phân Itô, tích phân Sugeno). Vì vậy, công cụ chính là các phép tính vi-tích phân ngẫu nhiên với các phương pháp Toán hiện đại: *(1) Toán mờ* (Tương quan, hồi quy mờ, phân tích mờ và giải mờ), *(2) Thống kê bootstrap* (jackknife, bootstrap khối, bootstrap dừng), và *(3) Lý thuyết về quá trình khuếch tán ngẫu nhiên có nhảy*.

Luận án ứng dụng phương pháp Bootstrap trong phân tích chuỗi dữ liệu thời gian trong lĩnh vực khí tượng thủy văn. Với các mục tiêu cụ thể:

1) Nghiên cứu, Phân tích dữ liệu chuỗi thời gian và dự báo chuỗi thời gian mờ;

2) Đánh giá dữ liệu khí tượng thủy văn giai đoạn 1986 – 2015;

3) Ứng dụng trong nghiên cứu đánh giá dự báo biến đổi khí hậu tại Đồng Bằng Sông Cửu Long, từ 2018 đến 2022.

3. Đóng góp của luận án

Luận án đã đánh giá một cách chi tiết về đặc tính cực trị các phân phối xác suất, về khả năng ứng dụng bootstrap trong phân tích chuỗi thời gian khí tượng thuỷ văn, và các phương pháp giải phương trình vi phân ngẫu nhiên tuyến tính tổng quát trong trường hợp khuếch tán có nhảy. Cụ thể, luận án đã đóng góp về lý thuyết và khả năng ứng dụng:

a) Phân tích về các đặc tính của cực trị các phân phối xác suất

Lý thuyết: Đưa ra Định lý về Giá trị cực đại*:* Định lý về tổng các kỳ vọng và tổng các phương sai của các phân phối cực trị (EVD – Extreme Value Distributions).

Ứng dụng: Nghiên cứu Phân phối cực trị trong một số mô hình thủy văn tại Tây Nam Bộ, từ đó thực hiện các công việc

- ✓ Tính cực đại của mực nước tại sông Tiền qua Tân Châu (An Giang) thông qua dữ liệu thực tế từ 1975 đến 2017.
- ✓ Tính cực đại của độ mặn và lượng mưa qua Thành phố Cà Mau thông qua số liệu thực tế từ 1990 đến 2017.
- ✓ Dự báo về mực nước tại sông Tiền và sông Hậu qua Tân Châu (An Giang) thông qua phân tích mờ các dữ liệu thủy văn từ 2018 đến 2022.

b) Phân tích một cách có hệ thống về việc áp dụng thống kê Bootstrap cho dữ liệu về chuỗi thời gian

Lý thuyết: tổng quan về phương pháp Bootstrap, là một hướng tiếp cận mới của lý thuyết Thống kê về xử lý dữ liệu. Luận án đã phân tích các loại Bootstrap khối (MBB, NBB, SBB, SB) và liên hệ giữa các phương pháp này trong xử lý dữ liệu, đặc biệt là trong mô hình tuyến tính về chuỗi thời gian.

Ứng dụng: đã đưa ra các thuật giải cho những dạng Bootstrap khác nhau, tương thích với các mô hình tuyến tính cụ thể trên số liệu thực tế về dòng chảy, lượng mưa và độ mặn tại Cà Mau và An Giang để minh chứng cho các phương pháp lý thuyết.

c) Đưa ra phương pháp giải phương trình vi phân ngẫu nhiên tuyến tính tổng quát trong trường hợp khuếch tán có nhảy.

Lý thuyết: Phương trình vi phân ngẫu nhiên thường được xét đến trong thực tế là dạng tuyến tính, luận án đã tìm cách giải dạng PTVP này. Đóng góp chính ở phần này là phương pháp tách nhiệm dựa vào nghiệm của phương trình vi phân tuyến tính thuần nhất.

Hướng áp dụng: giải bài toán của dữ liệu thủy văn (dòng khếch tán có nhảy), được xét cùng với các biến động (kể cả dị thường: lũ, bão, ngăn đập, vỡ đê) và các yếu tố có liên quan đến con người (ngăn đập, mở đập).

4. Bố cục của luận án

Sau phần mở đầu đã nêu các vấn đề tổng quan về luận án, nội dung chính được trình bày gồm 4 chương theo cấu trúc sau:

Chương 1, cơ sở toán học, các phương pháp và các kết quả nghiên cứu lý thuyết theo hướng kinh điển về chuỗi thời gian.

Chương 2, trình bày các hướng mới về chuỗi thời gian mờ cùng các phân tích khoa học mang tính dự báo một số vấn đề về dữ liệu thủy văn tại ĐBSCL.

Chương 3, đưa ra một cách tiếp cận hiện đại và hiệu quả về thống kê bootstrap cho các dạng dữ liệu hiếm và khó thu thập trong thực tế cùng những đánh giá có giá trị về việc xử dụng chúng.

Chương 4, phân tích các quá trình ngẫu nhiên, giải phương trình vi phân ngẫu nhiên liên tục và có nhảy. Đây là các dữ liệu có liên quan đến những loại biến động, rất gần với chuỗi thời gian. Kết thúc bằng việc giải phương trình vi phân ngẫu nhiên tuyến tính thuần nhất, giải phương trình vi phân ngẫu nhiên tuyến tính tổng quát với phương pháp tách nghiệm đề xuất, để tiếp tục giải bài toán của dữ liệu thủy văn xét cùng với các biến động ngẫu nhiên.

Phần cuối luận án đã chỉ ra được những hướng nghiên cứu khả thi về mặt lý thuyết và những vấn đề có thể áp dụng được trong thực tế từ các kết quả thu được.

Về nội dung nghiên cứu cụ thể xin được trình bày như sau

Chương 1 CÁC PHƯƠNG PHÁP TOÁN THỐNG KÊ KINH ĐIỂN TRONG XỬ LÝ DỮ LIỆU CHUỖI THỜI GIAN

1.1 Khái niệm cơ bản về dữ liệu chuỗi thời gian

Chuỗi thời gian là tập hợp gồm các số liệu có cùng khái niệm và phạm vi được thu thập liên tục và thường kỳ. Các giá trị quan sát theo thời gian của đại lượng Y được ký hiệu là Y_1, Y_2, …, Y_t, …, Y_n, với Y_t là giá trị quan sát của Y tại thời điểm t.

Căn cứ vào đặc điểm thời gian, thường chia dãy số thời gian thành 2 loại: (i) Dãy số thời kỳ biểu hiện sự thay đổi của hiện tượng qua từng thời kỳ nhất định, (ii) Dãy số thời điểm biểu hiện mặt lượng của hiện tượng vào một thời điểm cụ thể.

1.2 Thành phần của dữ liệu chuỗi thời gian, gồm Chu kỳ (Period - P_t), Mùa (Seasonal-S_t), Xu hướng (Trend-T_t), Bất thường (Irregular-I_t), có thể được kết hợp nhiều cách, theo dạng thức: $Y_t=\alpha_1\boldsymbol{T_t^{\beta_1}}\oplus\alpha_2\boldsymbol{S_t^{\beta_2}}\oplus\alpha_3\boldsymbol{C_t^{\beta_3}}\oplus\alpha_4\boldsymbol{I_t^{\beta_4}}$, trong đó, α_i là trọng số (i=1,2,3,4), β_j là hệ số biến đổi mũ (j=1,2,3,4).

1.3 Đặc tính của dữ liệu chuỗi thời gian

Tính dừng của chuỗi thời gian được thể hiện trong các suy luận về quá khứ hoặc tương lai của quan sát, với 03 đặc trưng

1) Kỳ vọng $E(X_t) = \mu$ (thường là hằng số hữu hạn),

2) Phương sai $Var\ (X_t) \leq \infty$,

3) Hiệp phương sai $Cov\ (X_t, X_{t+s}) = \gamma_s$

Hàm tự tương quan, chuỗi thời gian thường xét qua hàm tự tương quan ACF (A*utoCorrelation Function*), không phụ thuộc vào thời gian trong chuỗi mà phụ thuộc vào khoảng thời gian giữa các quan sát, ký hiệu là *l* và gọi là độ trễ (lag), đơn giản hơn *l* là trễ.

Quá trình dừng mạnh (dừng theo nghĩa hẹp), tính dừng có nghĩa là quá trình đạt đến một loại cân bằng thống kê và phân phối của quá trình không thay đổi nhiều, và **Qúa trình dừng yếu** (dừng theo nghĩa rộng).

Toán tử lùi, toán tử tiến

Toán tử lùi L liên kết với quá trình $\{X_t, t \in Z\}$ là quá trình $\{Y_t, t \in Z\}$ sao cho $Y_t = LX_t = X_{t-1.}$

Nếu L là toán tử tuyến tính, khả nghịch thì toán tử nghịch đảo $L^{-1} = T$ gọi là toán tử tiến T, được định nghĩa bởi: $TX_t = X_{t+1.}$

Trong một số chuỗi thời gian, thành phần mùa và thành phần bất thường thay đổi lớn làm cho việc xác định thành phần xu hướng và thành phần chu kỳ gặp nhiều khó khăn. Dùng phương pháp làm trơn dữ liệu để làm giảm sự thay đổi lớn.

1.4 Các phương pháp làm trơn dữ liệu

Phương pháp trung bình trượt, có Trung bình trượt đơn,Trung bình trượt kép, Trung bình trượt trung tâm; Phương pháp hàm mũ, có Hàm mũ đơn, Hàm mũ kép, Hiệu quả nhất là mô hình Box-Jenkins.

1.5 Phân tích các mô hình thống kê về dữ liệu chuỗi thời gian thông dụng, cụ thể: Mô hình tự hồi quy (AutoRegressive - AR), Mô hình trung bình trượt (Moving Averages - MA), Mô hình tự hồi quy trung bình trượt (Autoregressive Moving Average - ARMA), Mô hình tự hồi quy tích hợp trung bình trượt (Auto Regressive Integrated Moving Average - ARIMA), Phương pháp phân tích mô hình thống kê Box-Jenkins.

1.6 Tiêu chuẩn đánh giá mô hình

Một mô hình tốt phải là một mô hình có khả năng dự báo với độ chính xác cao. Để đánh giá mức độ chính xác về dự báo của mô hình đòi hỏi phải có các số liệu theo hai hướng: hoặc đưa thêm số liệu mới qua điều tra bổ sung để làm mẫu kiểm tra; hoặc phân chia mẫu hiện có thành hai mẫu con. Việc tách mẫu phải đảm bảo: không làm thay đổi nhiều đến kết quả dựa trên mẫu khởi động, đồng thời phải đủ số quan sát cho mẫu kiểm tra để đánh giá được khả năng dự báo của mô hình. Hai tiêu chuẩn được phân tích và sử dụng: Tiêu chuẩn thông tin Akaike (AIC) và Tiêu chuẩn thông tin Bayes (BIC) đươc sử dụng để lựa chọn mô hình trong các mô hình khác nhau và so sánh các mô hình dùng để dự báo

Ngoài các tiêu chuẩn trên, ta còn dùng số liệu của một số năm gần nhất để so sánh với số liệu dự đoán được. Nếu số liệu dự đoán từ một mô hình nào đó gần đúng với dữ liệu năm gần nhất so sánh thì chứng tỏ rằng mô hình đó có xu hướng phù hợp để có thể dự báo.

Chương 2. SỬ DỤNG TOÁN MỜ TRONG PHÂN TÍCH DỮ LIỆU VÀ DỰ BÁO CHUỖI THỜI GIAN

2.1. Cơ bản về chuỗi thời gian mờ

Chuỗi thời gian là một chuỗi các điểm dữ liệu, được đo theo từng khoảng khắc thời gian liền nhau theo một tần suất thời gian thống nhất. *Phân tích* chuỗi thời gian bao gồm các phương pháp để phân tích dữ liệu chuỗi thời gian, từ đó trích xuất ra được các thuộc tính thống kê có ý nghĩa và các đặc điểm của dữ liệu. *Dự báo* chuỗi thời gian là việc sử dụng mô hình để dự đoán các sự kiện thời gian dựa vào các sự kiện đã biết trong quá khứ để từ đó dự đoán các điểm dữ liệu trước khi nó xảy ra.

Chuỗi thời gian có thể được xem như một tập hợp dữ liệu được quan sát tại một thời điểm riêng biệt. Thông tin có thể được suy ra từ các mẫu của các quan sát trong quá khứ và có thể được sử dụng để dự báo các giá trị trong tương lai của chuỗi. Tuy nhiên dữ liệu có thể chưa chính xác và không đầy đủ, để giải quyết vấn đề này chúng ta giải quyết bằng phương pháp chuỗi thời gian mờ, dựa trên 4 khái niệm của giải tích mờ *1) Chuỗi thời gian mờ, 2) Quan hệ mờ, 3) Bậc của quan hệ mờ, 4) Bất biến thời gian của chuỗi thời gian mờ.*

2.2 Một số mô hình *được phân tích để sử dụng: Mô hình Abbasov-Mamedova ứng dụng chuỗi thời gian mờ để dự báo dữ liệu theo 6 bước, Mô hình của Chen và Hsu với 4 bước và 3 nguyên tắc, Mô hình của Chen và Hsu, gồm 5 bước và 3 quy tắc giải mờ, và một số mô hình khác như mô hình Heuristic, mô hình của Singh, mô hình Liu H. T, mô hình Saxena P., và S. Easo.*

2.3 Dự báo đỉnh mặn theo phương pháp toán mờ tại các trạm đo của tỉnh Cà Mau.

Ý tưởng bài toán: vấn đề xâm nhập mặn của tỉnh Cà Mau chủ yếu qua ba sông chính đổ ra biển: sông Gành Hào, sông Ông Đốc, và sông Cửa Lớn. Vì thế, mức độ mặn cũng như sự xâm nhập mặn bên trong các vùng tỉnh Cà Mau có thể xác định dựa vào kết quả phân tích độ mặn tại 3 trạm quan trắc này.

Từ bảng Số liệu đỉnh mặn tại các trạm đo Gành Hào, Cà Mau, Ông Đốc giai đoạn 2000-2017, cho thấy có sự biến động, trong đó giai đoạn 2009-2011 có sự thay đổi đáng kể.

Tóm tắt ý tưởng, chia dữ liệu (18 năm) thành hai phần: Tập huấn luyện và tập kiểm tra với tỉ lệ lần lượt là 80% (14 năm) và 20% (4 năm). Tập huấn luyện được sử dụng để xây dựng mô hình, tập kiểm tra được sử dụng để đánh giá các mô hình được xây dựng từ tập huấn luyện.

Thuật toán 1: Dự báo đỉnh mặn

Algorithm 1

Input: dữ liệu lần lượt của tập huấn luyện (80%), tập kiểm tra (20%)

Bắt đầu

1) Làm trơn
2) Mờ hóa với ARIMA, AM và IFTS
3) Tính các tham số

Kết thúc

Ouput: dữ liệu đã được xử lý, sử dụng cho việc dự báo, đánh giá.

Thực hiện tất cả các mô hình trên toàn bộ dữ liệu, lựa chọn mô hình tốt nhất dựa trên các tham số đánh giá. Dự báo đỉnh mặn cho các trạm đo đến năm 2022. Các tính toán của chương này, sử dụng gói AnalyzeTS với các hướng dẫn sử dụng trình bày trong Mô hình dự báo chuỗi thời gian và chi tiết các đoạn mã được trình bày trong phần Phụ lục.

Bảng 1. Dự báo đỉnh mặn tại các trạm đo Gành Hào, Cà Mau, Ông Đốc 2018 đến 2022

Năm	Gành Hào	Cà Mau	Ông Đốc
2018	34.85	36.60	39.62
2019	35.42	36.65	40.03
2020	36.06	36.70	40.40
2021	36.69	36.75	40.76
2022	37.29	36.80	41.11

Đánh giá kết quả dự báo đỉnh mặn khu vực ĐBSCL: Từ kết quả thực hiện cho dữ liệu đỉnh mặn tại 3 trạm đo Gành Hào, Cà Mau và Ông Đốc của tỉnh Cà Mau cho thấy đỉnh mặn của cả ba đều có khuynh hướng tăng trong thời gian sắp tới, trong đó đỉnh mặn tại trạm đo Gành Hào có khuynh hướng tăng nhiều nhất.

2.4. Phân tích dữ liệu trong bài toán về phân phối cực trị

Phân phối cực trị là một lớp phân phối quan trọng trong lý thuyết xác suất, được ứng dụng để mô hình hóa và giải các bài toán thủy văn như tính đỉnh lũ, đỉnh mặn, dòng chảy hạn; tính sự nghẽn mạch trong các dòng tín hiệu, hoặc trong điều tiết giao thông.

Bài toán cực đại và cực tiểu, Cho $X_1, X_2, \dots, X_n$, là dãy các đại lượng ngẫu nhiên độc lập, với cực đại $M_n = \max\{X_1, X_2, \dots, X_n\}$ và cực tiểu, $m_n = \min\{X_1, X_2, \dots, X_n\}$

Từ đó xét phân phối giới hạn cho biến động của cực đại và cực tiểu, thông qua khái niệm về miền hút và qua giới hạn của các hàm phân phối tích lũy tương ứng $F_{M_n}(x)$, $F_{m_n}(x)$.

Phân phối cực trị $G(x)$ sẽ là một trong ba loại: (i) Gumbel (phân phối mũ kép), (ii) Fréchet, (iii) Weibull.

Tiếp tục phát triển theo hướng nghiên cứu này ta thu được định lý khá lý thú về tổng các kỳ vọng và tổng các phương sai của các dạng phân phối giới hạn sau.

Định lý 1: Cho $\{\xi_I; i = 1,2, \dots\}$ là dãy các đại lượng ngẫu nhiên độc lập, thuộc miền hút max của $H_{\beta_i}(x, \lambda_I, \delta_i) \equiv H_I$ và $\{\eta_I; i = 1,2, \dots\}$, là dãy các đại lượng ngẫu nhiên độc lập, thuộc miền hút min của $L_{\beta_i}(x, \lambda_I, \delta_i) \equiv L_I$, khi đó

$$a)\ \sum_{i=1}^{n}(EH_i + EL_i) = 2\sum_{i=1}^{n}\lambda_i.$$

$b)\ \sum_{i=1}^{n}(VarH_i + VarL_i\ ) =$

$$\begin{cases} \dfrac{\pi^2}{3}\displaystyle\sum_{i=1}^{n}\delta_i^2, & \text{nếu } \mathrm{H_i}\sim\mathrm{MaxGD};\ \mathrm{L_i}\sim\mathrm{MinGD}, \\ 2\displaystyle\sum_{i=1}^{n}\delta_i^2\left[\Gamma\left(1+\dfrac{2}{\beta_i}\right)-\Gamma^2\left(1+\dfrac{1}{\beta_i}\right)\right], & \text{nếu}\mathrm{H_i}\sim\mathrm{MaxWD};\ \mathrm{L_i}\sim\mathrm{MinWD}, \\ 2\displaystyle\sum_{i=1}^{n}\delta_i^2\left[\Gamma\left(1-\dfrac{2}{\beta_i}\right)-\Gamma^2\left(1-\dfrac{1}{\beta_i}\right)\right], & \text{nếu}\mathrm{H_i}\sim\mathrm{MaxFD};\ \mathrm{L_i}\sim\mathrm{MinFD}. \end{cases}$$

2.5. Ứng dụng phân phối cực trị dự báo thủy văn tại Tây Nam bộ

Trên cơ sở nghiên cứu đã thấy được các phân phối cực hạn sẽ có dạng phân phối Gumbel thường dùng trong các mô hình về mưa và dòng chảy lũ, dạng phân phối Weibull thường dùng trong các mô hình về dòng chảy kiệt và hạn hán. Xét một số mô hình để đánh giá xây dựng quy hoạch vùng kinh tế của các tỉnh và khu vực Tây Nam bộ. Kết quả nghiên cứu của tác giả là các mô hình này đều có phân phối Gumbel (phân tích số liệu và quy luật đã chỉ ra trong luận án với các đặc trưng:

i) Kỳ vọng $EX = \mu + 0{,}577216\,\sigma$, $0{,}577216 \equiv$ hằng số Euler,

ii) Phương sai $Var\ X = \frac{(\pi\sigma)^2}{6}$,

iii) Bằng phương pháp moment ta có: $\begin{cases} \bar{X} = \mu + 0{,}577216\,\sigma, \\ S^2 = \frac{(\pi\sigma)^2}{6} \end{cases}$

Suy ra: $\begin{cases} \tilde{\mu} \approx \bar{X} - 0{,}4501\,S. \\ \tilde{\sigma} \approx 0{,}7797\,S. \end{cases}$, $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$; $S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$.

2.5.1 Dự báo cực đại cho mực nước sông Tiền qua Tân Châu, An Giang. *Phân tích* số liệu Đài Khí tượng Thủy văn Nam bộ (từ 1976 đến 2017) tại Trạm thủy văn trên sông Tiền qua Tân Châu, An Giang.

Thuật toán 2: Dự báo cực đại cho mực nước

Algorithm 2

Input: k, μ^0, σ^0

Bắt đầu

1) Xây dựng hàm hợp lý $L(\mu, \sigma)$, chọn $\hat{\mu}$ và $\hat{\sigma}$ thỏa
$(\partial\Lambda/\partial\mu = 0$ và $\partial\Lambda/\partial\sigma = 0)$
2) Vòng lặp thuật toán Newton – Raphson, đến khi
$$\Delta_j = \left(\mu^{(j+1)} - \mu^{(j)}\right)^2 + \left(\sigma^{(j+1)} - \sigma^{(j)}\right)^2 < k.$$
3) Hàm phân phối cực đại được xác định
$$F_2(x) \approx \exp\left\{-\exp\left\{\frac{-(x-375.3042)}{69.59}\right\}\right\}.$$
4) Đánh giá

Kết thúc

Ouput: Mực nước đã được xử lý qua hàm phân phối được xác định

Nghiên cứu đã thực hiện việc đánh giá, tại trạm Tân Châu, An Giang với $k = 10^{-4}$, $\mu^0 = 397{,}9874, \sigma^0 = 50{,}00598$, áp dụng thuật toán 2 ta sẽ có kết quả trong các bước lặp thể hiện trong bảng 2.

Bảng 2: Mực nước Tân Châu, An Giang

Bước j	μ^j	σ^0	Δ_j	10^{-4}
0	379.9874	50.00598		
1	377.0478	57.715	51.516	$>10^{-4}$
2	376.0177	64.818	15.47	$>10^{-4}$
3	375.4537	68.71	0.7335	$>10^{-4}$
4	375.3104	69.5544	0.00148	$>10^{-4}$
5	375.3042	*69.58668*	$4.1\text{x}.\ 10^{-5}$	$<10^{-4}$

Hàm phân phối cực đại trong bài toán này sẽ có dạng:
$$F_2(x) \approx \exp\left\{-\exp\left\{\frac{-(x-375.3042)}{69.59}\right\}\right\}.$$

Áp dụng thuật toán 2, cho bài toán về tính phân phối cực đại của lượng mưa tại trạm Tân Châu, An Giang với $k = 10^{-4}$, $\mu^0 =$

73.4567; $\sigma^0 = 16.2605$, có kết quả trong các bước lặp thể hiện số liệu trong bảng 3.

Bảng 3: Lượng mưa tại Tân Châu, An Giang

Bước j	μ^j	σ^0	Δ_j	10^{-4}
0	73.4567	16.2603		
1	72.92	17.997	3.303	$>10^{-4}$
2	72.725	18.777	0.647	$>10^{-4}$
3	72.698	18.889	0.0134	$>10^{-4}$
4	72.69766	18.891	$3.8\text{x}10^{-4}$	$< 10^{-4}$

Hàm phân phối cực đại trong bài toán này sẽ có dạng

$$F_3(x) \approx \exp\left\{-\exp\left\{\frac{-(x - 72.69766)}{18.891}\right\}\right\}$$

Sau khi phân tích thành phần, đặc tính chuỗi thời gian, theo các mô hình; với việc chứng minh định lý, và thực hiện việc xử lý theo dữ liệu, đã cho thấy ý nghĩa của việc phân tích dữ liệu. Trong đó có vai trò của công tác dự báo liên quan đến độ mặn, lượng mưa, mực nước sông cho khu vực ĐBSCL. Việc dự báo này góp phần trong việc định hướng phát triển văn hóa, xã hội và kinh tế cho vùng ĐBSCL.

Chương 3. PHƯƠNG PHÁP THỐNG KÊ BOOTSTRAP TRONG PHÂN TÍCH DỮ LIỆU CHUỖI THỜI GIAN

3.1 Phương pháp bootstrap trong thống kê

Các giá trị trong một mẫu đã thu được có thể được lặp lại một cách ngẫu nhiên, phương pháp bootstrap không chỉ xem một mẫu như một thể hiện mà xét nó như một tổng thể hay đại diện của tổng thể.

Cho mẫu ngẫu nhiên cỡ n, $\{X_1, X_2, \dots X_n\}$ từ hàm phân phối F chưa biết. Trong thống kê ta thường coi $X_i; \forall i = 1,2, \dots, n$ là những đại lượng ngẫu nhiên độc lập và có cùng phân phối F. Mỗi phép thử ngẫu nhiên là một thể hiện của mẫu tạo ra một giá trị cụ thể $(x_1, x_2, \dots x_n)$ trong R^n.

Xét thống kê $T_n(x_1, x_2, \dots x_n)$ phụ thuộc vào dữ liệu ngẫu nhiên trong mẫu, và phân phối của thống kê $G_n(x) = P(T_n \le x)$ thông thường phụ thuộc vào $F(x)$, để mô tả mối phụ thuộc đó ta thường viết $G_n(x, F)$. Trong trường hợp khi phân phối F chưa biết hoặc chưa xác định được thường ước lượng bởi hàm phân phối thực nghiệm (EDF – empirical distribution function) $\widehat{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le x)$ $P(A)$ có thể xấp xỉ bởi $\hat{P}(A)$, đối với mọi $A \in$ ß.

Quá trình lấy mẫu từ phân phối thực nghiệm $\widehat{P}(A)$ được thể hiện theo thuật giải Bootstrap độc lập có cùng phân phối.

Bootstrap có một số đặc tính phù hợp để phân tích dữ liệu chuỗi thời gian:

Tính vững (consistency) *Cho $\widehat{F_n}$ là phân phối thực nghiệm của F. Ước lượng bootstrap $G_n(X, \widehat{F_n})$ của $G_n(X, F)$ là vững nếu:*

$$\sup_X \left| G_n(X, \widehat{F_n}) - G_\infty(X, \widehat{F_n}) \right| \xrightarrow{p} 0 \qquad (3.1)$$

hội tụ theo xác suất.

Tính đúng đắn tiệm cận (Asymtotic validity of the bootstrap)

Với H_0 là giả thiết thống kê cần kiểm định, khi đó bootstrap là tiệm cận đúng nếu: $T_n^* \xrightarrow{d^*} G_\infty(X, \widehat{F_0})$ (3.2)

hội tụ theo phân phối.

3.2 Các phương pháp Bootstrap khả dụng cho dữ liệu phụ thuộc

*Phương pháp jackknife***,** cho một xấp xỉ đơn giản, ước tính sai số chuẩn và độ lệch sai số tiêu chuẩn $\widehat{se}_{jack} = \left[\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(.)}\right)^2\right]^{\frac{1}{2}}$

Khoảng tin cậy có thể được lấy từ hàm hợp lý, bằng cách sử dụng xấp xỉ $\mathcal{L}(\theta_0) = 2\left(L(\hat{\theta}_0) - L(\theta_0)\right) \sim \chi_2^2$.

Kỹ thuật bootstrap khối (block bootstrap) cho mỗi một khối cấu trúc của chuỗi mang tính xu hướng của chuỗi thời gian dừng

Bootstrap khối không chồng (Non-overlapping Block Bootstrap), cho trường hợp chuỗi thời gian đơn biến.

Bootstrap tự hồi quy sàng (AR-Sieve Bootstrap), Bootstrap tự hồi quy sàng – AR-SB được thiết lập cho mô hình chuỗi thời gian tự hồi quy (autoregressive – AR) $AR(p)$.

Mô hình chuỗi thời gian tuyến tính, nếu nó biểu diễn được dưới dạng: $x_t = \mu + \sum_{i=0}^{\infty} \psi_i a_{t-i}$

3.4 Đánh giá Bootstrap khối trong mô hình chuỗi thời gian

Thuật toán 3: Mô phỏng dữ liệu từ lý thuyết sang thực nghiệm

Algorithm 3

Input: Chuỗi thời gian lý thuyết
Bắt đầu

1) *arima.sim, với* ε_t là chuỗi nhiễu trắng độc lập và có cùng phân phối $N(0,1)$, kỳ vọng mẫu thực tế bằng không.

2) AR sinh bởi mô hình $x_t = \varphi_1 x_{t1} + \varphi_2 x_{t2} + \varepsilon_t$, với các tham số φ_1, φ_2;

3) MA sinh bởi mô hình $x_t = \theta_1 \varepsilon_{t1} + \theta_2 x_{t2} + \varepsilon_t$ với các tham số θ_1, θ_2;

4) Lặp khối $sd(\hat{\theta}) = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}\left(\hat{\theta}_i^* - \bar{\theta}^*\right)^2}$

5) Đánh giá chiều dài chuỗi thời gian và chiều dài khối

Kết thúc

Ouput: Chuỗi thời gian thực nghiệm

Kết quả: thực hiện các phương pháp lấy mẫu khối lặp MBB, NBB, CBB, CB trong xác định khoảng tin cậy cho giá trị θ. Gía trị thực của trung bình mẫu bằng không khi mô phỏng chuỗi thời gian. Với mỗi ước lượng bootstrap và với p-giá trị chọn $\alpha = 0{,}05$, xác định khoảng tin cậy $\left[\hat{\theta}^*_{\left(\frac{\alpha}{2}\right)}, \hat{\theta}^*_{\left(1-\frac{\alpha}{2}\right)}\right]$, (Kết quả theo mô hình M1 với số lặp bằng 1000) chỉ ra kết quả theo mô hình M1

Tiếp theo, với độ dài khối bằng 10, theo mô hình M1 ta sẽ có các kết quả thể hiện trong Bảng 34 (Kết quả theo mô hình M1 với độ dài khoảng bằng 10).

Qua các số liệu thu được, thấy rằng các phương pháp MBB và CBB cho kết quả tốt hơn NBB và SB trong trường hợp này. Thuật toán lần lượt sẽ có các dữ liệu cho dạng mô hình M2, M3, M4.

Từ các kết quả trên ta xác định ước lượng bootstrap theo các cách chọn khối khác nhau tương ứng với các công thức thống kê ta sẽ có các ước lượng tham số theo mô hình M1 dưới dạng Bảng (các ước lượng tham số theo mô hình M1).

Kết quả trên, thể hiện rằng cỡ mẫu và cấu trúc phụ thuộc đóng vai trò quan trọng trong việc chọn mô hình và tối ưu thuật giải.

Các phương pháp lấy mẫu lặp lại thường được sử dụng để suy luận về các tham số mô hình. Cỡ mẫu và cấu trúc phụ thuộc đóng vai trò trong việc chọn thủ tục bootstrap tối ưu. Nghiên cứu đã đưa ra những đánh giá và so sánh các dạng bootstrap khối trong phân tích mô hình. Đồng thời sử dụng chúng trong các mô phỏng với những kết quả cụ thể.

Chương 4. GIẢI TÍCH NGẪU NHIÊN TRONG BÀI TOÁN ĐÁNH GIÁ VÀ DỰ BÁO DỮ LIỆU

4.1 Tóm tắt về quá trình ngẫu nhiên

Quá trình ngẫu nhiên X_t, là họ các đại lượng ngẫu nhiên $(X_t, t \in T)$ xác định trên cùng một không gian xác suất $(\Omega, \mathcal{F}, P)$.

Bộ lọc của quá trình ngẫu nhiên trên $(\Omega, \mathcal{F})$ là họ hàm tập không giảm $\{\mathcal{F}_t, t \in T\}$ của: $\sigma -$đại số $\mathcal{F}$: $\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}$; $\forall\ s \leq t$.

Các quá trình ngẫu nhiên đều được xét trên không gian xác suất có lọc tương ứng $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$

4.2 Quá trình Wiener và quá trình Poisson

Thông qua việc nghiên cứu quá trình Wiener $W(t)$ và tích phân Itô cùng với dạng suy rộng là tích phân Ito-Levy, đã cho một hướng tiếp

cận hiện đại với lớp các quá trình ngẫu nhiên liên tục và có nhảy từ đó giải quyết được nhiều bài toán thực tế như các bài toán khuếch tán và dẫn truyền trong môi trường có nhiễu của các loại sóng điện từ.
Xuất phát điểm trong vấn đề này đã xây dựng mô hình về phương trình vi phân ngẫu nhiên tuyến tính tổng quát và nghiên cứu về sự tồn tại nghiệm cùng với cách giải đúng của chúng.
Bên cạnh đó là phương pháp số để giải gần đúng thông qua các độ đo ngẫu nhiên liên tục và độ đo Poisson $P(dt, dz)$ có nhảy. Bằng hai hướng giải quyết như vậy có thể nói bài toán khuếch tán đã được giải quyết phần quan trọng vấn đề cần mở rộng là xét các đang nhiễu phức tạp trong các môi trường dẫn truyền thực tế.

4.3 Giải phương trình vi phân ngẫu nhiên tuyến tính thuần nhất

Phương trình vi phân tuyến tính Itô – Levy là những quá trình ngẫu nhiên cho trước dạng:

$$\mathrm{d}X(t) = [\alpha(t,\omega)X(t^-) + A(t,\omega)]\mathrm{d}t + [\beta(t,\omega)X(t^-) + B(t,\omega)]\mathrm{d}W(t) + \int_{(R_0)^{n_2}} [\gamma(t,z,\omega)X(t^-) + G(t,z,\omega)]\overline{N}\,(\mathrm{d}t,\mathrm{d}z)$$

với: $\gamma(t,z,\omega) > -1;\ \forall (t,z,\omega) \in [0,\infty\} \times R_0 \times\ \Omega$

Việc chứng minh thông qua *Phương trình vi phân ngẫu nhiên tuyến tính thuần nhất, có dạng:*

$$dX(t) = X(t^-)\left[\alpha(t,\omega)dt + \beta(t,\omega)dW(t) + \int_{R_0} \gamma(t,z,\omega)\overline{N}(dt,dz)\right]$$

Xét $X_1(t) = F\big(t, H(t)\big); t \geq 0$ với $F(t,x) = e^x$ và $H(t)$ xác định bởi:

$$H(t) = \int_0^t \left[\alpha(s,\omega) - \frac{1}{2}\beta^2(s,\omega) + \int_{R_0} \log\big(1+\gamma(s,z,\omega)\big) - \gamma(s,z,\omega)v(dz)\right] ds + \int_0^1 \beta(s,\omega)dW(s) + \int_0^1 \int_{R_0} \log\big(1+\gamma(s,z,\omega)\big)\overline{N}(ds,dz)$$

Áp dụng công thức Ito cho $X_1(t) = F\big(t, H(t)\big)$, ta sẽ thu được được nghiệm của phương trình là:

$dX_1(t) = e^{H(t)}\left[\left(\alpha(t,\omega) - \frac{1}{2}\beta^2(t,\omega) + \int_{R_0}\left[\log\left(1+\gamma(t,z,\omega)\right) - \gamma(t,z,\omega)\right]v(dz)\right)dt\right]$

$+e^{H(t)}\left[\frac{1}{2}\beta^2(t,\omega)dt + \beta(t,\omega)dW(t)\right] + \int_{R_0} e^{H(t)}\left[\gamma\left((t,z,\omega) - \log\left(1+\gamma(t,z,\omega)\right)\right)\right]v(dz)dt \;\; + + \int_{R_0} e^{H(t^-)}\gamma(t,z,\omega)\tilde{N}(dt,dz) =$

$X_1(t^-)\left[\alpha(t,\omega)dt + \beta(t,\omega)dW(t) + \int_{R_0}\gamma(t,z,\omega)\tilde{N}(dt,dz)\right]$ ■

4.4 Phương pháp tách nghiệm

Phương trình vi phân ngẫu nhiên tuyến tính tổng quát, còn gọi phương trình Itô-Levy. Đề xuất phương pháp tách nghiệm để tìm nghiệm phương trình tuyến tính, nghĩa là tìm nghiệm dưới dạng tích

$$X(t) = X_1(t^-).X_2(t^-) \tag{4.1}$$

trong đó, $X_1(t)$ là nghiệm của phương trình tuyến tính thuần nhất tương ứng

$$\frac{dX_1(t)}{X_1(t^-)} = \left[\alpha(t,\omega)dt + \beta(t,\omega)dW(t) + \int_{R_0}\gamma(t,z,\omega)\bar{N}(dt,dz)\right] \tag{4.2}$$

và $X_2(t)$ là nghiệm của phương trình

$$dX_2(t) = A^*(t,\omega)dt + B^*(t,\omega)dW(t) + \int_{R_0} G^*(t,z,\omega)\tilde{N}(dt,dz)$$

Với điều kiện $X_2(0) = x_0$, trong đó $A^*(t,\omega)$; $B^*(t,\omega)$; $G^*(t,z,\omega)$

Từ phương trình tuyến tính thuần nhất, ta có nghiệm $X_1(t^-)$ của (4.2) cho bởi hệ thức $X_1(t) = \exp\left\{\int_0^t\left[\alpha(s,\omega) - \frac{1}{2}\beta^2(s,\omega) + \int_{R_0}\log\left(1+\gamma(s,z,\omega)\right) - \gamma(s,z,\omega)v(dz)\right]ds + \int_0^1\beta(s,\omega)dW(s) + \int_0^1\int_{R_0}\log\left(1+\gamma(s,t,\omega)\right)\bar{N}(ds,dz)\right\}$ (4.3)

Áp dụng vi phân cho tích hai hàm ngẫu nhiên

$X(t) = X_1(t^-).X_2(t^-)$, ta được:

$d(X(t) = d\left(X_1(t^-).X_2(t^-)\right) = X_1(t^-).dX_2(t) + .X_2(t^-)dX_1(t) + \beta(t,\omega)\,X_1(t^-)B^*(t,\omega)dt + \int_{R_0}\gamma(t,z,\omega)X_1(t^-)G^*(t,z,\omega)\tilde{N}(dt,dz) =$

$\alpha(t,\omega)X_1(t^-)X_2(t^-)dt +$

$\beta(t,\omega)X_1(t^-)X_2(t^-)dW(t) + \int_{R_0}\gamma(t,z,\omega)X_1(t^-)X_2(t^-)\tilde{N}(dt,dz) +$

$X_1(t^-)A^*(t,\omega)dt \quad + X_1(t^-)B^*(t,\omega)dW(t) +$
$X_1(t^-)\int_R \ G^*(t,z,\omega)\tilde{N}(dt,dz) + \beta(t,\omega)X_1(t^-)B^*(t,\omega)dt +$
$\gamma(t,z,\omega)X_1(t^-)G^*(t,z,\omega)N(dt,dz).$

Mặt khác $X(t)$ là nghiệm của phương trình tuyến tính $dX(t) =$ $[\alpha(t,\omega)X(t^-) + A(t,\omega)]dt + [\beta(t,\omega)X(t^-) + B(t,\omega)]dW(t) +$ $\int_{(R_0)^{n_2}}[\gamma(t,z,\omega)X(t^-) + G(t,z,\omega)]\overline{N}\,(dt,dz)$ (4.4)

từ đó so sánh giữa (4.4) và (4.1), ta thu được hệ phương trình:

$$\begin{cases} A(t,\omega) = X_1(t^-)\left[A^*(t,\omega) + B(t,\omega)B^*(t,\omega) + \int_{R_0} \gamma(t,z,\omega)G(t,z,\omega)v(dz)\right] \\ B(t,\omega) = X_1(t^-)B^*(t,\omega) \\ \int_{R_0} G(t,z,\omega)\tilde{N}(dt,dz) = X_1(t^-)\int_{R_0} \big(1+\gamma(t,z,\omega)\big)G^*(t,z,\omega)\tilde{N}(dt,dz) \end{cases}$$

Suy ra:

$$\begin{cases} A^*(t,\omega) = \frac{1}{X_1(t^-)}\left[A(t,\omega) - B(t,\omega)\beta(t,\omega) - \int_{R_0} \frac{\gamma(t,z,\omega)G(t,z,\omega)}{1+\gamma(t,z,\omega)}v(dz)\right] \\ B^*(t,\omega) = \frac{B(t,\omega)}{X_1(t^-)} \\ G^*(t,z,\omega) = \frac{G(t,z,\omega)}{X_1(t^-)\big(1+\gamma(t,z,\omega)\big)} \end{cases}$$

Đặt $X_1(t)$ cho bởi (4.3), và các biểu thức của $A^*(t,\omega); B^*(t,\omega); G^*(t,z,\omega)$ đã xác định được vào (4.1), ta sẽ có nghiệm phương trình đã cho. ■

4.6.3 Định hướng dự báo trong tài chính

Vận dụng giải phương trình vi phân ngẫu nhiên tuyến tính theo phương pháp tách nghiệm trong bài toán về tài sản và phương án đầu tư trên thị trường bằng các công cụ của giải tích ngẫu nhiên.

Loại tài sản phi rủi ro với biến động giá $X_0(t)$ được xét qua phương trình vi phân: $dX_0(t) = \lambda(t)X_0(t)dt\,;\ X_0(0) = 1\,; t \in [0,T]$

Loại tài sản rủi ro với biến động giá $X_1(t)$ thường được xét qua phương trình vi phân ngẫu nhiên có dạng:

$dX_1(t) = \alpha(t)X_1(t)dt + \beta(t)X_1(t)dB_t(\omega)\,;\ \ X_1(0) > 0, t \in [0,T]$

trong đó: $\lambda(t)$; $\alpha(t)$; $\beta(t)$ là những hàm tất định mang các giá trị dương khi ta mua tài sản, mang giá trị âm khi bán tài sản.

Ta ký hiệu: $\tau_0(t); \tau_1(t);\ t \in [0, T]$; là những đơn vị vốn đầu tư cho loại tài sản phi rủi ro và rủi ro tương ứng. Khi đó ta gọi: $\phi = \{(\tau; X)(t)\} \coloneqq \Big(\big(\tau_0(t), \tau_1(t)\big); \big(X_0(t), X_1(t)\big)\Big)$ là một phương án đầu tư (một danh mục đầu tư - portfolio) với tổng giá trị tài sản tại thời điểm t bằng: $V^{\tau}(t) = \tau_0(t)X_0(t) + \tau_1(t)X_1(t)$

Phương án đầu tư được gọi là tự hạch toán (*self-financing portfolio*) muốn tăng đầu tư vào một chứng khoán nào đó phải giảm đầu tư vào chứng khoán khác, không làm tăng hoặc giảm vốn đầu tư, nghĩa là:

$\tau_0(t)X_0(t) + \tau_1(t)X_1(t) = \tilde{\tau}_0(t)X_0(t) + \tilde{\tau}_1(t)X_1(t) \Rightarrow \big(\tau_0(t) - \tilde{\tau}_0(t)\big)X_0(t) + \big(\tau_1(t) - \tilde{\tau}_1(t)\big)X_1(t) = 0$

Ta đặt: $\Delta\tau_0(t) = \big(\tau_0(t) - \tilde{\tau}_0(t)\big)$;

$\Delta\tau_1(t) = \big(\tau_1(t) - \tilde{\tau}_1(t)\big) \Rightarrow \Delta\tau_0(t)X_0(t) + \Delta\tau_1(t)X_1(t) = 0.$

Viết dưới dạng vi phân: $X_0(t)d\tau_0(t) + X_1(t)d\tau_1(t) = 0.$

Mặt khác do: $V^{\tau}(t) = \tau_0(t)X_0(t) + \tau_1(t)X_1(t)$ với $\tau_0(t), \tau_1(t)$ là những hàm tất định, ta sẽ thu được $dV^{\tau}(t) = \tau_0(t)dX_0(t) + \tau_1(t)dX_1(t) + X_0(t)d\tau_0(t) + X_1(t)d\tau_1(t) \Rightarrow dV^{\tau}(t) = \tau_0(t)dX_0(t) + \tau_1(t)dX_1(t).$

Một phương án đầu tư $\{(\tau; X)(t)\}$, là một tự tài trợ khi và chỉ khi:

$$dV^{\tau}(t) = \tau_0(t)dX_0(t) + \tau_1(t)dX_1(t).$$

Từ tổng giá trị tài sản tại thời điểm t, ta rút ra: $\tau_0(t) = \frac{V^{\tau}(t) - \tau_1(t)X_1(t)}{X_0(t)}$

Từ điều kiện về tự tài trợ sẽ dẫn đến hệ thức:

$dV^{\tau}(t) = \lambda(t)\big(V^{\tau}(t) - \tau_1(t)X_1(t)\big)dt + \tau_1(t)dX_1(t)$

Dựa vào phương trình trên, có được:

$dV^{\tau}(t) = \big[\lambda(t)V^{\tau}(t) + \big(\alpha(t) - \lambda(t)\big)\tau_1(t)X_1(t)\big]dt + \beta(t)\tau_1(t)X_1(t)dB_t$

Điều này cho thấy phương trình trên là một phương trình vi phân ngẫu nhiên tuyến tính, có thể sử dụng phương pháp tách nghiệm.

Thông qua giải phương trình vi phân tuyến tính Itô-Levy, giải phương trình vi phân ngẫu nhiên tuyến tính thuần nhất, giải phương trình vi phân ngẫu nhiên tuyến tính tổng quát, chúng ta có những đánh giá trong phân tích dữ liệu trong tài chính.

KẾT LUẬN & KIẾN NGHỊ

KẾT LUẬN

Dữ liệu chuỗi thời gian là một trong những dạng thường gặp nhất trong tự nhiên. Việc phân tích, đánh giá, dự báo về nó là một bài toán rộng và khó của Toán học và Công nghệ thông tin. Kết quả đánh giá và dự báo dữ liệu chuỗi thời gian rất có ý nghĩa trong các nghiên cứu khoa học và thực tế.

Luận án đã phân tích dữ liệu chuỗi thời gian trong các đánh giá và dự báo, với kết quả đạt được cụ thể.

(1) Phân tích dữ liệu chuỗi thời gian ***theo các phương pháp kinh điển*** *của lý thuyết Xác suất và Thống kê*, theo dạng các mô hình hồi quy trung bình trượt tích hợp phối hợp với các dạng phân phối cực trị của chuỗi. Kết quả này được trình bày trong *chương 1, và mở rộng trong chương 4*, khi xét đến các dạng dữ liệu ngẫu nhiên liên tục, là một số bài toán cực trị trong thủy văn tại một số tỉnh Miền Tây Nam Bộ. Dữ liệu thu được thường là giá trị của một loại quá trình ngẫu nhiên xác định và có liên quan đến những loại biến động như nhiễu trắng, kích động có nhảy, nguồn tán xạ điện từ,… Nghiên cứu đã mở rộng sang các loại dữ liệu bởi các quá trình ngẫu nhiên liên tục hoặc rời rạc (chuỗi thời gian chỉ là dạng quá trình ngẫu nhiên với thời gian rời rạc) trong một môi trường phức hợp có nhiễu và có biến động nhảy.

(2) Phân tích về *dữ liệu chuỗi thời gian* ***theo các phương pháp mới của lý thuyết Xác suất và Thống kê Toán học,*** đó là: *Phương pháp toán mờ*, theo các mô hình khác nhau do tính đa dạng của các bài toán thường gặp trong thực tế (kinh tế, xã hội, công nghệ). Kết quả lý thuyết trên các dữ liệu thực tế được trình bày trong *chương 2*. Phân tích dữ liệu chuỗi thời gian theo một hướng mới của thống kê hiện nay là các *phương pháp Thống kê bootstrap*, phương pháp này tập trung vào quá trình phân tích lặp các dữ liệu đã có để sử dụng sức mạnh của máy tính khi mà tính toán lý thuyết không thực hiện được. Các nghiên cứu tập trung trình bày trong *chương 3,* với kết quả lý thuyết và ứng dụng trong bộ dữ liệu khí tượng thủy văn vùng Tây Nam bộ.

(3) Phân tích dữ liệu chuỗi thời gian theo một *hướng rộng và tổng quát nhất là bằng các quan điểm của Giải tích ngẫu nhiên*, từ đó có thể giải quyết triệt để được các bài toán phức hợp của thực tế sinh ra các dữ liệu ngẫu nhiên (trong bài toán về vật lý lượng tử hoặc trong các vấn đề của kinh tế vĩ mô). Kết quả nghiên cứu và hướng ứng dụng đã được trình bày trong *chương 4.*

Với các kết quả trên đều có ý nghĩa lớn về mặt lý thuyết và ứng dụng, khi giải quyết các bài toán ứng dụng để dự báo cần phải sử dụng phối hợp các nghiên cứu này trong một chỉnh thể hỗ trợ. Trong đó, việc phân tích dữ liệu chuỗi thời gian theo bootstrap, một hướng mới của thống kê, là một hướng của tương lai vì hướng này sử dụng được các thế mạnh của máy tính và trí tuệ nhân tạo, tránh được sự phụ thuộc máy móc vào khai phá dữ liệu trong thống kê kinh điển.

KIẾN NGHỊ

Từ những kết quả đã đạt được, luận án là tiền đề để tiếp tục giải quyết các hạn chế và tập trung vào việc chuyên sâu bootstrap để giải quyết bài toán thiếu dữ liệu thực tế trong một số lĩnh vực khác nhau, đặc biệt là các bài toán với biến ngẫu nhiên khó xác định.