TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
Chuyên ngành: KHOA HỌC MÁY TÍNH
Mã số: 62.48.01.01

LUẬN ÁN TIẾN SĨ

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)

NCS: Đặng Kiên Cường
CBHD: TS. Trần Tích Phước
TS. Dương Tôn Đảm

NỘI DUNG

01

LÝ DO, MỤC TIÊU

Tính cấp thiết của Luận án

- Dữ liệu chuỗi thời gian
  - Quản lý thiên tai, Dự báo thiên tai (Khí tượng thủy văn)
  - Khí tượng thủy văn dữ liệu lớn (>= 30 năm)
  - 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
- Trong QL 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

MỤC TIÊU

- Mục tiêu tổng quát: Phân tích, đánh giá và dự báo chuỗi thời gian KTTV nhằm hỗ trợ quản lý
- Mục tiêu cụ thể:
  - Nghiên cứu về tập dữ liệu trong biến động theo thời gian, để tìm ra *quy luật hoặc những đặc tính cơ bản* của tập dữ liệu.
  - Xây dựng *mô hình dự báo* trên cơ sở các quy luật hoặc các đặc tính của tập dữ liệu thực tế và tiến hành huấn luyện, kiểm tra bằng các thuật toán phù hợp.
  - Phân tích tập dữ liệu bằng *các phương pháp mới*, đó là việc *tích hợp* toán thống kê *kinh điển và hiện đại.*

02

TỔNG QUAN NGHIÊN CỨU

PHÂN TÍCH, ĐÁNH GIÁ, DỰ BÁO

- Một trong những vấn đề quan trọng nhất của dữ liệu đó là phân tích và dự báo dữ liệu.

1. Hướng nghiên cứu kinh điển trong xác suất và thống kê như Lý thuyết tương quan và hồi quy với các phương pháp ARMA, ARIMA, phân tích PCA, phân tích phương sai,… được nghiên cứu ban đầu bởi Pearson, Bayes, Holt-Winters.
2. Phát triển bởi Box-Jenkins và Van der Vaart, Chen H,... mở rộng sang các dạng tiệm cận và toán mờ trong thống kê.

PHÂN TÍCH, ĐÁNH GIÁ, DỰ BÁO (1)

3. Cạnh đó là các phương pháp thống kê Bootstrap để khắc phục những khiếm khuyết trong thu thập dữ liệu mẫu từ những khái niệm lặp có hoàn của B. Efron (1990). Phương pháp Bootstrap trở nên một công cụ rất hữu ích khi nghiên cứu về chuỗi thời gian, đặc biệt là các dạng Bootstrap khối. Trong đó phải kể đến:

- Thuật toán tổng hợp – bootstrap aggregating được Breiman giới thiệu vào năm 1996;
- Phương pháp Bergmeir C. (2016) tạo lập bootstrap từ phần còn lại của nó qua sự phân hủy STL “Seasonal and Trend decomposition using Loess”
- Phương pháp Laurinec P. (2019) tạo lập boostrap dựa trên K-means clustering.

Định hướng nghiên cứu

Trên cơ sở nghiên cứu các Quy luật và đặc tính của các dữ liệu ngẫu nhiên trong chuỗi thời gian (Luật phân phối cực trị EVD cùng các đặc tính của nó)

- Dữ liệu thủy văn tại ĐBSCL qua các dòng chảy chính và với những biến động dị thường (bão, lũ, ngăn dòng, xây đập) và trong xu thế biến đổi khí hậu hiên nay.
- Bài toán dự báo về chuỗi thời gian có thể sử dụng các phương pháp mới của Thống kê toán để nâng cao hiệu quả và hạn chế tác hại. Qua đó sẽ nâng được các giá trị về xử lý dữ liệu về mặt lý thuyết và cả thực tiễn.
- Nghiên cứu đã thu đươc các kết quả phù hợp với mục tiêu theo các định hướng trên.

NGHIÊN CỨU CÓ LIÊN QUAN

- Nguyễn Văn Thắng, “Nghiên cứu xây dựng hệ thống dự báo, cảnh báo hạn hán cho Việt Nam với thời hạn đến 3 tháng”; 2016
- Phan Văn Tân (dịch), NXB ĐHQG HN, 2005. Lý thuyết xác suất, thống kê, lý thuyết hàm ngẫu nhiên, toán học quan trọng sử dụng trong khí tượng, thủy văn.
- Nguyễn Văn Thu, Nguyễn Đức Phương (2008), Ứng dụng phương pháp Bootstrap để nhận biết mức độ nguy hiểm của căn bệnh loãng xương.
- Hoàng Thị Diệp (2017), bootstrap cây tiến hóa là kĩ thuật phổ biến để xác định độ tin cậy cây tiến hóa, đề xuất phương pháp giải quyết: thời gian, độ chính xác, ảnh hưởng của vi phạm mô hình và hiện tượng đa phân, mở rộng cho dữ liệu.

NGHIÊN CỨU CÓ LIÊN QUAN

- Nick M., Das S., Simonovic S. P., The Comparison of GEV, Log-Pearson Type 3 and Gumbel Distributions in the Uppee Thames River Watershed under Global Climate Models, The University of Western Ontario; London, Ontario. Canada, R. No:77, 2011.
- Benstock D. , Extreme value analysis (EVA) of inspection data and its uncertainties, NTD & E Intrenational Vol: 87, 68-77, Elsevier, 2017.
- Carsten J., Christian H. W., Boostraping integer-valued autoregressive models, University of Mannheim, ***2017***, W-P 17-02.
- Gul Nisa , Farhat Iqbal, Bootstrapping the Li-Mak and McLeod-Li Portmanteau Tests for GARCH Models, The Journal of Middle East and North Africa Sciences, ***2018***; 4(01)

NGHIÊN CỨU CÓ LIÊN QUAN

- Carsten J., Christian H. W., Boostraping integer-valued autoregressive models, University of Mannheim, ***2017.***
- Arturo Kohatsu-Higa, Atsushi Takeuchi, Jump SDEs and the study of their densities, Springer Nature Singapore Pte Ltd, ***2019***
- Bergmeir, C., Hyndman, R. J., Koo, B., A note on the validity of cross-validation for evaluating autoregressive time series prediction, *Computational Statistics and Data Analysis,* 2018
- Anna E. Dudek , Block boostrap for periodcic characteristics of periodcically correlated time series, Journal of Nonparametric Statistcs, American Statistical Association, 2018.
- Gao M., Extreme value analysis and Risk Communication for a Changing Climate, Advances in Environmental Monitoring and Assessment . Intech Open, Edited by Suriyanarayanan Sarvajayakesavalu, 84-102, Published in London, UK, 2019.

03

DỮ LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

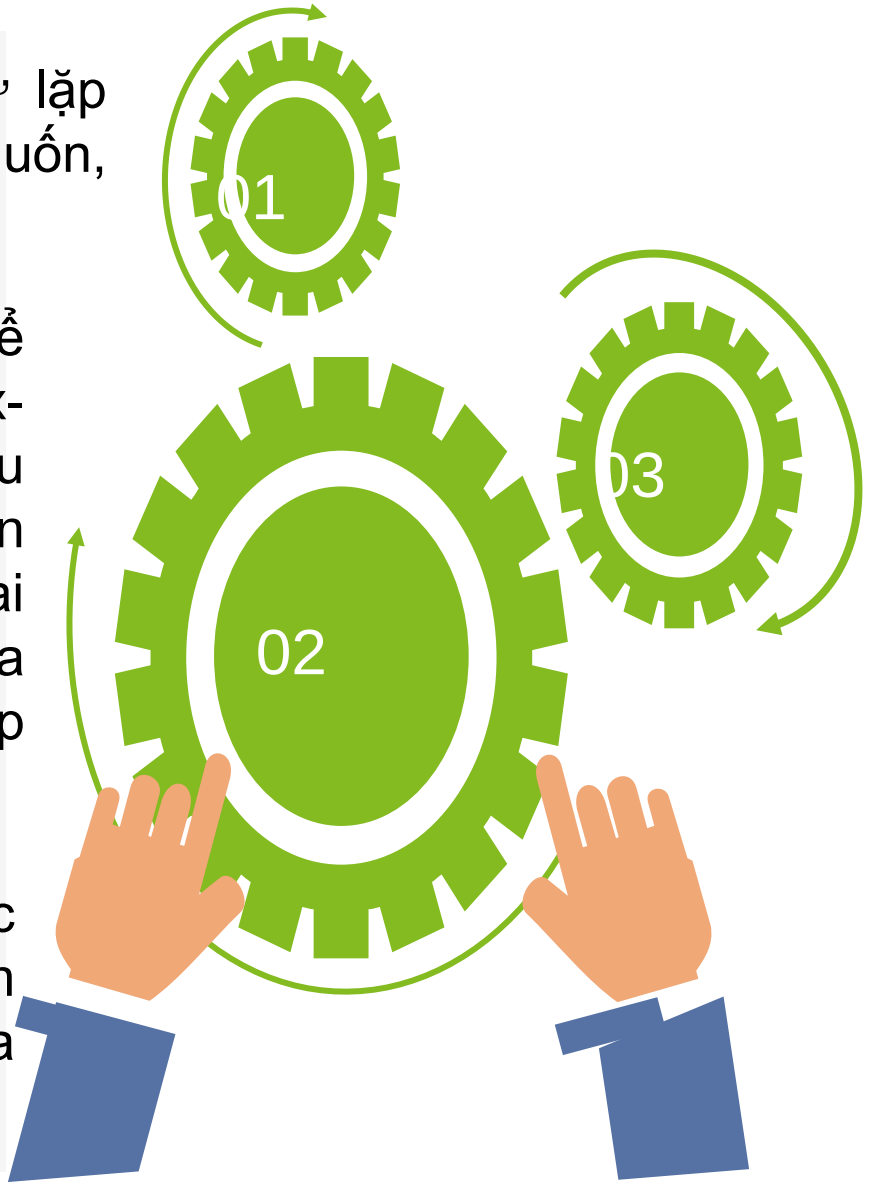
DỮ LIỆU

Loại dữ liệu	Mô tả	Nguồn thu thập
1. Lượng mưa	Biến số: Mưa, Tmax, Tmin, Tmean, ET, RH Giai đoạn: 1978/1986 – 2015	Đài Khí tượng Thuỷ văn Nam Bộ
2. Mực nước	Biến số: Nước, Tmax, Tmin, Tmean, Date	
	Giai đoạn: 1990-2017	Đài Khí tượng Thuỷ văn Nam Bộ
3. Độ mặn	Biến số: Mặn, Tmax, Tmin, Tmean, Date	Đài Khí tượng Thuỷ văn Nam Bộ
	Giai đoạn: 2000-2017	
4. Dữ liệu toàn cầu CRU TS4.02	Biến số: Mưa, Tmax, Tmin, Tmean Giai đoạn: 1901-2017, 1951-2017, 1981-2017	Climatic Research Unit (University of East Anglia – UK)

Cập nhật 12/2017

Xử lý dữ liệu

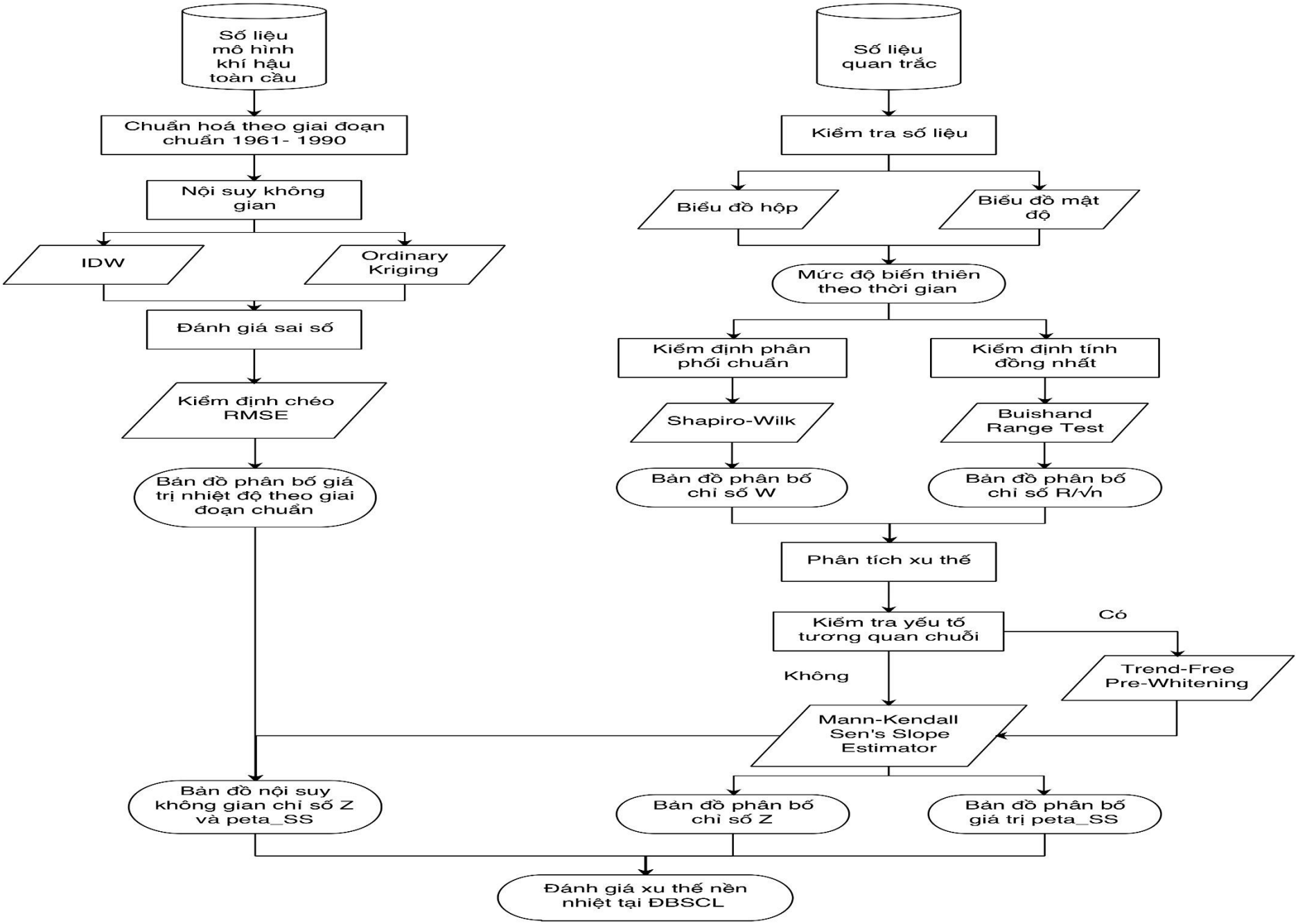
- Thiếu dữ liệu do: không có sự lặp lại, vấn đề không mong muốn, không có điều kiện để thử.
- Từ mô hình ARMA, ARIMA thể hiện trong phương pháp Box-Jenkins tích hợp với xử lý dữ liệu dưới dạng bootstrap: chỉ dựa trên 1 mẫu (sample), tiến hành lặp lại (trên 1.000 lần với sự hỗ trợ của máy tính) để thay thế cho tập tổng thể (population)
- Từ nhận dạng quy luật và thực hiện dự báo, xác định được kích cỡ của khối và tốc độ hội tụ của khối



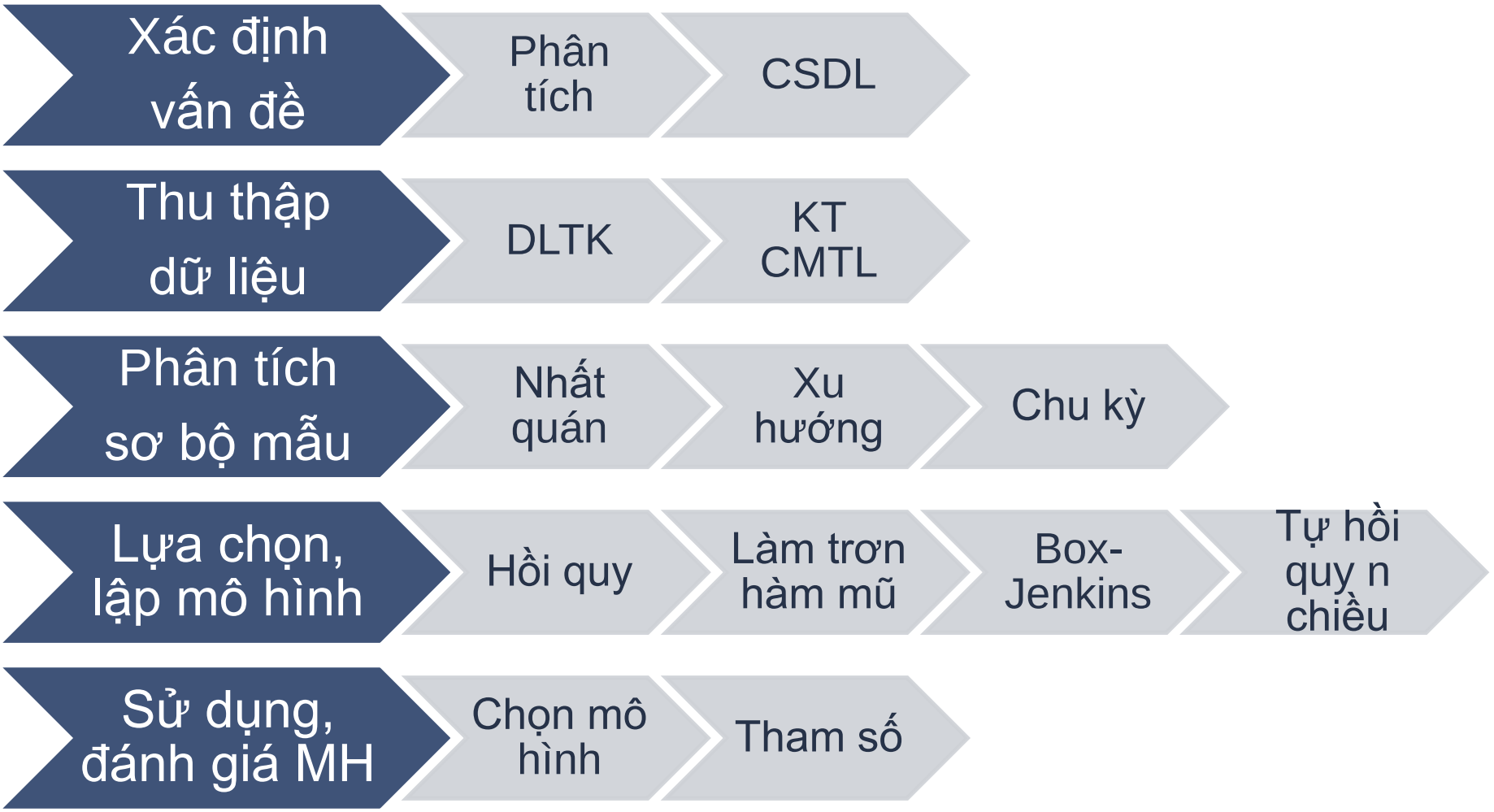
Phương Pháp nghiên cứu

- 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,…).
- Công cụ chính là các phép tính ***vi-tích phân ngẫu nhiên*** với các phương pháp Toán hiện đại:
  - *Toán mờ* (Tương quan, hồi quy mờ, phân tích mờ và giải mờ)
  - Thống kê *bootstrap* (jackknife, bootstrap khối, bootstrap dừng,…)
  - Lý thuyết về quá trình *khuếch tán ngẫu nhiên* có nhảy

Thuật toán phân tích dữ liệu



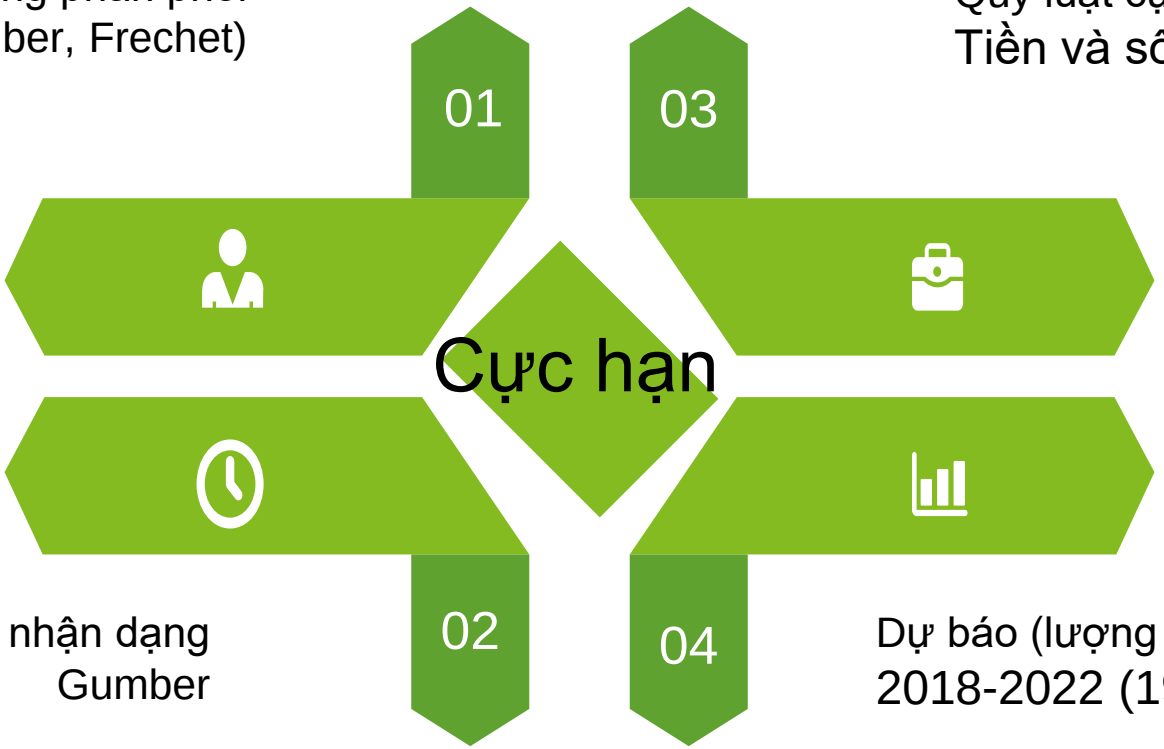
Nghiên cứu dự báo



Đặc tính của dữ liệu

	Dữ liệu tất định	Dữ liệu ngẫu nhiên
Quan hệ hàm	$f(t,x)$: $R^2 \to R$	$f(t,\omega)$: $R \times \Omega \to R$
Công cụ xử lý	Giải tích thực: Vi-tích phân hàm tất định Xấp xỉ và giới hạn với topô trong KG thực R^n Mô phỏng hàm thực…	Giải tích ngẫu nhiên: Vi-tích phân hàm ngẫu nhiên Xấp xỉ và các dạng giới hạn trong KG Xác suất nhiều chiều Mô phỏng ngẫu nhiên Monter-Carlo…
Dự báo	Dự báo điểm, khoảng tất định Cực trị của hàm	Dự báo qua độ tin cậy XS Dự báo về quy luật của cực trị (EVD)

Bài toán Cực hạn



Quá trình ngẫu nhiên Ito-Levy

PTVPNN biến động

Lũ, kiệt

Yếu tố có liên quan đến con người: xây đập, phá đập

Ngẫu nhiên (từ yếu tố thiên nhiên: lũ, bão, triều cường)

i) Giải đúng, thực hiện thông qua phương pháp tách nghiệm
ii) Giải thông qua máy tính, theo phương pháp số.

(*Trình bày tại Hội nghị khoa học ĐHTN 2019, đăng trên TC KHCN 2019*)

Thuật toán 1: Dự báo đỉnh mặn

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.

trang 48 của LA

Thuật toán 2: Dự báo cực đại mực nước

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

$$(\partial\Lambda/\partial\mu = 0 \text{ và } \partial\Lambda/\partial\sigma = 0)$$

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

trang 67 của LA

Thuật toán 3: Mô phỏng Bootstrap

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

Algorithm 3

Input: Chuỗi thời gian lý thuyết

Bắt đầu

1) sử dụng hàm *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 $sd(\hat{\theta}) = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(\hat{\theta}_i^* - \bar{\theta}^*)^2}$

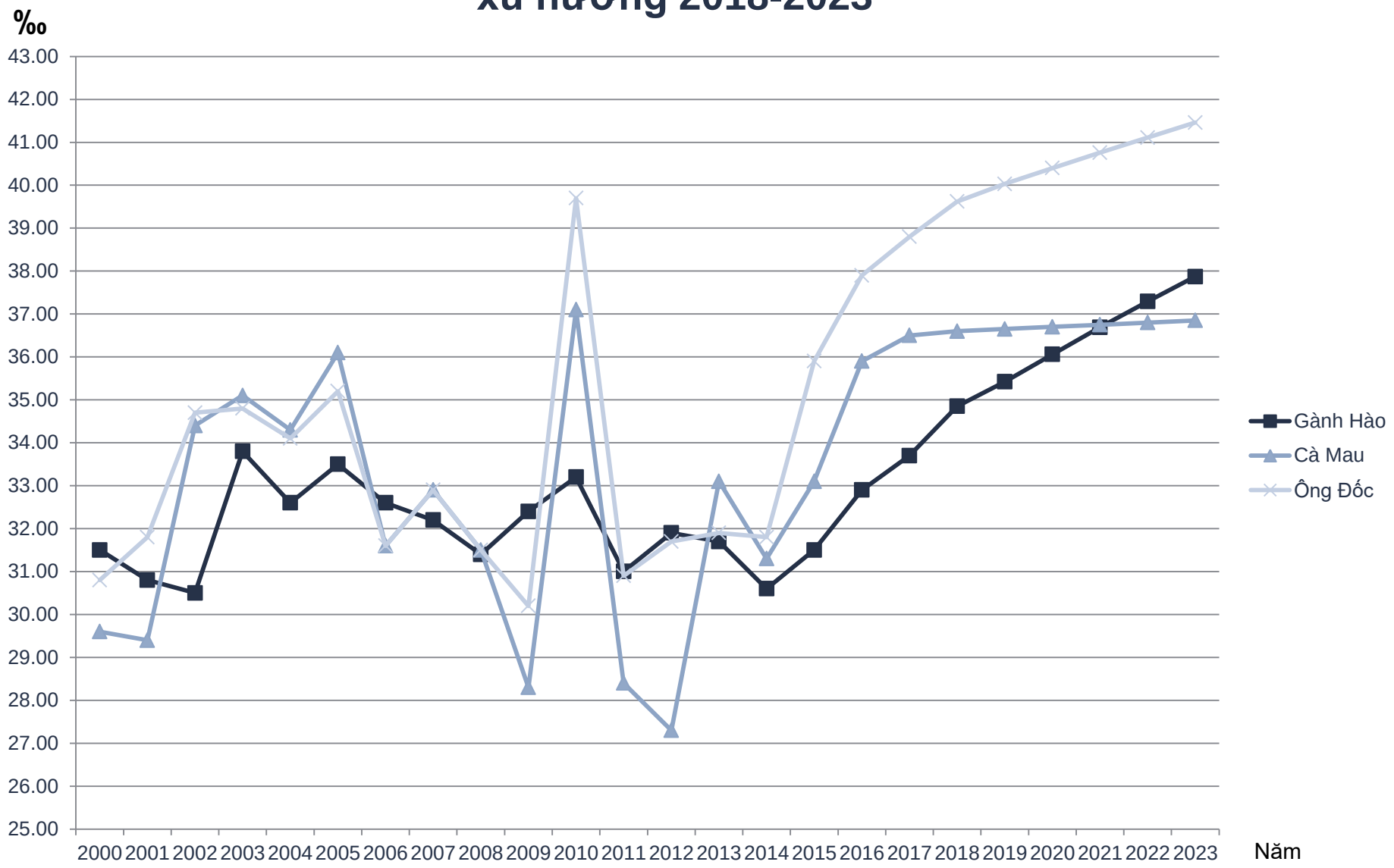
5) Đánh giá chiều dài chuỗi thời gian và chiều dài khối

Kết thúc

Ouput: Chuỗi thời gian thực nghiệm

trang 88 của LA

Đỉnh mặn tại 3 trạm đo (2000-2017), xu hướng 2018-2023



Hoạt động dự báo

Dữ liệu

- Ngẫu nhiên, Liên tục, Rời rạc, nhiều chiều

Quy luật

- μ^j, σ^j

Mô hình

- $F_1(\mathrm{x}) \approx \exp\left\{-\exp\left\{\frac{-(\mathrm{x}-30{,}6767)}{2{,}605}\right\}\right\}$
- $(\mathrm{x}) \approx \exp\left\{-\exp\left\{\frac{-(\mathrm{x}-375.3042)}{69.59}\right\}\right\}$
- $F_3(\mathrm{x}) \approx \exp\left\{-\exp\left\{\frac{-(\mathrm{x}-72.69766)}{18.891}\right\}\right\}$

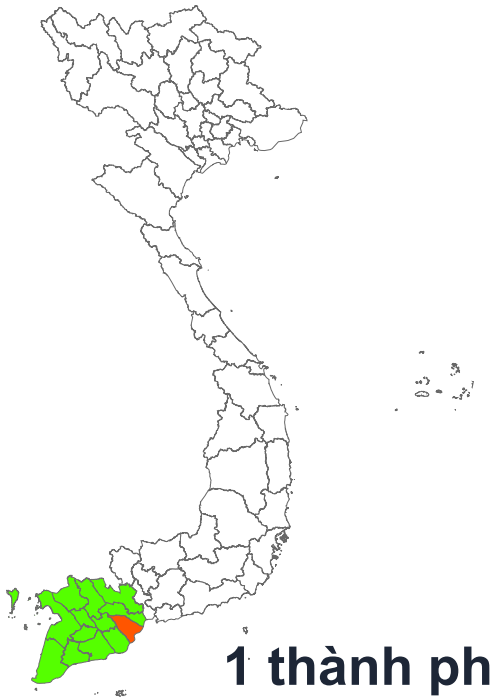
Đánh giá, Dự báo

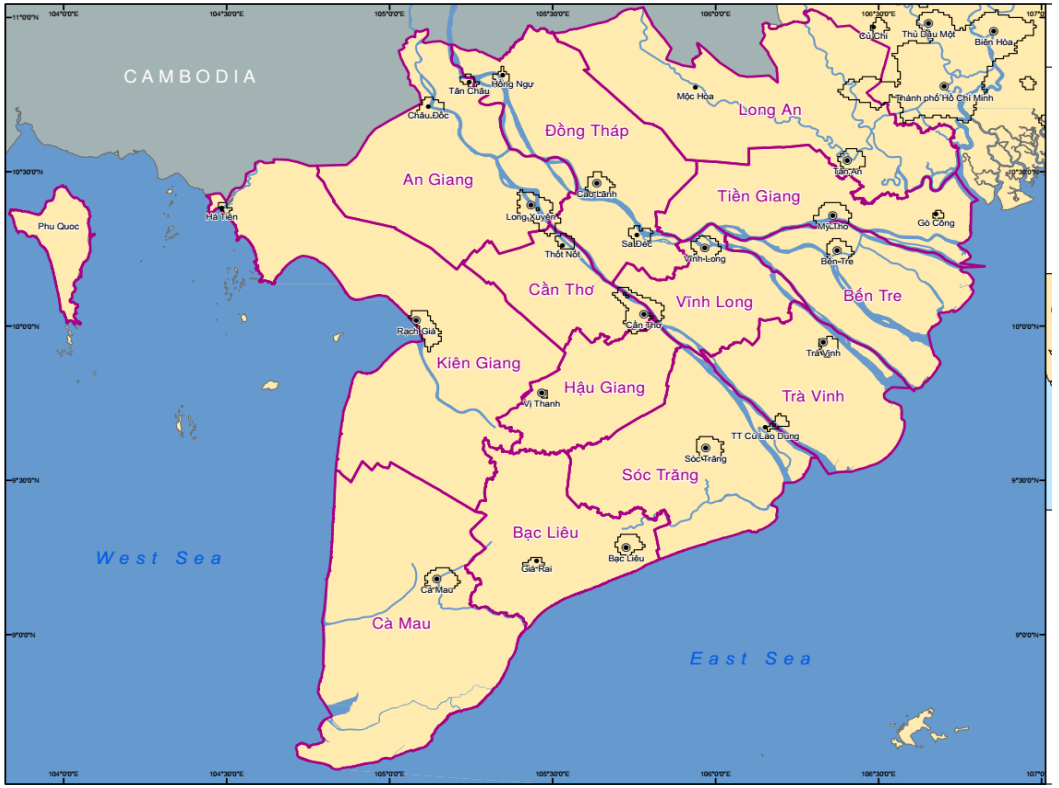
- Ngắn hạn, dài hạn, cấp thời

trang 66, 68, 69 của LA

Đồng bằng sông Cửu Long (ĐBSCL)

Vị trí địa lý





1 thành phố, 12 tỉnh

Diện tích tự nhiên: 40.548,2 km^2 (12,2% cả nước)

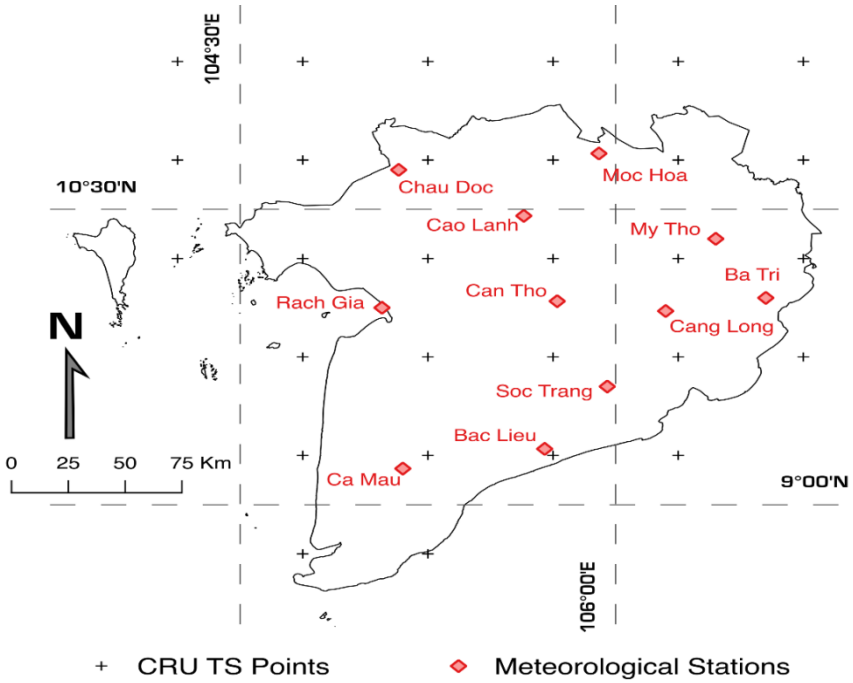
3 loại đất chính (phèn > phù sa > xám)

2 nhánh sông Tiền, Hậu (lưu vực Mekong)

Bờ biển dài 732 km

Dữ liệu thu thập

Dữ liệu	Mô tả	Nguồn thu thập
Dữ liệu quan trắc	Biến số: Mưa, Tmax, Tmin, Tmean, ET, RH Giai đoạn: 1978/1986 - 2015	Đài Khí tượng Thuỷ văn Nam Bộ
Dữ liệu toàn cầu CRU TS4.02 (High-resolution gridded dataset)	Biến số: Mưa, Tmax, Tmin, Tmean Độ phân giải không gian: 0.5º x 0.5º Giai đoạn: 1901-2017, 1951-2017, 1981-2017 (doi: 10.1002/joc.3711)	Climatic Research Unit (University of East Anglia – UK)



04

KẾT QUẢ NGHIÊN CỨU

1. Tập dữ liệu trong biến động theo thời gian

- Các kết quả cụ thể:
- Các phân phối cực đại cho lượng mưa, mực nước, độ mặn, phân bố lượng mưa, biến thiên lượng mưa

1.1 Phân phối cực đại của độ mặn tại Cà Mau

Bước j	μ^j	σ^j	Δ_j	$< k = 10^{-4}$
0	30,77	2,337		
1	30,6953	2,549	0,0458	$>10^{-4}$
2	30,6,778	2,6016	$3{,}07.10^{-3}$	$>10^{-4}$
3	30,6767	2,6049	$1{,}27.10^{-5}$	$<10^{-4}$

Hàm phân phối cực đại độ mặn có dạng:

$$F_1(x) \approx \exp\left\{-\exp\left\{\frac{-(x-30{,}6767)}{2{,}605}\right\}\right\}$$

"The second Vietnam International Applied Mathematics Conference, VIAMC 2017". Dang Kien Cuong, Duong Ton Dam, Duong Ton Thai Duong, "Study on hydrological data of some areas in Mekong Delta from extreme value approach",

1.2 Phân phối cực đại cho mực nước sông Tiền qua 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	$4.1x.10^{-5}$	$<10^{-4}$

Hàm phân phối cực đại mực nước có dạng:

$$F_2(x) \approx \exp\left\{-\exp\left\{\frac{-(x-375.3042)}{69.59}\right\}\right\}$$

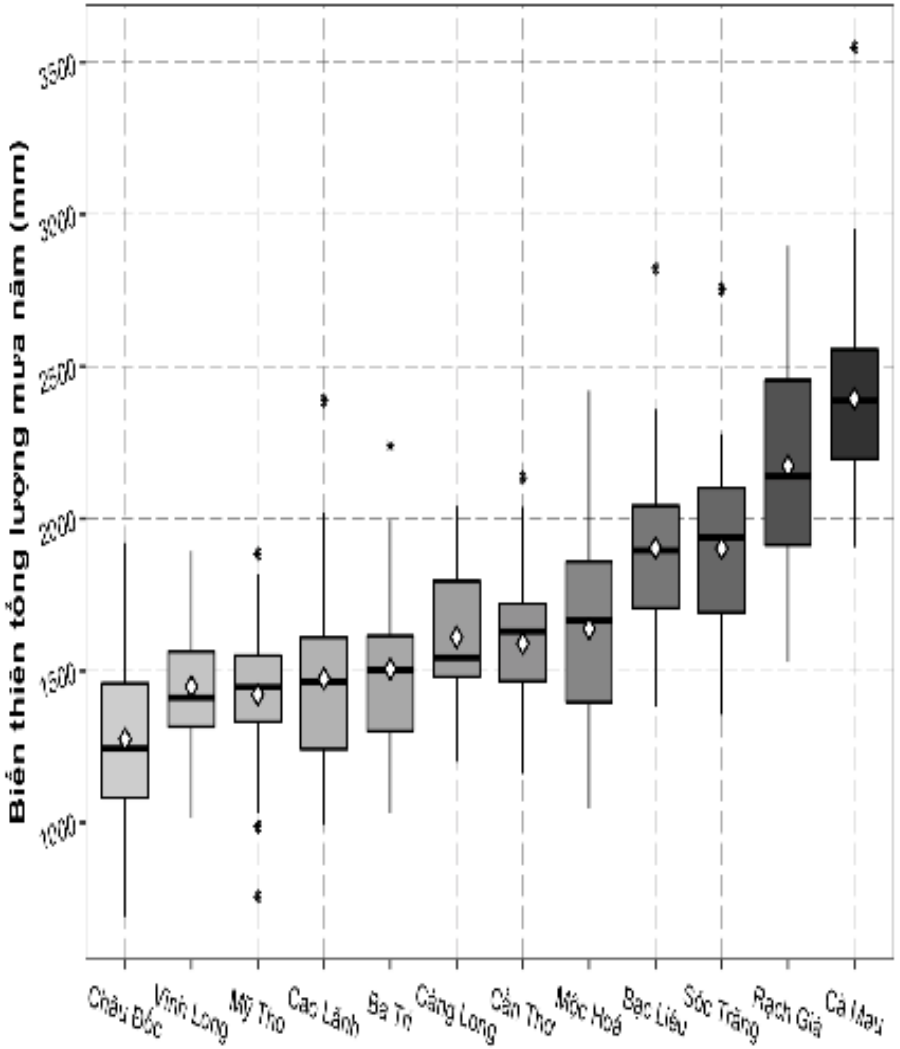
1.3 Phân phối cực đại lượng mưa tại Tân Châu, An Giang

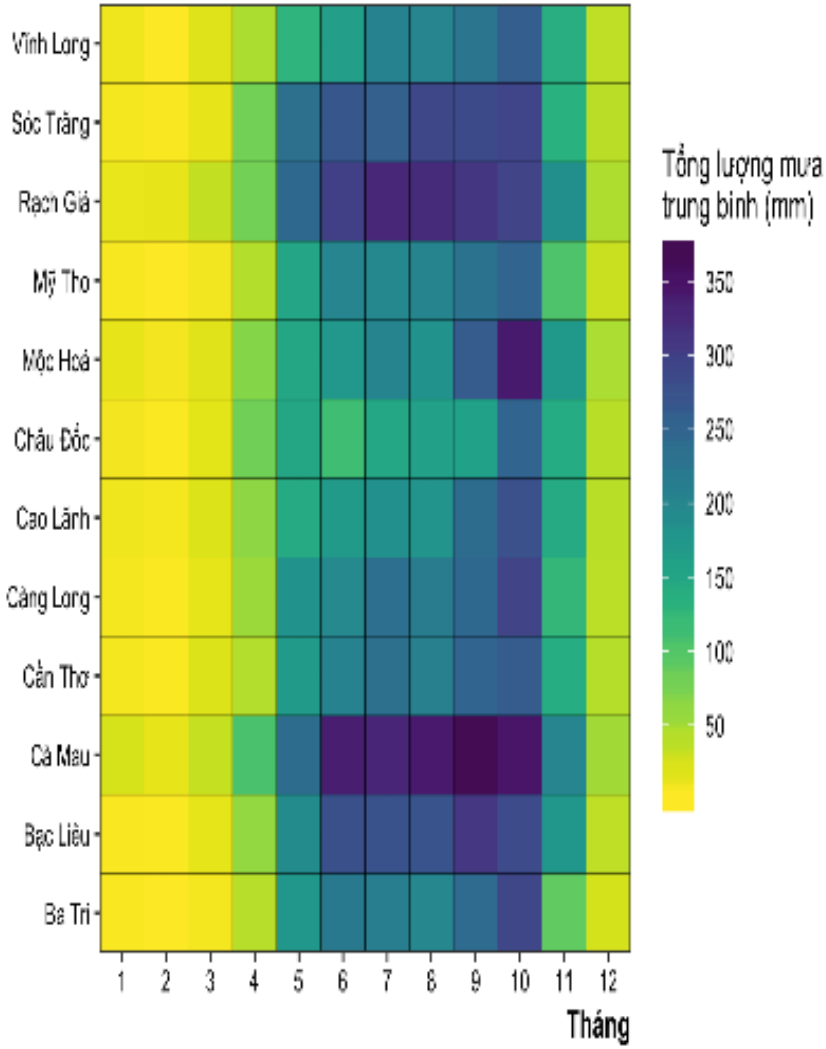
Bước j	μ^j	σ^0	Δ_j	10^{-4}
0	73.4567	16.2603		
1	72.92	17.997	3.303	$>10^{-4}$
2	72.725	18.777	0.647	$>10^{-4}$
3	72.698	18.889	0.0134	$>10^{-4}$
4	72.69766	18.891	$3.8x10^{-4}$	$< 10^{-4}$

Hàm phân phối cực đại lượng mưa có dạng:

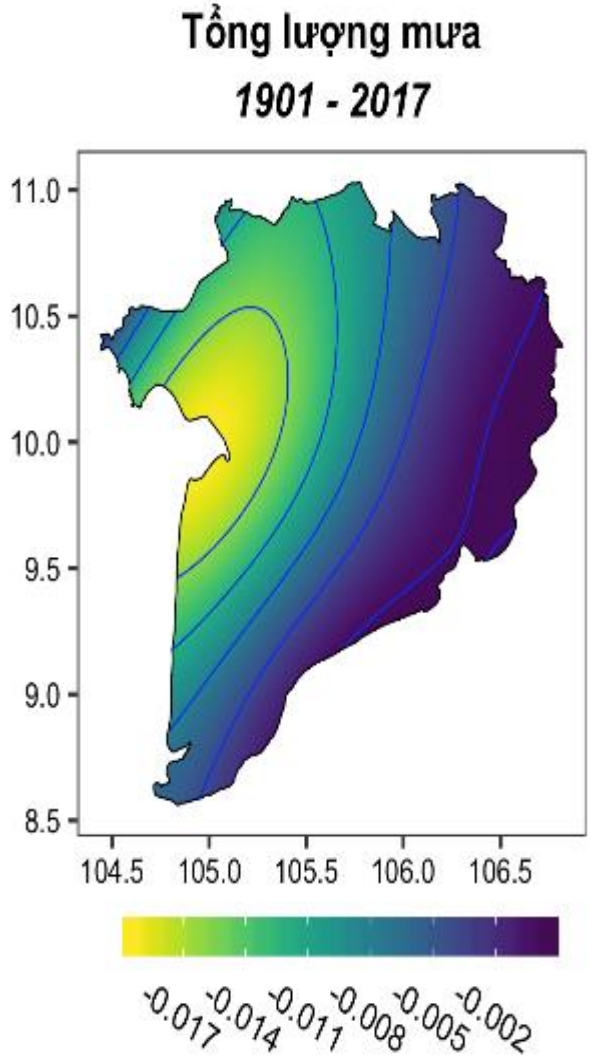
$$F_3(\mathrm{x}) \approx \exp\left\{-\exp\left\{\frac{-(\mathrm{x}-72.69766)}{18.891}\right\}\right\}$$

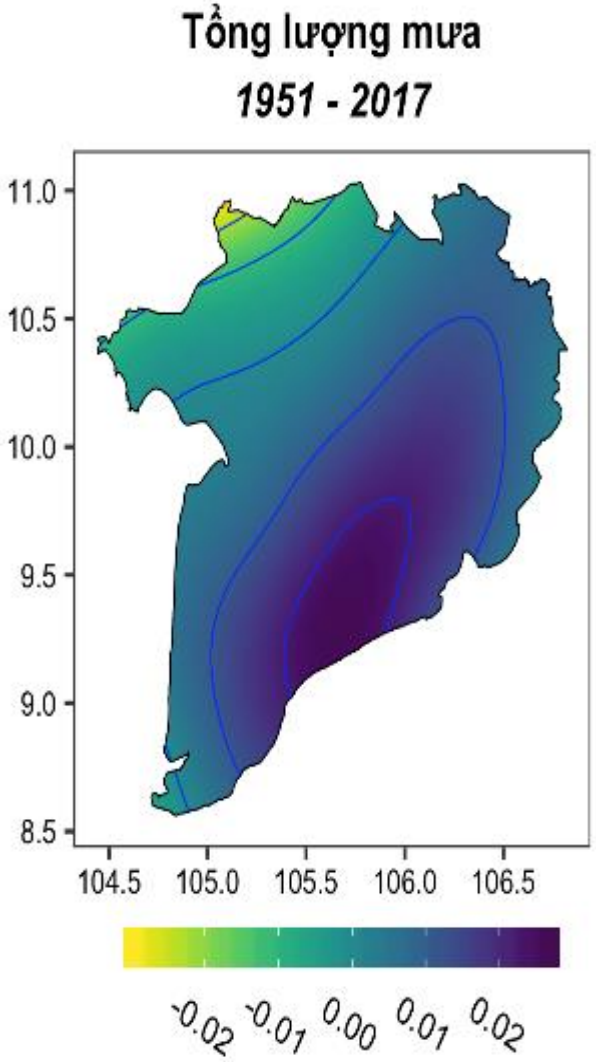
1.4 Phân tích Biến thiên tổng lượng mưa năm và các tháng tại ĐBSCL

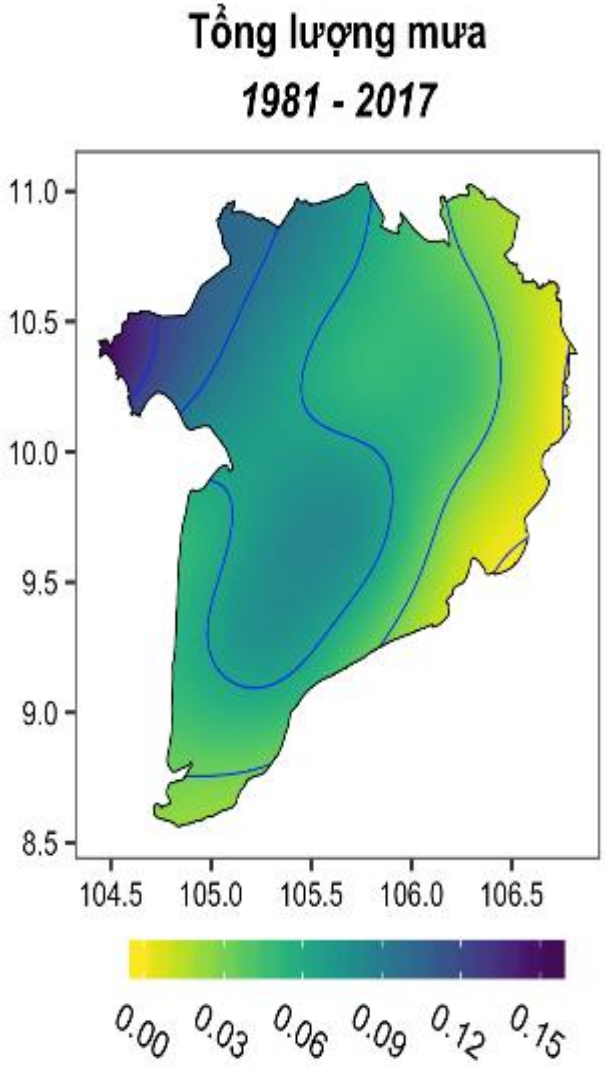




1.5 Phân bố xu thế tổng lượng mưa năm qua các giai đoạn so với thời kỳ chuẩn 1961–1990 (Hệ số dốc Sen: %/năm)







2. Xây dựng *mô hình dự báo* trên cơ sở các quy luật hoặc các đặc tính của tập dữ liệu thực tế và tiến hành huấn luyện, kiểm tra bằng các thuật toán phù hợp

- Thuật toán 1
- Thuật toán 2
- ***Quy luật*** của tập dữ liệu ngẫu nhiên: xác định các tham số trong phân phối Gumbel bằng phương pháp *ước lượng Hợp lý cực đạ*i; Đánh giá và chính xác hóa giá trị các tham số bằng *thuật giải Newton – Raphson*

Kết quả phân tích dự báo

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
2023	37.87	36.85	41.46

3. Phân tích tập dữ liệu bằng *các phương pháp mới*: *tích hợp* toán thống kê *kinh điển và hiện đại.*

- Nghiên cứu được kết quả về mặt lý thuyết, cũng như dựa trên lý thuyết về quy luật để thực hiện ứng dụng dự báo (chỉ ra được quy luật Gumbel trong phân tích GEV cùng các tham số phù hợp).
- Nghiên cứu đã thực hiện phương pháp luận trong việc so sánh các block bootstrap trong đánh giá thống kê, với việc đưa ra được nhận xét với 2 loại tốt MBB, CBB (và 2 loại không tốt trong môt số phân tích, dựa trên tốc độ hội tụ và khoảng cách hội tụ. Phân tích được thực hiện theo các dạng tích hợp của toán thống kê.

KẾT QUẢ

- **Mô hình toán học** để dự đoán các hiện tượng khí hậu, thủy văn, với các minh chứng ở các tỉnh An Giang và Cà Mau ([CT1],[CT4])
- **Thuật toán** khai thác dữ liệu, thực hiện trên dữ liệu khí tượng thủy văn để từ đó dự báo nền nhiệt, xu hướng nhiệt (kết quả trong công trình [CT1], [CT5], [CT6]).
- Nghiên cứu các vấn đề liên quan đến dữ liệu, và xử lý dữ liệu, trong đó lưu ý vấn đề dữ liệu lớn ([CT2]).
- Nghiên cứu về khuếch tán Ito-Levy, ngẫu nhiên, cũng như các bài toán về dữ liệu không đầy đủ, để từ đó hỗ trợ trong các dự báo mặn, lũ ([CT3]).

Định lý 1: Gía trị cực hạn

Cho $\{\xi_i; i = 1,2,\ldots\}$ *là dãy các đại lượng ngẫu nhiên độc lập, chúng thuộc miền hút max của* $H_{\beta_i}(x, \lambda_i, \delta_i) \equiv H_i$ *và* $\{\eta_i; i = 1,2,\ldots\}$, *là dãy các đại lượng ngẫu nhiên độc lập, chúng thuộc miền hút min của* $L_{\beta_i}(x, \lambda_i, \delta_i) \equiv L_i$, *khi đó ta sẽ có:*

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

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

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

[CT1] Dang Kien Cuong, Duong Ton Dam, Duong Ton Thai Duong, and Nguyen Kim Loi, NguyenSonVo, 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.

Phương pháp tách nghiệm

- Phương trình vi phân ngẫu nhiên khuếch tán-nhảy tuyến tính, theo dạng:

$$dX(t) = [\alpha(t)X(t^-) + A(t)]dt + [\beta(t)X(t^-) + B(t)]dW(t) + \int_{R_0} [\gamma(t,z)X(t^-) + G(t,z)]\bar{N}(dt,dz) \quad (1)$$

với một tập các hàm liên tục ngẫu nhiên $\{\alpha, \beta, \gamma, A, B, G\}$ và giả sử rằng quá trình Poisson bù $\bar{N}(t,z)$ độc lập với quá trình Wiener W(t).

- Xuất phát từ các công thức Ito-Hermite cho quá trình Ito-Hermite và cho lớp quá trình Ito-Levy, nghiên cứu trình bày kết quả sự tích hợp vi phân ngẫu nhiên đa chiều cho quá trình Ito-Hermite. Đưa ra phương pháp tách nghiệm để giải phương trình vi phân khuếch tán-nhảy tuyến tính.

Phương pháp tách nghiệm

Giải (1) bằng tách theo dạng tích

$$X(t) = X_1(t^-).X_2(t^-) \qquad (2)$$

$X_1(t)$ là nghiệm của phương trình tuyến tính thuần nhất tương ứng, xác định trong (3)

$X_2(t)$ là nghiệm của phương trình:

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

trong đó $A^*(t); B^*(t); G^*(t,z)$ là những hàm xác định trong (4)

Phương pháp tách nghiệm

- *Phương trình vi phân ngẫu nhiên tuyến tính thuần nhất có dạng:*

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

- *Nghiệm có dạng* (3)

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

Phương pháp tách nghiệm

$$\begin{cases} A^*(t,\omega) = \dfrac{1}{X_1(t^-)}\left[A(t,\omega) - B(t,\omega)\beta(t,\omega) - \int_{R_0} \dfrac{\gamma(t,z,\omega)G(t,z,\omega)}{1+\gamma(t,z,\omega)}\nu(dz)\right] \\ B^*(t,\omega) = \dfrac{B(t,\omega)}{X_1(t^-)} \\ G^*(t,z,\omega) = \dfrac{G(t,z,\omega)}{X_1(t^-)\left(1+\gamma(t,z,\omega)\right)} \end{cases} \quad (4)$$

[CT3] 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

CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ

EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems Research Article **EAI.EU**

Extreme Value Distributions in Hydrological Analysis in the Mekong Delta: A Case Study in Ca Mau and An Giang Provinces, Vietnam

Dang Kien Cuong[1,*], Duong Ton Dam[2], Duong Ton Thai Duong[3], Nguyen Kim Loi[1], Nguyen-Son Vo[4], and Ayse Kortun[5]

[1]Nong Lam University, Ho Chi Minh City, Vietnam
[2]University of Information Technology, Vietnam National University, Ho Chi Minh City, Vietnam
[3]Department of Academic Affair, Vietnam National University, Ho Chi Minh City, Vietnam
[4]Institute of Fundamental and Applied Sciences (IFAS), Duy Tan University, Ho Chi Minh City, Vietnam
[5]Queen's University Belfast, United Kingdom

Abstract

Climate change poses a critical risk to the sustainable development of many regions in Vietnam, especially in the Mekong River. In this paper, we show the specific extreme value distributions of rainfall, flow, and crest of salinity based on the hydrological data from 1975 to 2017 in An Giang and Ca Mau provinces in the Mekong Delta. We also derive a theoretical model and validate its accuracy compared to the empirical data over the years. The results demonstrate that the extremely high flows increase in both magnitude and frequency, while the extremely low ones are projected to occur less often under the climate change. The results can further help the local governments reduce the risk of lack water in dry season, control the salinization, and avoid the threat of flooding in the downstream of the Mekong Delta.

Received on 06 April 2019; accepted on 12 April 2019; published on 13 June 2019
Keywords: Extreme value distribution, Gumbel distribution, max-domain of attraction, maximum likelihood estimation, Newton – Raphson method, Mekong Delta.

doi:10.4108/eai.13-6-2019.159122

1. Introduction

The Mekong Delta is located at the end and in the lowest region of the Mekong river, in the far south of Vietnam. The delta consists of 13 provinces in a triangle form of 3.9 million hectares beginning from Tien Giang province in the east, to An Giang and Kien Giang provinces in the northwest, and down to Ca Mau province in the southernmost tip of Vietnam. The delta people, especially in the Plain of Reeds and the Long Xuyen quadrangle, are affected by seriously seasonal flooding of 3m depth. In addition, in the dry season, over 1.4 million hectares of the coastal regions in the delta are under the effect of salt water intrusion. The Mekong Delta has been facing many challenges due to not only climate change but also hydropower causing more droughts and saltwater intrusion [1].

1. Dang Kien Cuong, Duong Ton Dam, Duong Ton Thai Duong, 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.

CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ

Journal of Mechanics Engineering and Automation 9 (2019) 236-242
doi: 10.17265/2159-5275/2019.07.004

Applications of Bootstrap in Analyzing General Extreme Value Distributions

Dang Kien Cuong[1], Duong Ton Dam[2], Duong Ton Thai Duong[3] and Ngo Thuan Du[4]
1. Information Technology, Nong Lam University, Ho Chi Minh City, Vietnam
2. University of Information Technology, Vietnam National University Ho Chi Minh City, Vietnam
3. Vietnam National University of Ho Chi Minh City, Vietnam
4. CanTho University, Vietnam

Abstract: The bootstrap method is one of the new ways of studying statistical math which this article uses but is a major tool for studying and evaluating the values of parameters in probability distribution. Our research is concerned overview of the theory of infinite distribution functions. The tool to deal with the problems raised in the paper is the mathematical methods of random analysis (theory of random process and multivariate statistics). In this article, we introduce the new function to find out the bias and standard error with jackknife method for Generalized Extreme Value distributions.

Key words: Bootstrap method, time series, block bootstrap jackknife method, generalized extreme value distributions.

1. Basic Framework

1.1 Block Bootstrap Methods for Time Series

Let $X_1, X_2, \ldots, X_n$ be independent and identically random variables with distribution function F. The first step in extreme value theory is to investigate the distribution of $M_n = \max(X_1, X_2, \ldots, X_n)$ as $n \to \infty$. Suppose having sequences of constants $a_n > 0$, $b_n \in R$ such that:

$$P(M_n \le a_n x + b_n) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x), \forall x \in C(G) \quad (1)$$

where $G(x)$ is a non-degenerate distribution function, $C(G)$ is the set of all continuity points of $G(x)$. Limit distribution functions $G(x)$ satisfying Eq. (1) are the well known extreme value three types of distributions (Frechet, Weibull, and Gumbel distributions).

The generalized extreme value (GEV) family of distribution is:

$$G(X) = e^{-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}}, \left\{x: 1 + \xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\} \quad (2)$$

where μ is a location parameter ($\mu \in R$), σ is a scale parameter ($\sigma > 0$), and ξ is the extreme value shape parameter.

Observations $X_1, X_2, \ldots, X_n$ (realisations of a stationary process) are not independent, dependence in time series is relatively simple example of dependent data. Block bootstrap methods for time series data and spatial data have been put forward by Efron, Hall, Radovanov, B., and Marcikie A in Refs. [7, 9-10] among others. See four types of bootstrap methods.

- Moving Block Bootstrap (MBB)

Blocks length l, starting at X_l : $B_l = (X_l, X_{l+1}, \ldots, X_{l+l-1})$. To get a bootstrap sample:

– Draw with replacement $B_1^*, B_{1+1}^*, \ldots, B_k^*$ from $B_1, B_2, \ldots, B_{n-l+1}$.

– Concatenate the blocks $B_1^*, B_2^*, \ldots, B_k^*$ to give the bootstrap sample $X_1^*, X_2^*, \ldots, X_{kl}^*$.

– $l = 1$, corresponds to the classical i.i.d bootstrap.

– Blocks in the MBB may overlap.

- Non-overlapping Block Bootstrap (NBB)

Blocks of length l: $B_1 = (X_1, X_2, \ldots, X_l); B_2 =$

Corresponding author: Dang Kien Cuong, Ph.D. student, lecture, research fields: data science, business intelligence, data mining, education management, machine learning.

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.

CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ

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.

CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ

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“, Second Vietnam international Applied Mathematics Conference (VIAMC 2017), Information and Communications Publishing House, ISBN: 978-604-80-0608-2.

05

KẾT LUẬN

KẾT LUẬN

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 LUẬN

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 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 (như trong bài toán về vật lý lượng tử hoặc trong các vấn đề của kinh tế vĩ mô,…).

CHÂN THÀNH CẢM ƠN QUÝ THẦY CÔ

Góp ý của Phản biện và Hội đồng

1. Những điểm chưa rõ trong luận án, chưa thể hiện rõ khi trình bày: đóng góp của LA, các nghiên cứu trong LA.

- NCS: đã thực hiện theo ý kiến

2. Các công trình công bố, có nội dung giống nhau, ít có liên quan trực tiếp đến luận án, CT6, và CT1, nội dung gần giống nhau, CT3 không có liên quan đến luận án.

- NCS: đã chọn lọc lại CT

3. Tập danh mục công trình chưa chọn lọc, chưa đầy đủ minh chứng, theo quy định, chưa sắp xếp thứ tự

- NCS: đã làm lại tập DMCT theo quy chuẩn

4. Tài liệu tham khảo chưa cập nhật khai phá dữ liệu, khai phá dữ liệu chuỗi thời gian, sắp xếp tài liệu tham khảo chưa chuẩn, chưa có trích dẫn, thiếu trong danh mục.

- NCS: đã bổ sung TLTK

Góp ý của Thầy Cô, Chuyên gia

1. Tổng quan của bài toán, và xác định các mục tiêu nghiên cứu, mô tả liên quan đến dữ liệu, làm nổi bật vấn đề nghiên cứu, tài liệu tham khảo

- NCS: đã thể hiện lại tổng quan, 3 mục tiêu nghiên cứu, bổ sung thêm TLTK: 05 chuỗi thời gian, 08 bootstrap

2. So sánh phương pháp nghiên cứu với một trong những phương pháp khác

- NCS: giải quyết được một số vấn đề của Machine learning như Cluster Analys (K-mean, Clustering Algorithms,…), PCA (Independent Component, Dimension Reduction,…)

Góp ý của Thầy Cô, Chuyên gia

3. Điều chỉnh lại các thuật toán, theo hướng công nghệ thông tin, cũng như thực hiện cách trình bày liên quan đến khoa học máy tính.

- NCS: đã thể hiện lại thuật toán

4. Tinh gọn thêm nữa các công bố.

- NCS: đưa 02 CT không còn liên quan ra ngoài LA

5. Thể hiện rõ kết quả nghiên cứu

- NCS: đã thể hiện rõ 3 kết quả