Introducing the Official Thailand R Local Community: R x TH

Nathakhun Wiroonsri

Organizer

Materials

Slides

Outline

- Our mission and aims
- Our team
- Past contributions and awards
- How to join us
- Q&A
- Alumni Talk
- Workshop

Our Mission

R x TH มีพันธกิจหลักในการรวบรวมผู้ใช้ และ ขยายฐานของผู้ใช้งาน R ในประเทศไทย และสร้างความรู้ความเข้าใจถึงประสิทธิภาพและประโยชน์ของภาษา R ทั้งในภาคการศึกษา และภาคธุรกิจ

R User Groups

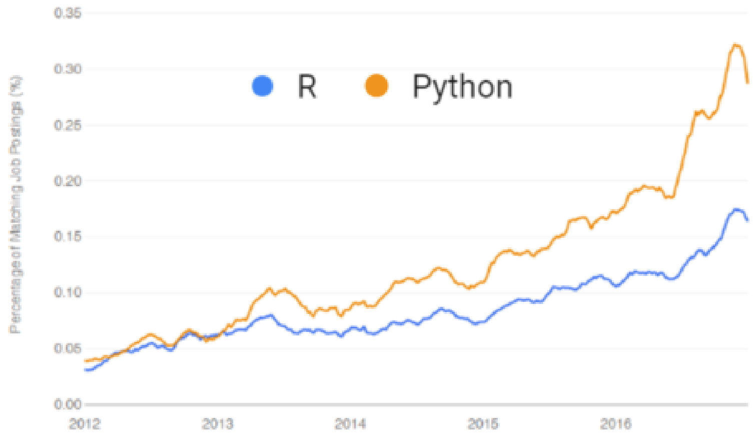
Members	Groups	Countries
75,570	91	39

Our Aims

- เพื่อรวบรวมผู้ใช้งาน R ให้รับรู้ถึงการใช้งานที่กว้างขวางในประเทศไทย
- เพื่อพัฒนาความรู้ให้กับผู้ใช้งาน R
- เพื่อแนะนำ R และขยายฐานผู้งานใหม่
- เพื่อร่วมกันพัฒนางานวิจัย และ ซอฟต์แวร์แพ็คเกจ

Our Plan

- R x TH
  - Newcomers
    - Introduction blogs and vlogs
    - Basic workshops
  - Experienced R users
    - Advanced workshops
    - Research collaboration



https://www.codementor.io/@sayantinideb/r-vs-python-best-programming-language-for-data-science-and-analysis-te05xgx98

Our Team

Nathakhun Wiroonsri (Nat)
Assistant Professor
Organizer

Wasamon Jantai (Mam)
Lecturer
Chula Co-organizer

Onthada Preedasawakul (WW)
Master student
Co-organizer

Noppanon Teangthae (Pure)
Master student
Co-organizer

Our Team

Raywat Tanadkithirun
Member, Assistant Professor

Porntip Dechpichai
Member, Lecturer

Thanet Chitsuphaphan
Member, Lecturer

Chutiphan Charoensuk
Member, Master student

Hafsah Tabassum
Member, Post-doc

Natapon Aeimsri
Member, Master student

Sponsors

Past contributions: Packages

CRAN: Package UniversalCVI | R x TH | Meetup

cran.r-project.org/web/packages/UniversalCVI/index.html

UniversalCVI: Hard and Soft Cluster Validity Indices

Algorithms for checking the accuracy of a clustering result with known classes, computing cluster validity indices, and generating plots for comparing them. The package is compatible with K-means, fuzzy C means, EM clustering, and hierarchical clustering (single, average, and complete linkage). The details of the indices in this package can be found in: C. Alok. (2010) <https://hdl.handle.net/10603/93443>, J. C. Bezdek, M. Moshtaghi, T. Runkler, C. Leckie (2016) <doi:10.1109/TFUZZ.2016.2540063>, T. Calinski, J. Harabasz (1974) <doi:10.1080/03610927408827101>, C. H. Chou, M. C. Su, E. Lai (2004) <doi:10.1007/s10044-004-0218-1>, D. L. Davies, D. W. Bouldin (1979) <doi:10.1109/TPAMI.1979.4766909>, J. C. Dunn (1973) <doi:10.1080/01969727308546046>, F. Haouas, Z. Ben Dhiaf, A. Hammouda, B. Solaiman (2017) <doi:10.1109/FUZZ-IEEE.2017.8015651>, M. Kim, R. S. Ramakrishna (2005) <doi:10.1016/j.patrec.2005.04.007>, S. H. Kwon (1998) <doi:10.1049/EL:19981523>, S. H. Kwon, J. Kim, S. H. Son (2021) <doi:10.1049/ell2.12249>, G. W. Miligan (1980) <doi:10.1007/BF02293907>, M. K. Pakhira, S. Bandyopadhyay, U. Maulik (2004) <doi:10.1016/j.patcog.2003.06.005>, M. Popescu, J. C. Bezdek, T. C. Havens, J. M. Keller (2013) <doi:10.1109/TSMCB.2012.2205679>, S. Saitta, B. Raphael, I. Smith (2007) <doi:10.1007/978-3-540-73499-4_14>, A. Starczewski (2017) <doi:10.1007/s10044-015-0525-8>, Y. Tang, F. Sun, Z. Sun (2005) <doi:10.1109/ACC.2005.1470111>, N. Wiroonsri (2024) <doi:10.1016/j.patcog.2023.109910>, N. Wiroonsri, O. Preedasawakul (2023) <doi:10.48550/arXiv.2308.14785>, C. H. Wu, C. S. Ouyang, L. W. Chen, L. W. Lu (2015) <doi:10.1109/TFUZZ.2014.2322495> and X. Xie, G. Beni (1991) <doi:10.1109/34.85677>.

Version:	1.1.2
Depends:	R (≥ 2.10)
Imports:	e1071, mclust
Published:	2024-03-31
DOI:	10.32614/CRAN.package.UniversalCVI
Author:	Nathakhun Wiroonsri [cre, aut], Onthada Preedasawakul [aut]
Maintainer:	Nathakhun Wiroonsri <nathakhun.wir at kmutt.ac.th>
License:	GPL (≥ 3)

downloads 3134

CRAN: Package BayesCVI | R x TH | Meetup

cran.r-project.org/web/packages/BayesCVI/index.html

BayesCVI: Bayesian Cluster Validity Index

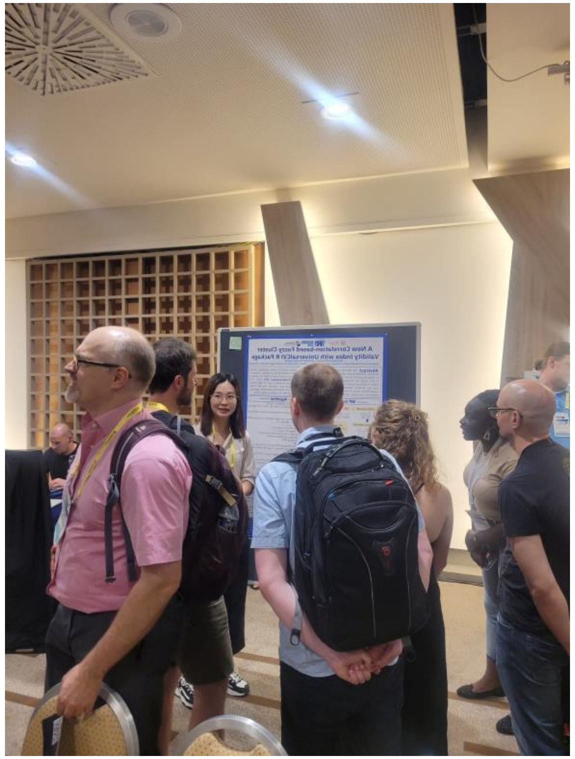
Algorithms for computing and generating plots with and without error bars for Bayesian cluster validity index (BCVI) (O. Preedasawakul, and N. Wiroonsri, A Bayesian Cluster Validity Index, Computational Statistics & Data Analysis, 202, 108053, 2025. <doi:10.1016/j.csda.2024.108053>) based on several underlying cluster validity indexes (CVIs) including Calinski-Harabasz, Chou-Su-Lai, Davies-Bouldin, Dunn, Pakhira-Bandyopadhyay-Maulik, Point biserial correlation, the score function, Starczewski, and Wiroonsri indices for hard clustering, and Correlation Cluster Validity, the generalized C, HF, KWON, KWON2, Modified Pakhira-Bandyopadhyay-Maulik, Pakhira-Bandyopadhyay-Maulik, Tang, Wiroonsri-Preedasawakul, Wu-Li, and Xie-Beni indices for soft clustering. The package is compatible with K-means, fuzzy C means, EM clustering, and hierarchical clustering (single, average, and complete linkage). Though BCVI is compatible with any underlying existing CVIs, we recommend users to use either WI or WP as the underlying CVI.

Version:	1.0.1
Depends:	R (≥ 2.10)
Imports:	e1071, mclust, ggplot2, UniversalCVI
Published:	2024-09-04
DOI:	10.32614/CRAN.package.BayesCVI
Author:	Nathakhun Wiroonsri [aut], Onthada Preedasawakul [cre, aut]
Maintainer:	Onthada Preedasawakul <o.preedasawakul at gmail.com>
License:	GPL (≥ 3)

downloads 1841

Past contributions: Talks

Past contributions: Conferences



- ICSDS 2023 by IMS, Lisbon, Portugal
  - useR! 2024, Salzburg, Austria
- ICSDS 2024 by IMS, Nice, France

Statistics & Data Science
King Mongkut's University of Technology Thonburi

R×TH R consortium

Past contributions: research papers

Computational Statistics and Data Analysis 202 (2025) 108053

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A Bayesian cluster validity index

Onthada Preedasawakul [1], Nathakhun Wiroonsri [*,2]

Mathematics and Statistics with Applications Research Group (MaSA), Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, Thailand

ARTICLE INFO

Keywords:
Cluster analysis
CVI
Dirichlet
Fuzzy c-means
K-means
MRI

ABSTRACT

Selecting the appropriate number of clusters is a critical step in applying clustering algorithms. To assist in this process, various cluster validity indices (CVIs) have been developed. These indices are designed to identify the optimal number of clusters within a dataset. However, users may not always seek the absolute optimal number of clusters but rather a secondary option that better aligns with their specific applications. This realization has led us to introduce a Bayesian cluster validity index (BCVI), which builds upon existing indices. The BCVI utilizes either Dirichlet or generalized Dirichlet priors, resulting in the same posterior distribution. The proposed BCVI is evaluated using the Calinski-Harabasz, CVNN, Davies–Bouldin, silhouette, Starczewski, and Wiroonsri indices for hard clustering and the KWON2, Wiroonsri–Preedasawakul, and Xie–Beni indices for soft clustering as underlying indices. The performance of the proposed BCVI with that of the original underlying indices has been compared. The BCVI offers clear advantages in situations where user expertise is valuable, allowing users to specify their desired range for the final number of clusters. To illustrate this, experiments classified into three different scenarios are conducted. Additionally, the practical applicability of the proposed approach through real-world datasets, such as MRI brain tumor images are presented. These tools are published as a recent R package 'BayesCVI'.

1. Introduction

Cluster analysis is a well-known unsupervised learning tool in statistical and machine learning. It is used to split observations into groups with similar behaviors (refer to the book by James et al. (2023) for a review). Researchers apply cluster analysis to solve problems in various fields, ranging from social science to outer space. There are different types of clustering algorithms, including centroid-based clustering (such as K-means and fuzzy c-means (FCM)), hierarchical clustering (which includes single linkage, complete linkage, group average agglomerative, and Ward's criterion), density-based clustering (such as DBSCAN, DENCLUE, and OPTICS), probabilistic clustering (such as EM), grid-based clustering (such as CLIQUE, MAFIA, ENCLUS, and OptiGrid), and spectral clustering (see Aggarwal and Reddy (2014) for more detailed information on these techniques). Recently, there has been significant attention given to deep learning clustering (Min et al., 2018) and 3D point cloud clustering (Xie et al., 2020; Guo et al., 2021; Chen et al., 2023). Some examples of 3D point cloud techniques include PointNet, PointNet++, DGCNN, and RandLA-Net.

* Corresponding author at: 126 Pracha Uthit Rd., Bangmod, Thung Khru, Bangkok 10140, Thailand.
E-mail addresses: o.preedasawakul@gmail.com (O. Preedasawakul), nathakhun.wir@kmutt.ac.th (N. Wiroonsri).

1 This author has been supported by Petchra Pra Jom Klao Master's Degree Research Scholarship from King Mongkut's University of Technology Thonburi, Grant number: 32/2567.
2 This author has been financially supported by National Research Council of Thailand (NRCT), Grant number: N42A660991 (2023).

https://doi.org/10.1016/j.csda.2024.108053
Received 9 April 2024; Received in revised form 26 July 2024; Accepted 23 August 2024
Available online 30 August 2024

Pattern Recognition 145 (2024) 109910

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Clustering performance analysis using a new correlation-based cluster validity index

Nathakhun Wiroonsri [*]

Mathematics and Statistics with Applications Research Group (MaSA), Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, Thailand

ARTICLE INFO

Keywords:
Clustering algorithm
Cluster validity measure
Data partitions
Correlation
Pattern recognition
Marketing

ABSTRACT

There are various cluster validity indices used for evaluating clustering results. One of the main objectives of using these indices is to seek the optimal unknown number of clusters. Some indices work well for clusters with different densities, sizes, and shapes. Yet, one shared weakness of those validity indices is that they often provide only one optimal number of clusters. That number is unknown in real-world problems, and there might be more than one possible option. We develop a new cluster validity index based on a correlation between an actual distance between a pair of data points and a centroid distance of clusters that the two points occupy. Our proposed index constantly yields several local peaks and overcomes the previously stated weakness. Several experiments in different scenarios, including UCI real-world data sets, have been conducted to compare the proposed validity index with several well-known ones.

1. Introduction

Cluster analysis is one of the most popular unsupervised tools in statistical and machine learning (see [1,2] for an overview of the method.) Researchers apply it to solve problems in various fields, such as medicine, social science, image processing, and biology. Currently, in the world of big data, it plays a very significant role in marketing. One of the most known techniques is to categorize customers to determine an effective business strategy for each group, where the number of groups is unknown (see [3,4] and references therein.) There is no true number of classes in this area, and there is more than one potential option. These facts motivated us to write this paper.

There are two main goals of developing a new theory in cluster analysis. The first one is to develop a new cluster algorithm, such as k-means [5,6], Hierarchical clustering [7,8], fuzzy c-means [9,10], EM algorithm [11], DBSCAN [12], and many more modern techniques, including state-of-the-art deep clustering methods [13,14], point cloud segmentation [15,16] and recently dissimilarity-based random forest clustering [17]. The second goal is to develop a new cluster measurement to evaluate the cluster quality or find the optimal number of clusters. This field has attracted great attention for about a half-century. Several works developed indices for crisp clustering as detailed later in Section 2, and many works developed ones for fuzzy clustering (see [18–25], and [26] for example.) During the process of preparing this manuscript, including the duration of the review process, several works have developed new indices worth mentioning. The work [27] proposed a new index based on a newly defined within cluster distance and its adjustment. The work [28] proposed an object-based cluster validation using densities.

In this work, we introduce a new cluster validity index for crisp clustering based on a correlation coefficient, such as Pearson, Kendall, and Spearman between the distances of each pair of points and each pair of centroids of the clusters that the two points occupy. The correlation introduced here can also check whether the selected clustering result has met a user's expectation.

Several known cluster validity indices handle the cases where each cluster has a different shape, size, and density. The developed validity indices completed their performance checks based on known class data sets to see whether or not they indicate the true number of classes as the optimum. Some of those data sets have multiple classes that are very close as if they are from the same class regardless of the background knowledge. It could be that other factors affect the classes excluded from the data set. The natural question is, "How can we be certain that those indices provide an appropriate number of clusters if the true class is unknown in reality?" That is, switching to another specific application, assuming that the data has a similar pattern as just mentioned, "How can we be certain that the same pattern in this new application should separate into the same number of clusters?"

A problem here is that most indices sometimes give a clear optimum only at a specific number of clusters. Using this information, we tend to

* Correspondence to: King Mongkut's University of Technology Thonburi, 126 Pracha Uthit Rd, Bang Mot, Thung Khru, Bangkok 10140, Thailand.
E-mail address: nathakhun.wir@kmutt.ac.th.

https://doi.org/10.1016/j.patcog.2023.109910
Received 18 March 2021; Received in revised form 30 April 2023; Accepted 22 August 2023
Available online 24 August 2023

Methodology and Computing in Applied Probability (2023) 25:19
https://doi.org/10.1007/s11009-023-10004-7

Normal Approximation for Fire Incident Simulation Using Permanental Cox Processes

Dawud Thongtha [1] · Nathakhun Wiroonsri [1]

Received: 21 April 2022 / Revised: 26 September 2022 / Accepted: 29 December 2022 /
Published online: 11 February 2023

Abstract
Estimating the number of natural disasters benefits the insurance industry in terms of risk management. However, the estimation process is complicated due to the fact that there are many factors affecting the number of such incidents. In this work, we propose a Normal approximation technique for associated point processes for estimating the number of natural disasters under the following two assumptions: 1) the incident counts in any two distinct areas are positively associated and 2) the association between these counts in two distinct areas decays exponentially with respect to distance outside some small local neighborhood. Under the stated assumptions, we extend previous results for the Normal approximation technique for associated point processes, i.e., the establishment of non-asymptotic L^1 bounds for the functionals of these processes. Then we apply this new result to permanental Cox processes that are known to be positively associated. Finally, we apply our Normal approximation results for permanental Cox processes to Thailand's fire data from 2007 to 2020, which was collected by the Geo-Informatics and Space Technology Development Agency of Thailand.

Keywords Correlation inequality · Cox process · Local dependence · Random fields · Natural disaster · Positive association

AMS 2010 Subject Classifications Primary 60F05 · 82B30 · 60G60

1 Introduction

Probability and statistical models have been widely used in most business sectors, including the insurance business. They are major tools for estimating loss, claim severity, claim count and claim probabilities, as these factors significantly affect their business operations. Since claim occurrences are used in determining policy premiums and are usually not predictable, claim simulation and prediction are some of the main tools for handling the

✉ Nathakhun Wiroonsri
nathakhun.wir@kmutt.ac.th

1 Mathematics and Statistics with Applications Research Group, Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, Thailand

Springer

Past contributions: research papers

Smart Agricultural Technology 8 (2024) 100435

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

ELSEVIER

Determination of crop water requirements and potential evapotranspiration for sustainable coffee farming in response to future climate change scenarios

Usa Wannasingha Humphries [a,1], Muhammad Waqas [b,c,1], Phyo Thandar Hlaing [b,c], Angkool Wangwongchai [a], Porntip Dechpichai [a,*]

[a] Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT), Bangkok,10140, Thailand
[b] The Joint Graduate School of Energy and Environment (JGSEE), King Mongkut's University of Technology Thonburi (KMUTT), Bangkok, 10140, Thailand
[c] Center of Excellence on Energy Technology and Environment (CEE), Ministry of Higher Education, Science, Research and Innovation, Bangkok, Thailand

ARTICLE INFO

Keywords:
Climate change
Coffee
Downscsling
Crop water requirement
Potential evapotranspiration

ABSTRACT

Climate change (CC) is causing a significant threat to agriculture, a sector complicatedly tied to natural resources. Changes in precipitation patterns, atmospheric water content, and rising temperatures intensely affect global agriculture, especially in tropical regions. In this intense CC scenario, potential evapotranspiration (PET) and crop water requirement (CWR) are critical components of agricultural water management. This study evaluates the future impact of CC on precipitation, CWR, and PET in different provinces of Thailand's northern and northeastern regions. Three bias correction methods (Delta (DT), Empirical Quantile Mapping (EQM), and Quantile Mapping (QM)) were employed for precipitation downscaling from the CanESM5 CMIP6-GCM across selected 13 coffee farms with different coffee species. Arabica and Robusta coffee were carefully selected for this analysis. The DT method demonstrated superiority, exhibiting lower RMSE and higher correlation coefficients than EQM and QM. Farm-specific assessments illuminated water demand's complex dynamics during critical growth stages, showcasing variable CWR and PET. During the blooming stage in N-F1, CWR ranged from 16.7 to 33.7 mm/stage, highlighting the variability in water needs. Projected CC impacts on Arabica and Robusta coffee farms in Chiang Rai and Sisaket presented challenges, emphasizing farm-specific strategies to address potential water deficits or surpluses during critical growth phases. Projected 2023, 2028, and 2033 precipitation demonstrated incongruities with CWR and PET. The findings emphasize the crucial role of farm-specific adaptive strategies in mitigating the impacts of changing precipitation patterns on coffee cultivation.

1. Introduction

Climate change (CC) challenges humanity, affecting various sectors such as the agriculture industry and services [1,2]. The central emphasis when evaluating the pivotal dimension of 21st-century CC revolves around examining significant changes in global warming. Various research endeavors have substantiated the increase in the mean surface temperature. [3]. According to the Intergovernmental Panel on CC (IPCC) report, temperatures on land and ocean surfaces exhibited an average increase ranging from 0.3 to 0.6 °C between 1900 and 1995, with a discernible upward trend of approximately 0.2–0.3 °C over 40 years [4]. Over centuries, the Earth's climate has experienced a gradual warming, marked by a 0.74 °C increase in average temperature [1]. The IPCC (2018) predictions suggest a continued upward trajectory in global warming [5]. To access these predictions, Global climate models (GCMs) are extensively employed across various disciplines, including

List of Abbreviations: CC, Climate Change; CWR, Crop Water Requirement; PET, Potential Evapotranspiration; DT, Delta method; EQM, Empirical Quantile Mapping; QM, Quantile Mapping; GCM, Global Circulation Model; RMSE, Root Mean Square Error; MBE, Mean Bias Error; IPCC, Intergovernmental Panel on Climate Change; CMIP6, Coupled Model Intercomparison Project Phase 6; CMIP5, Coupled Model Intercomparison Project Phase 5; ETc, Crop Evapotranspiration; Eto, Reference Evapotranspiration; Kc, Crop Coefficient; FAO, Food and Agriculture Organization; CDF, Cumulative Distribution Function; TMD, Thai Meteorological Department; ESGF- WCRP, Earth System Grid Federation – World Climate Research Programme.

* Corresponding author.
E-mail address: porntip.dec@kmutt.ac.th (P. Dechpichai).

[1] These authors contributed equally to this work.

Quality & Quantity (2023) 57:509–539
https://doi.org/10.1007/s11135-022-01366-0

Do elderly want to work? Modeling elderly's decision to fight aging Thailand

Krittiya Kantachote [1] · Nathakhun Wiroonsri [2]

Accepted: 26 February 2022 / Published online: 29 March 2022

Abstract

Thailand has entered into an aging society since the year 2000. Using the 2017 Survey of the Older Persons in Thailand collected by National Statistical Office of Thailand, this study uses cross tabulation, random forest with variable importance measure and lasso logistic regression to examine factors that have effects on the elderly's decision to remain in the labor market after retirement. This study reveals that these following variables: age, education level, healthcare eligibility, marital status, health condition, total assets, gender, residential type, percent of elderly in the household, and number of children have strong influences on an elderly's desire to continue work. By knowing which factors contribute to the elderly wish to continue work in the market, this research allows for future prediction of the labor market that can accommodate elderly in Thailand. Our final models of random forest and lasso logistic regression provide prediction accuracy of 68.19 and 69.58 percent on the elderly's desire to work, respectively. This study has a significant impact as policy-makers can utilize our models in predicting elderly's desire to work after retirement age and design a labor market that can accommodate elderly in Thailand in the future.

Keywords Aging society · Elderly retirement · Lasso logistic regression · Machine learning · Random forest · Thailand

1 Introduction

1.1 Elderly

Countries around the world consider an individual to be an elderly at a different age. Elderly are divided into different stages, for example, 65-74, 75-84, and those above 85. According to the Thailand Act on the Elderly (2003), it determines that an elderly is an individual who is 60 or above (Ministry of Social Development and Human Security 2003).

✉ Nathakhun Wiroonsri
nathakhun.wir@kmutt.ac.th

[1] Department of Sociology, Srinakharinwirot University, Bangkok, Thailand

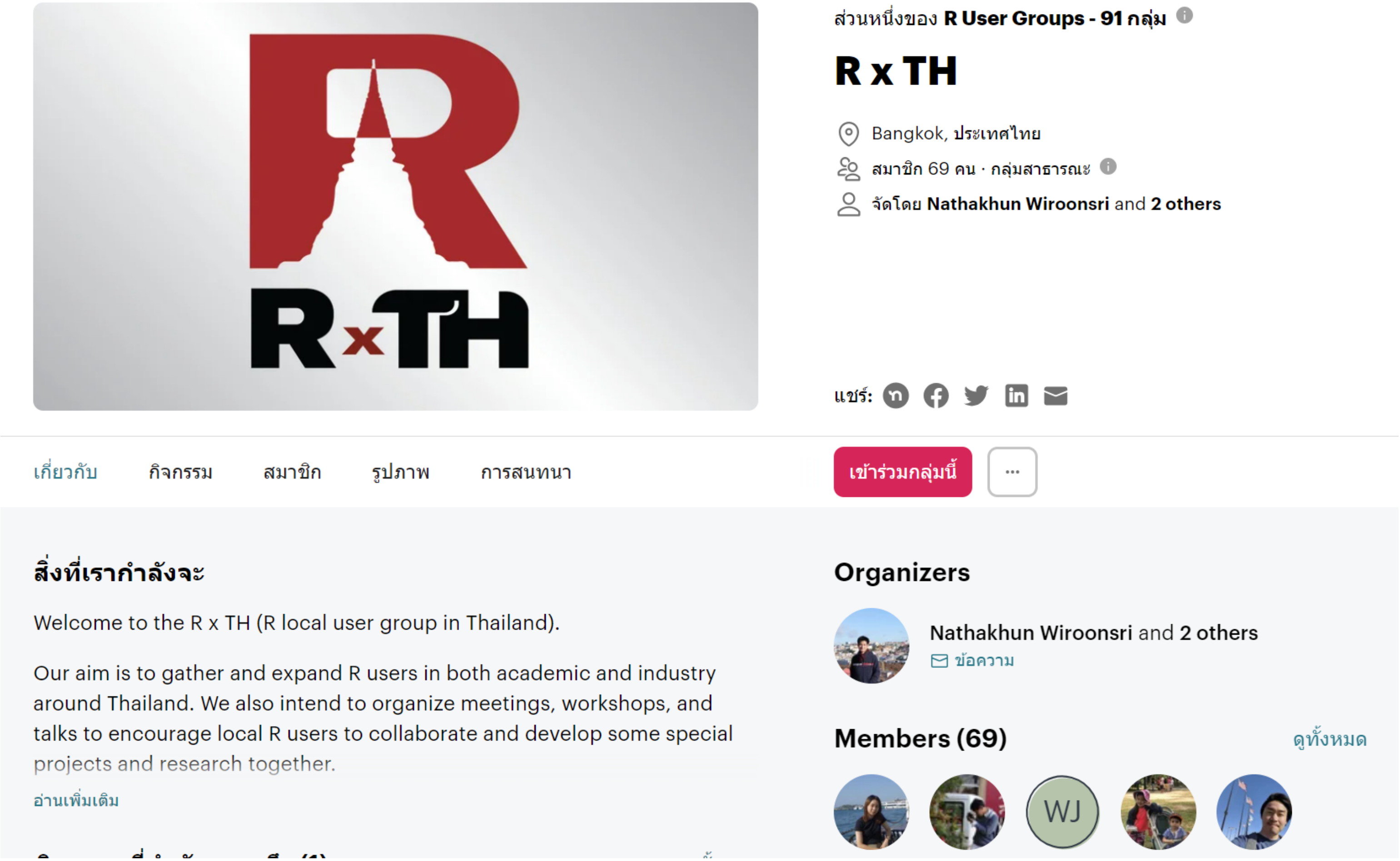
[2] Mathematics and Statistics with Applications Research Group, Department of Mathematics, King Mongkut's University of Technology Thonburi, 126 Pracha Uthit Rd, Bang Mot Thung Khru, Bangkok 10140, Thailand

Springer

Past contributions: Awards

Please join us

https://www.meetup.com/r-x-th/



Evaluation

THANK YOU.

ANY QUESTIONS?