

Hamid R. Arabnia
Leonidas Deligiannidis
Farzan Shenavarmasouleh
Farid Ghareh Mohammadi
Soheyla Amirian (Eds.)

Communications in Computer and Information Science

2969

Emerging Trends in Data Science and Cyber Warfare

12th International Conference on Computational Science
and Computational Intelligence, CSCI 2025
Las Vegas, NV, USA, December 3–5, 2025
Proceedings, Part III

Part 3

Communications in Computer and Information Science

2969

Series Editors

Gang Li , *School of Information Technology, Deakin University, Burwood, VIC, Australia*

Joaquim Filipe , *Polytechnic Institute of Setúbal, Setúbal, Portugal*

Zhiwei Xu, *Chinese Academy of Sciences, Beijing, China*

Rationale

The CCIS series is devoted to the publication of proceedings of computer science conferences. Its aim is to efficiently disseminate original research results in informatics in printed and electronic form. While the focus is on publication of peer-reviewed full papers presenting mature work, inclusion of reviewed short papers reporting on work in progress is welcome, too. Besides globally relevant meetings with internationally representative program committees guaranteeing a strict peer-reviewing and paper selection process, conferences run by societies or of high regional or national relevance are also considered for publication.

Topics

The topical scope of CCIS spans the entire spectrum of informatics ranging from foundational topics in the theory of computing to information and communications science and technology and a broad variety of interdisciplinary application fields.

Information for Volume Editors and Authors

Publication in CCIS is free of charge. No royalties are paid, however, we offer registered conference participants temporary free access to the online version of the conference proceedings on SpringerLink (<http://link.springer.com>) by means of an http referrer from the conference website and/or a number of complimentary printed copies, as specified in the official acceptance email of the event.

CCIS proceedings can be published in time for distribution at conferences or as post-proceedings, and delivered in the form of printed books and/or electronically as USBs and/or e-content licenses for accessing proceedings at SpringerLink. Furthermore, CCIS proceedings are included in the CCIS electronic book series hosted in the SpringerLink digital library at <http://link.springer.com/bookseries/7899>. Conferences publishing in CCIS are allowed to use our online conference service (Meteor) for managing the whole proceedings lifecycle (from submission and reviewing to preparing for publication) free of charge.

Publication process

The language of publication is exclusively English. Authors publishing in CCIS have to sign the Springer CCIS copyright transfer form, however, they are free to use their material published in CCIS for substantially changed, more elaborate subsequent publications elsewhere. For the preparation of the camera-ready papers/files, authors have to strictly adhere to the Springer CCIS Authors' Instructions and are strongly encouraged to use the CCIS LaTeX style files or templates.

Abstracting/Indexing

CCIS is abstracted/indexed in DBLP, Google Scholar, EI-Compendex, Mathematical Reviews, SCImago, Scopus. CCIS volumes are also submitted for the inclusion in ISI Proceedings.

How to start

To start the evaluation of your proposal for inclusion in the CCIS series, please send an e-mail to ccis@springer.com

Hamid R. Arabnia · Leonidas Deligiannidis ·
Farzan Shenavarmasouleh ·
Farid Ghareh Mohammadi · Soheyla Amirian
Editors

Emerging Trends in Data Science and Cyber Warfare

12th International Conference on Computational Science
and Computational Intelligence, CSCI 2025
Las Vegas, NV, USA, December 3–5, 2025
Proceedings, Part III

Editors

Hamid R. Arabnia 
University of Georgia
Athens, GA, USA

Leonidas Deligiannidis 
Wentworth Institute of Technology
Boston, MA, USA

Farzan Shenavarmasouleh 
Medialab Inc.
Lawrenceville, GA, USA

Farid Ghareh Mohammadi 
Verify Radiology Images Consultants
Houston, TX, USA

Soheyla Amirian 
Pace University
New York, NY, USA

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-032-28459-4 ISBN 978-3-032-28460-0 (eBook)
<https://doi.org/10.1007/978-3-032-28460-0>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2027

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Unless otherwise expressly agreed by the Publisher in writing, the Publisher does not authorize use or reproduction of all or any part of this work in any manner for the purpose of training, fine-tuning, evaluating, or developing artificial intelligence models, technologies, or systems. In accordance with Article 4(3) of the EU Directive 2019/790 on copyright and related rights in the Digital Single Market, the Publisher expressly reserves all rights with respect to this work under the text and data mining exception of the directive.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

It is our great pleasure to present this collection of research papers from the 12th International Conference on Computational Science and Computational Intelligence (CSCI 2025). This volume, titled “Emerging Trends in Data Science and Cyber Warfare”, comprises selected papers accepted in the Research Tracks on Big Data and Data Science (CSCI-RTBD); and Cyber Warfare, Cyber Defense, & Cyber Security (CSCI-RTCW). The conference was held in Las Vegas, Nevada, USA, from December 3 to December 5, 2025.

The CSCI 2025 International Conference brought together papers from a diverse array of communities, including researchers from universities, corporations, and government agencies. Accepted papers are published by Springer Nature, and the proceedings showcase solutions to key challenges in various critical areas of Computational Science and Computational Intelligence.

Computational Science (CS) is the study of addressing problems that are impossible to solve (or difficult to solve) without computers. CS can be considered to be the bridge between computer science and other sciences. The field is interdisciplinary by nature and includes the use of advanced computing capabilities to understand and solve complex problems. In short, CS is the science of using computers to do science. Computational Intelligence (CI) is the study of computational methods that exhibit intelligence. These methods adapt to changing environments and changing goals. There is a significant overlap between the fields of CI and Artificial Intelligence (AI). However, there is also a difference: AI techniques often involve top-to-bottom methods (i.e., methods to reach solutions are imposed from the top) whereas CI techniques often involve bottom-up methods (i.e., solutions emerge from unstructured beginnings). An important part of CI includes a set of Nature-inspired computational approaches to address complex problems for which traditional methods are infeasible. Computational Science and Computational Intelligence both share the same objective: finding solutions to difficult problems. However, the methods to reach the solutions are different. The main objective of the CSCI Conference is to facilitate increased opportunities for cross-fertilization across CS and CI.

Considering the above broad outline, the CSCI 2025 International Conference was composed of the following focused Research Tracks:

Artificial Intelligence (CSCI-RTAI); Big Data and Data Science (CSCI-RTBD); Computational Science (CSCI-RTCS); Computational Intelligence (CSCI-RTCI); Computational Biology (CSCI-RTCB); Cyber Warfare, Cyber Defense, & Cyber Security (CSCI-RTCW); Signal & Image Processing, Computer Vision & Pattern Recognition (CSCI-RTPC); Smart Cities and Smart Mobility (CSCI-RTSC); Education - CS & CE (CSCI-RTED); Health Informatics and Medical Systems (CSCI-RTHI); Mobile Computing, Wireless Networks, & Security (CSCI-RTMC); Software Engineering (CSCI-RTSE); Internet of Things & Internet of Everything (CSCI-RTOT); Social Network Analysis, Social Media, & Mining (CSCI-RTNA); Cloud Computing and Data Centers

(CSCI-RTCC); Parallel & Distributed Computing (CSCI-RTPD); and Large Language Models (LLM). The scope of each track can be found at: <https://www.american-cse.org/csci2025/topics> and <https://www.american-cse.org/LLM2025/topics>

An important objective of the CSCI 2025 International Conference and its associated Research Tracks was to foster opportunities for cross-fertilization between the fields of Computational Science and Computational Intelligence. The CSCI Conference is deeply committed to promoting diversity and eliminating discrimination, both in its role as a conference organizer and as a service provider. Our goal is to create an inclusive culture that respects and values differences, promotes dignity, equality, and diversity, and encourages individuals to reach their full potential. We are also dedicated, wherever possible, to organizing a conference that represents the global community. We sincerely hope that we have succeeded in achieving these important objectives.

The Steering Committee and the Program Committee would like to extend their gratitude to all the authors who submitted papers for consideration. CSCI 2025 received submissions from 48 countries, with approximately 48% of them coming from outside the United States. Each submitted paper underwent a rigorous peer-review process, with at least two experts (an average of 2.5 referees per paper) evaluating the submissions based on originality, significance, clarity, impact, and soundness. In cases where reviewers' recommendations were contradictory, a program committee member was tasked with making the final decision, often consulting additional referees for further guidance. The CSCI Conference followed the guidelines of COPE (Committee on Publication Ethics):

- Typical submissions underwent a single-blind peer review process, in which the authors remain unaware of the identities of the reviewers, while the reviewers are informed of the authors' identities.
- Papers authored by one or more members of the program committee, including co-chairs, were subjected to a double-blind peer review process, ensuring that neither the authors nor the reviewers were aware of each other's identities or affiliations.

The Research Track on Big Data and Data Science (CSCI-RTBD) received 98 submissions, of which 19 papers were accepted, resulting in a paper acceptance rate of 19%; we selected only 13 papers for this volume. The Research Track on Cyber Warfare, Cyber Defense, & Cyber Security (CSCI-RTCW) received 184 submissions, of which 39 papers were accepted, resulting in a paper acceptance rate of 21%; we selected only 27 papers for this volume. A total of 40 papers are included in this volume.

We are deeply grateful to the many colleagues who contributed their time and effort to organizing the CSCI 2025 Conference. In particular, we extend our thanks to the members of the Program Committee, the Steering Committee, and the referees. We would also like to express our appreciation to the primary sponsor of the conference, the American Council on Science & Education. The list of members of the Program Committee for each track can be found at: <https://www.american-cse.org/csci2025/committees>

We extend our heartfelt gratitude to all the speakers and authors for their valuable contributions. We would also like to thank the following individuals and organizations for their support: the staff at the Luxor Hotel (an MGM property, conference/meeting department) and the staff of Springer for their assistance in various aspects of the event.

We are pleased to present the proceedings of CSCI 2025 (selected papers of CSCI-RTBD and CSCI-RTCW Research Tracks). These proceedings represent a collection of outstanding research contributions that reflect the diversity and depth of work in Data Science and Cyber Warfare.

December 2025

Hamid R. Arabnia
Leonidas Deligiannidis
Farzan Shenavarmasouleh
Farid Ghareh Mohammadi
Soheyla Amirian



Q-Gini: A Multidimensional Inequality Index with Isolation of Its Causal Drivers

Ray Hashemi¹(✉), Omid M. Ardakani², Rushmila Shabneen¹, and Azita G. Bahrami³

¹ Department of Computer Science, Georgia Southern University, Savannah, GA, USA
{rayhashemi, rs24222}@georgiasouthern.edu

² Department of Economics, Georgia Southern University, Statesboro, GA, USA
oardakani@georgiasouthern.edu

³ IT Consultation, Savannah, GA, USA

Abstract. The *univariate* Gini (U-Gini) index is the gold standard for quantifying the inequality for a population in reference to a socioeconomic factor such as income, wealth, health, education, etc. However, the U-Gini's focus on a single factor (dimension) does not fully capture inequality in a population. We introduce a new Gini index (Q-Gini) delivering the inequality for multi-socioeconomic factors. The Minimum Squared error approach is used to validate the Q-Gini index by establishing its performance superiority over the U-Gini index. A Modified Neighborhood Clustering is introduced to reduce the multidimensions of the dataset into a set of neighborhoods and then validate the new Q-Gini for the neighborhoods. We also isolate the causal impacts of socioeconomic factors on the Q-Gini. Such isolation is essential because the causal drivers represent targets for intervention by policymakers. The results revealed the performance superiority of Q-Gini over U-Gini in both cases of the non-reduced and reduced dimensions of the dataset. Also, the isolated causal impacts validated the Q-Gini's sensitivity to these fundamental socioeconomic forces.

Keywords: Gini Index · Multidimension Gini Index · Causal impact

1 Introduction

The use of the Gini index (to be precise, univariate Gini—U-Gini) is common for quantifying the inequality for a given attribute (dimension), A , of a population. The U-Gini index for A is calculated using formula (1),

$$U - Gini_A = 1 - L/E \quad (1)$$

where L is the shaded area under the distribution of the attribute A (Lorenza curve), and E is the area under the equality line for the attribute A (triangle of abc) [1]. This line has a 45-degree angle with both the X-axis and the Y-axis to represent the equality of (x, y) coordinates for every point on this line, Fig. 1.

The U-Gini index is the gold standard for quantifying the inequality within a population in reference to a socioeconomic factor such as income, wealth, health, education,

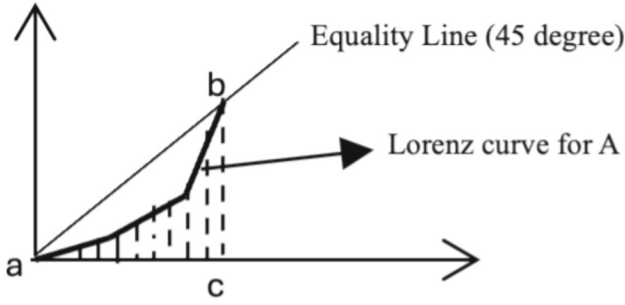


Fig. 1. Equality Income Line and Distribution Curve for Attribute A.

etc. [2, 3]. However, the U-Gini's focus on a single factor (dimension) does not fully capture inequality in a population, because inequality manifests across several interconnected socioeconomic attributes [4, 5]. In this paper, we propose a Gini index capable of delivering the inequality in reference to multi-socioeconomic factors—a comprehensive Gini (Q-Gini) index.

We also investigate the dimensions' reduction effect on Q-Gini. In addition, we isolate the causal impacts of socioeconomic factors on the Q-Gini because (a) these factors exhibit high correlations among themselves, which can lead to spurious inference, and (b) causal drivers represent targets for intervention by policymakers.

The goal of this research effort has four prongs: (i) Introducing the Q-Gini index as a representative of inequality for a multidimensional dataset, (ii) Validating the Q-Gini by comparing its performance with U-Gini, (iii) investigating and validating the effect of the dimension reduction on Q-Gini, and (iv) Delivering the causal drivers of Q-Gini in the dataset.

The rest of the paper is organized as follows: The previous works are the subject of Sect. 2. The Methodology is introduced in Sect. 3. The Empirical results are presented in Sect. 4. The Conclusion and future research are covered in Sect. 5.

2 Previous Works

Inequality measurement is foundational in socioeconomic analysis. It is a known fact that the U-Gini index is used to quantify the inequality for a given socioeconomic index. This is also another known fact that inequality is influenced by more than one socioeconomic index. Therefore, one can claim that the U-Gini index does not fully capture inequality. There is a need for multidimensional extensions of the Gini index. Although such extensions exist [6, 7], their computational validation remains underdeveloped. In particular, challenges remain in aligning composite indices with univariate benchmarks.

The Gini index [1] measures single-attribute Seminal extensions to multiple dimensions include Normative Multidimensional Ginis [6], which generalized the Gini to multiple attributes but omitted computational validation. Distance-Based Approaches [7] proposed Euclidean aggregation yet relied on arbitrary weights. Clustering-Aware Indices [8] advocated attribute grouping but offered no empirical validation protocol. Self-Organizing Map (SOM) for Socioeconomic Clustering [9] established SOMs as

topology-preserving tools for high-dimensional data reduction. Applications show efficacy in inequality contexts (e.g., Deboeck for wealth segmentation [10]). Validation of Composite Indices pioneered Monte Carlo sensitivity analysis [11], while Maggino emphasized predictive accuracy as a validation criterion, both of which are underutilized in Gini literature [12].

We propose a comprehensive inequality measure through hierarchical clustering. A novel Mean Squared Error (MSE)-driven comparison of the proposed measure against traditional univariate Gini is employed.

3 Methodology

The details for the four prongs of the goal are presented in the following four subsections.

3.1 Q-Gini Index for a Multidimensional Dataset

To calculate the Q-Gini index for a multidimensional dataset, we build a tree structure out of the dimensions of the dataset in such a way that the root of the tree, as a dimension, be the representative of all dimensions of the dataset. The Gini index for the root dimension is considered the Q-Gini index for the multidimensional dataset.

The tree is a dendrogram [13] that is built using the algorithm Dendrogram, Fig. 2. We explain the algorithm through an example.

Let us assume that we have five attributes of $A_1, \dots, A_5$ in a dataset, and the goal is to calculate the Q-Gini index for the dataset. By applying the algorithm Dendrogram, we generate a tree out of the attributes. A possible instance of such a tree is shown in Fig. 3

Each node of the dendrogram (not leaves) represents a new attribute that is made up of its two children. To decide which two attributes to merged], the total Euclidean distances' square between every pair of attributes is calculated. The pair with the smallest distance is the winner. The reader needs to know that values for the pair of attributes must be of the same scale.

Therefore, all attributes of the dataset D , are normalized to the same scale of 1 to 9, prior to distance calculation using formula (2).

$$Nv_i = 1 + 8 \frac{v_i - \min(A_j)}{\max(A_j) - \min(A_j)} \quad (2)$$

where Nv_i is the normalized value of v_i in attribute A_j .

Let us assume that a pair of attributes (A_1, A_2) is the one with the smallest distance. The new attribute of B is generated, and the values for B are the average of the corresponding values in A_1 and A_2 . Attribute B replaces A_1 and A_2 in the dataset.

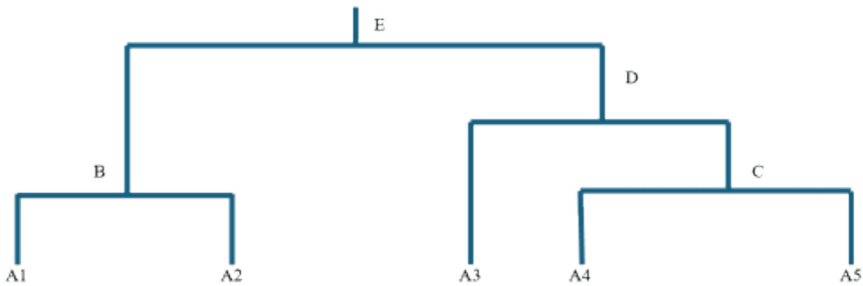
The process of finding a pair of attributes with the minimum distance, averaging them into a new attribute, and replacing the pair with the new attribute continues until the multidimensional dataset is converted into a one-dimensional dataset. This dimension is the root of the tree (E), and it represents all dimensions of the dataset.

Algorithm DENDROGRAM**Input :** n-dimensional dataset D, with attributes of $A_1, \dots, A_n$ **Output:** Dendrogram of D**Method:**

```

1- Normalize every attribute in D
2- Trans= Transposition (D);
3- Min =10;
4- While (|Trans| >1)
5-   For i= 1 to |Trans|-1
       For j = i+1 to |Trans|
           If Min > Euclidean distance(rowi, rowj);
               Then Min=Euclidean distance(rowi, rowj);
                   Pair = (rowi, rowj)
       End;
   End;
   New_row = Average of corresponding values of rows in the Pair.
   Remove from Trans the rows kept in the Pair;
   Insert New_row into the Trans;
   Min=10;
End;
6-End of Method;

```

Fig. 2. Algorithm DENDROGRAM**Fig. 3.** A Possible Dendrogram for the Five Attributes of $A_1, \dots, A_5$.**3.2 Validation of the Q-Gini**

File D is given with $n + 1$ attributes. One attribute (γ) in D carries the U-Gini value obtained for the population, in reference to one attribute, A, where A is not γ . We remove the attribute γ from the file D and make a new file for γ called D'. Now, D' and D have 1 and n attributes, respectively, and both have the same number of records.

Let TR and TS be a training set and a test set created out of the file (D). The |TR| and |TS| are equal to u_1 and u_2 , respectively such that $u_1 > u_2$ and $(u_1 + u_2) = |D|$. Both files have n attributes. We calculate the U-Gini index for the attribute A_i (for $i = 1$ to n)

of the TR. As a result, we have n U-Gini indices of $(G^{tr}_1, \dots, G^{tr}_n)$ –one per attribute. We also calculate Q-Gini for the TR file using the Dendrogram algorithm. The outcome is a Q-Gini index (Q^{tr}) for the entire attributes of TR file. The Information Gain (IG) for each attribute of TR is $(IG^{tr}_i = Q^{tr} - G^{tr}_i)$ (for $i = 1$ to n).

We select the same $u1$ records from D' and calculate the Q-Gini (Q^y) and information gain for the selected $u1$ records of D' , $(IG^{tr}_i = Q^{tr} - G^{tr}_i)$ (for $i = 1$ to n).

For file TS, we repeat the calculation of U-Gini for each attribute $(G^{ts}_1, \dots, G^{ts}_n)$, Q-Gini for the entire TS file (Q^{ts}), and information gain for each attribute $(IG^{ts}_i = Q^{ts} - G^{ts}_i)$ (for $i = 1$ to n).

The minimum squared errors in reference to TS are obtained using formulas (3) and (4).

$$MSE_{TR-Q-Gini} = \frac{1}{n} \left[\sum_{i=1}^n (IG^{tr}_i - IG^{ts}_i)^2 \right] \quad (3)$$

$$MSE_{g-Q-Gini} = \frac{1}{n} \left[\sum_{i=1}^n (IG_i - IG^{ts}_i)^2 \right] \quad (4)$$

The Min ($MSE_{TR-Q-Gini}$, $MSE_{g-Q-Gini}$) determines the superiority of one index over the other one. If the minimum squared error belongs to $MSE_{TR-Q-Gini}$, then the validity of Q-Gini is established.

3.3 Investigating and Validating the Effect of the Dimensions' Reduction of Q-Gini

The main question is whether a reduced version of an n -dimensional dataset of TR can be used, and if so, how the Q-Gini of the reduced dataset is comparable with Q^{tr} and Q^y . To answer this question, we reduce the n -dimensional dataset into an m -dimensional dataset, where $m < n$. The reduction is done by using a selected clustering method that satisfies two conditions of natural clustering and inclusion. The natural clustering condition says that the clustering must take place naturally and not by force. Therefore, those clustering techniques for which the number of final set of clusters, C , must be announced prior to the actual clustering process, perform a forced clustering because each record of the dataset has no choice but to be in one of the C clusters and the natural number of clusters (based on the hidden patterns) within the dataset may be higher, or lower than C .

The inclusion condition says that the centroid of cluster, c_i , must not be used for deciding on inclusion of a new record into c_i because the similarity is not established between two individual records of D , but between a record of D and the centroid of a cluster (a fabricated record).

A neighborhood system [14–17] with some modifications satisfies both conditions of natural clustering and inclusion. The modified version (*Modified-Neighborhood clustering*) This clustering has two parts: *Neighborhood building* and *Neighborhood merging* that are presented in the following two subsections.

Neighborhood Building. The dataset H , which is a copy of the TR file, has n dimensions $(A_1, \dots, A_n)$ and $u1$ records $(R_1, \dots, R_{u1})$. Each attribute of this file has a representative that is kept separately from H . The representative is a binary string with a length of n , and each bit is a position holder for attributes of A_1 to A_n . Initially, the binary string for attribute A_i contains one “1” value in the bit i of its string. We also need the dataset Shadow- H , which is a duplication of H . Attribute A_i in Shadow- H is considered a shadow of the attribute A_i in H , and it has $u1$ values of $(v_{1i}, \dots, v_{u1i})$.

Two attributes of $A_i (v_{1i}, \dots, v_{u1i})$ and $A_j (v_{1j}, \dots, v_{u1j})$ of H are considered neighbors, if they satisfy the following conditions:

For $k = 1$ to n :

1. $0 \leq |v_{ki} - v_{kj}| \leq \delta$, where $\delta \geq 1$ and
2. $\sum |v_{ki} - v_{kj}| > 0$

The real variable δ determines the closeness level of two attributes [14, 15]. The neighboring is established based on the closeness of the two attributes and not by closeness of A_i or A_j to a fabricated attribute (centroid of the cluster). However, we replace the shadow of A_i (in Shadow- H) with the averaging of the corresponding attribute values of A_i and A_j . (The shadow of A_i is a centroid of the two neighbors and is kept for later use. This centroid does not play any role in the formation of the final clusters.

We start finding neighbors for A_i of D (for $i = 1$ to $u1$). If the attribute A_j (where $j = i + 1$) of D is the neighbor of A_i , then A_j is in the neighborhood of A_i , and it is denoted as $N(A_i) = \{A_j\}$. The $N(A_i)$ is recorded by setting bit j to one in the bitstring of A_i . The bitstring of attribute A_j does not change. Therefore, each attribute in D has its own neighborhood, and all these neighborhoods are created using the same closeness level. All the neighborhoods are established prior to neighborhood merging.

For $N(A_i)$ the shadow of A_i is the centroid of this neighborhood. However, inclusion of attribute A_b , for example, in $N(A_i)$ is not influenced by the shadow of A_i , but by the closeness of the two attributes of A_i and A_b in H .

Neighborhood Merging. Two neighborhoods of $N(A_i)$ and $N(A_b)$ may merge if the following condition is satisfied:

$$\text{Bitstring}(A_i) \text{C} \text{Bitstring}(A_b)^3 d$$

The integer variable d represents a threshold for the number of common attributes between $N(A_i)$ and $N(A_b)$. If a merge happens between $N(A_i)$ and $N(A_b)$, then.

- a. $N(A_i) = N(A_i) \cup N(A_b)$,
- b. $\text{Bitstring}(A_i) = \text{Bitstring}(A_i) \cup \text{Bitstring}(A_b)$,
- c. $\text{Bitstring}(A_b) = 0$, and
- d. $\text{Shadow}(A_i) = \text{Averaging corresponding values for Shadows of } A_i \text{ and } A_b$.

Upon completion of possible merges for all neighborhoods, those neighborhoods with at least one “1” in their Bitstrings are considered a cluster, and there is one shadow attribute for each one of these neighborhoods that can represent the centroid of the neighborhood.

The outcome for applying the Modified Neighborhood clustering on the n -dimensional TR file is an m -dimensional dataset (named file Z). The number of attributes

for Z is m and $|Z| = |TR|$. We repeat the calculation of the U-Gini index for each attribute A_j (for $j = 1$ to m) in Z , generating $(G_1, \dots, G_m)$ —one per attribute. One can easily assign the new U-Gini values of $(G_1, \dots, G_m)$ to the n attributes of the file TR by assigning the same G_i value to those attributes of TR that belong to the cluster, C_i . We also use the Dendrogram algorithm to calculate Q-Gini for Z — Q^Z . In addition, the information gain for Z can be calculated and it is $(IG^Z_i = Q^Z - G_i)$ (for $i = 1$ to n). Furthermore, the MSE_z is calculated using formula 5.

$$MSE_z = \frac{1}{n} \left[\sum_{i=1}^n \left(IG_i^z - IG_i^{ts} \right)^2 \right] \quad (5)$$

The $\text{Min}(MSE_{tr}, MSE_y, MSE_z)$ determines the most superior Gini index and its validation.

3.4 Delivering the Causal Drivers of Q-Gini in the Dataset

To determine whether the relationship between dimensions and the Q-Gini is causal or merely correlational, we use the Instrumental Variable (IV). The IV method addresses endogeneity in models, which occurs when an explanatory variable is correlated with the error term. This correlation can bias the estimation of causal relationships. The IV method solves this by introducing an instrument. The IV is a variable that is correlated with the endogenous variable and uncorrelated with the error term.

Instrumental Variables Estimation. To examine potential causal effects of covariates on inequality, we specify a series of exactly identified Instrumental Variables (IV) models. The main challenge is the endogeneity of a regressor X , where $E[X\varepsilon] \neq 0$, which violates the standard linear model assumption and leads to biased Ordinary Least Squares (OLS) estimates. For each endogenous regressor X and a chosen instrument W , we estimate a Two-Stage Least Squares (2SLS) model for each outcome $G \in Q^{TR}, Q^Z$ as First Stage: $X = \pi_0 + \pi_1 W + v$, Second Stage: $G = \beta_0 + \beta_1 X_b + \varepsilon$, where X_b are the fitted values from the first-stage regression. The key identifying assumption is the exclusion restriction: the instrument W affects the outcome G only through its effect on the endogenous variable X [18].

To guard against heteroskedasticity, which is common in cross-sectional economic data, we use heteroskedasticity-robust standard errors using the HC1 estimator [19]. The covariance matrix is given by,

$$\widehat{Var}_{HC1}(\hat{\beta}_{2SLS}) = (\hat{X}^T \hat{X})^{-1} \left(\sum_{i=1}^n \hat{\varepsilon}_i^2 \hat{x}_i \hat{x}_i^T \right) (\hat{X}^T \hat{X})^{-1},$$

where $\hat{X}$ is the matrix of second-stage regressors (including the constant and the fitted values $\hat{X}$), and $\hat{\varepsilon}_i$ are the 2SLS residuals.

For each model, we report several diagnostics. The null hypothesis of a weak instrument is tested using the first-stage F-statistic on the excluded instrument Z . While a common rule-of-thumb suggests $F > 10$ indicates sufficient strength [20], this test is

technically undefined under exact identification, as the standard error of the 2SLS estimate is a function of the first-stage error and becomes infinite if the instrument is irrelevant. Results are reported where calculable. A Wu-Hausman test [21, 22] is conducted to test for endogeneity. It tests the null hypothesis that the OLS estimator is consistent and efficient against the alternative that the IV estimator is consistent. A low p-value suggests significant endogeneity, justifying the use of IV.

4 Empirical Results

In our empirical study, we employ an n-dimensional dataset from the U.S. Census Bureau, specifically the 2022 American Community Survey (ACS 5-year estimates). This dataset provides a granular view of socioeconomic variables at the county level across the United States. The variables include Median Household Income (MHI), Per Capita Income (PCI), Total Earnings (TE), Poverty Levels (POV), U-Gini Index of Income Inequality (U-GINI), Unemployment Rate (UR), Educational Attainment (EA), Median Gross Rent (MGR), Median Housing Value (MHV), Internet Access (INT), Health Insurance Coverage (HIC), and demographic breakdowns by race and ethnicity: Percentage of White Population (WP), Percentage of Black or African American Population (BP), and Percentage of Hispanic or Latino Population (HP).

After normalizing the data using formula (2), we calculated the minimum error differences using formulas (3), (4), and (5). The winner was the use of Q-Gini produced for TR produced from the dendrogram. The Q-Gini for the reduced dataset Z obtained by applying the Modified Neighborhood clustering was the second winner.

We employ Instrumental Variable (IV) regression to estimate the causal relationships between socioeconomic factors and the U-Gini. This addresses the endogeneity of key variables. Instead of ordinary linear models, which may provide biased estimates due to endogeneity, IV regression uses external instruments to isolate the exogenous variation in each independent variable. For each socioeconomic variable X_i , we selected a corresponding instrument Y_i that is correlated with X_i but not with the error term ϵ_i , in the Gini model. The IV regression model for each socioeconomic variable is given by

$$G_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad (6)$$

where G_i is the U-Gini index for population i , β_0 is the intercept, β_1 is the coefficient of X_i , and ϵ_i is the error term. The coefficients obtained from these models indicate the direction and magnitude of the relationship between each independent variable and income inequality.

Our objective is to determine whether the relationship between regressors and the Comprehensive Gini is causal or merely correlational. The IV method addresses endogeneity in models, which occurs when an explanatory variable is correlated with the error term. This correlation can bias the estimation of causal relationships. The IV method solves this by introducing an instrument. The IV is a variable that is correlated with the endogenous variable and uncorrelated with the error term.

This approach provides a framework to find a causal relationship between variables and the Comprehensive Gini. Each defines an endogenous variable and its corresponding

instrument, chosen based on their correlations from the correlation matrix (Table 1). For example, MHI is instrumented by PCI due to their strong correlation ($r = 0.88$) and PCI's relatively low correlation with the Gini coefficient ($r = -0.09$), making PCI a valid instrument that meets the relevance and exogeneity conditions. Similarly, TE is instrumented by UR, given their high correlation ($r = 0.98$), ensuring that UR strongly predicts TE but has a weak direct association with Gini ($r = 0.13$), addressing concerns of endogeneity. POV is instrumented by TE, as their correlation ($r = 0.95$) allows TE to serve as a strong predictor of POV while mitigating the risk of bias in the Gini model. In a similar vein, UR is instrumented by EA, and MGR is instrumented by MHV, both selected for their high correlations with the endogenous variables and minimal direct impact on the Gini coefficient. This method reduces the potential for omitted variable bias and reverse causality. The map of endogenous variables to their instruments is pre-specified based on theoretical considerations and observed correlations.

Table 1. Mapping of Endogenous Regressors to Instrument

Endogenous variable (X)	Instrument (W)
Median household income	Per capita income Total
earnings	Unemployment rate
Poverty rate	Total earnings
Unemployment rate	Educational attainment
Median gross rent	Median home value
Road infrastructure	Educational attainment
Black population %	Hispanic population %
Health insurance coverage	Educational attainment
Hispanic population %	Hispanic population %

The IV regression results indicate the causal impact of various socioeconomic factors on the U-Gini, Table 2. Each coefficient reflects the magnitude and direction of the causal effect of the respective variable on Gini. MHI exhibits a significant negative effect (-0.07 , $p\text{-value} < 0.001$), implying that higher median incomes reduce income inequality, confirming the causal relationship between household income and inequality. In contrast, TE and UR have strong positive effects (0.34 , $p\text{-value} < 0.001$), suggesting that increased earnings and unemployment rates exacerbate income inequality. Similarly, POV and INT also show significant positive impacts on Gini, indicating their role in widening inequality. BP and HP have large negative coefficients (-1.17 and -1.79 , respectively), with highly significant p -values, pointing to the potential disparities in income distribution across different racial demographics.

The following table (Table 3) presents the results of IV regressions for factors, providing estimates of causal relationships between these variables. Each dependent variable (MHI, TE, POV, UR, INT, BP, HIC, HP) is instrumented using other related variables. The rows with zero estimates or not statistically significant estimates are

Table 2. Causal Impacts on Gini

Variable	Estimate	Std. Error	Variable	Estimate	Std. Error
MHI	−0.07	0.01	INT	0.28	0.04
TE	0.34	0.04	BP	−1.17	0.32
POV	0.31	0.04	HIC	0.30	0.04
UR	0.34	0.05	HP	−1.79	0.43

removed. PCI has a strong positive effect on MHI (estimate: 0.89), indicating that higher per capita income leads to a higher median household income. POV, UR, and INT negatively affect MHI, suggesting that as poverty, unemployment, and lack of internet access rise, household income tends to fall. HIC shows a significant positive impact on MHI (estimate: 7.06), indicating that access to health insurance correlates with higher household income, likely reflecting better job quality and benefits in higher-income households.

POV has significant negative impact on TE (estimate: -0.14). Both HIC and INT show strong positive effects on total earnings. MHI and TE have strong negative effects on poverty (estimates: −0.01 and −2.27), confirming the expected causal relationship where increases in income and earnings reduce poverty levels. UR, INT, and HIC all have positive effects on POV, suggesting that higher unemployment and lack of services exacerbate poverty. POV and TE have significant positive effects on UR, with TE indicating that higher earnings can also reflect higher participation in the labor force, thus increasing unemployment in some cases due to labor market friction. INT and HIC reduce unemployment.

TE significantly increases INT (estimate: 1.04). UR has a negative impact on INT (estimate: −0.16), indicating that as unemployment rises, access to internet infrastructure diminishes, potentially due to affordability issues. POV and INT have significant positive impacts on BP, which may indicate higher vulnerability of this demographic to poverty and technological exclusion. MHI has a negative effect on BP (estimate: −0.46). TE and UR both have positive effects on HIC. POV and HIC have large positive effects on HP. UR and INT show negative effects.

Table 3 says that the variables that have a causal effect on UR, for example, are TE, POV, UR, and HIC. The table also says that the variables that have a causal effect on TE are UR, POV, and INT. Clearly, we have a loop between UR and TE. This is a big problem because the graph that will represent the causality among the variables must be an acyclic directed graph.

Table 3. Causal Estimates of Socioeconomic Variables

Dependent Variable	IV	Estimate	Std. Error
MHI	PCI	0.89	0.01
MHI	POV	−1.25	0.17
MHI	UR	−0.76	0.2
MHI	INT	−4.27	0.34
MHI	BP	−0.03	0.01
MHI	HIC	7.06	0.68
MHI	HP	0.01	0.001
TE	POV	−0.14	0.001
TE	UR	0.04	0.01
TE	INT	0.25	0.01
TE	HIC	0.84	0.01
POV	MHI	−0.01	0.001
POV	TE	−2.27	0.06
POV	UR	0.63	0.02
POV	INT	0.82	0.03
POV	HIC	1.8	0.06
UR	TE	0.42	0.06
UR	POV	0.44	0.01
UR	INT	−0.45	0.03
UR	HIC	0.53	0.06
INT	TE	1.04	0.03
INT	POV	0.21	0.01
INT	UR	−0.16	0.01
INT	HIC	−0.03	0.04
BP	MHI	−0.46	0.03
BP	TE	2.67	2.13
BP	POV	0.78	0.53
BP	UR	−0.6	0.63
BP	INT	3.42	1.05
BP	HIC	−5.65	2.11
BP	HP	−0.14	0.01
HIC	TE	0.85	0.01

(continued)

Table 3. (continued)

Dependent Variable	IV	Estimate	Std. Error
HIC	UR	0.05	0.01
HP	MHI	−0.45	0.03
HP	POV	3.23	0.62
HP	UR	−3.38	0.74
HP	INT	−5.77	1.23
HP	BP	−0.19	0.02
HP	HIC	6.73	2.48

5 Conclusion and Future Research

We introduced the Q-Gini index to quantify inequality across multiple socioeconomic dimensions. The Q-Gini provides a comprehensive evaluation by hierarchically aggregating all dimensions of a dataset into a single inequality dimension. This aspect is missing in the univariate U-Gini. Validation of the Q-Gini used an MSE protocol, demonstrating its superior performance over the conventional U-Gini index.

We also studied effect of dimensionality reduction on inequality measurement. By introducing modified-neighborhood system that satisfies natural clustering and inclusion conditions, we reduced the multidimensional data to representative neighborhoods. Q-Gini calculated on this reduced dataset ranked second among the most accurate measures in our validation framework.

The isolation of causal drivers of the Q-Gini index recognized high correlations among socioeconomic factors, which can lead to spurious inferences. We used IV models to distinguish correlation from causal impact and identify effective policy targets. The IV results confirm that variables such as median household income, total earnings, and poverty levels not only associate with but cause changes in multidimensional inequality. This validates the Q-Gini’s sensitivity to these fundamental socioeconomic forces.

For future research, we are considering developing a belief network. This would model the probabilistic relationships between socioeconomic variables and inequality.

Acknowledgments. Our special thanks go to our graduate students, Ms. Noushin Gauhar and Mr. Jeffrey Young for helping us when we needed.

References

1. Gini, C.: Variabilità e mutabilità [Variability and mutability]. Reprinted *Memorie di Metodologica Statistica* **1**, 211–382 (1912)
2. Atkinson, A.B.: On the measurement of inequality. *J. Econ. Theory* **2**(3), 244–263 (1970)
3. Ceriani, L., Verme, P.: The origins of the Gini index: extracts from *Variabilità e Mutabilità* (1912) by Corrado Gini. *J. Econ. Inequality* **16**(3), 421–443 (2018)

4. Decancq, K., Lugo, M.A.: Weights in multidimensional indices of well-being: an overview. *Economet. Rev.* **32**(1), 7–34 (2013)
5. UNDP. Human Development Report 2020. United Nations Development Program (2020)
6. Gajdos, T., Weymark, J.A.: Multidimensional generalized Gini indices. *Econ. Theor.* **26**(3), 471–496 (2005)
7. Andreoli, F., Zoli, C.: From unidimensional to multidimensional inequality: a review. *J. Econ. Surv.* **34**(4), 754–785 (2020)
8. Lugo, M.A.: Comparing multidimensional indices of inequality: methods and application. In: *Inequality and poverty* (Vol. 14, pp. 213–236). Emerald Group Publishing Limited (2007)
9. Kohonen, T.: Essential of the self-organizing map. *Neural Netw.* **37**, 52–65 (2013)
10. Deboeck, G.J.: Financial applications of self-organizing maps. *Neural Netw. World* **8**(2), 213–241 (1998)
11. Saisana, M., Saltelli, A., Tarantola, S.: Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *J. R. Stat. Soc. Ser. A* **168**(2), 307–323 (2005)
12. Maggino, F.: Developing indicators and managing the complexity. In: Maggino, F. (ed.) *Complexity in society: From indicators construction to their synthesis*, pp. 87–114. Springer (2017)
13. Swofford, D.L., Olsen, G.J., Waddell, P.J., Hillis, D.M.: “Phylogenetic inference”. In: Hillis, D.M., Moritz, C., Mable, B.K. (eds.) *Molecular Systematics*, 2nd edition. Sunderland, MA: Sinauer, pp. 407–514 (1996)
14. Yao, Y.Y.: Rough Sets, neighborhood Systems and Granular computing. *Proceedings of 1999 IEEE Canadian Conference on Electrical and Computer Engineering*, vol. 3, pp. 1553–1558 (1999)
15. Hashemi, R., Danley, J., Tyler, A., Slikker, W., Paule M.: Quality of information granulation: kohonen self-organizing map vs. neighborhood system. *Proceedings of the Fourth International Joint Conference on Information Sciences*, George Georgiou, Janikow, C., Yao, Y. (eds.). Research Triangle Park, NC, pp 294–297 (1998)
16. Hashemi, R., Bahrami, A., Tyler, N., Antonelli, M., Dahlqvist, B.: Property preservation in reduction of data volume for mining: a neighborhood system approach. *The Third International Conference on Data Analytics (DATA ANALYTICS 2014)*, Laux, F., Pardalos, P.M., Crolotte, A. (eds.), Rome, Italy, pp. 105–111 (2014)
17. Hashemi, R., Bahrami, A., Smith, M., Tyler, N., Antonelli, M., Clapp, S.: Discovery of predictive neighborly rules from neighborhood systems. *International Conference on Information and Knowledge Engineering (IKE’13)*, Las Vegas, Nevada, pp. 119–125 (2013)
18. Angrist, J.D., Kugler, A.D.: Rural windfall or a new resource curse? Coca, income, and civil conflict in Colombia. *Rev. Econ. Stat.* **90**(2), 191–215 (2008)
19. White, H.: A Heteroskedasticity-Consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: J. Econ. Soc.* 817–838 (1980)
20. Staiger, D., Stock, J.: To Econometrica. *Econometrica* **65**(3), 557–586 (1997)
21. Wu, D.M.: Alternative tests of independence between stochastic regressors and disturbances. *Econometrica: J. Econ. Soc.* 733–750 (1973)
22. Hausman, J.A.: Specification tests in econometrics. *Econometrica: J. Econ. Soc.* 1251–1271 (1978)