Norma: A Framework for Finding Threshold Associations Between Continuous Variables Using Point-wise Functions

Md Mahin [1] and Christoph F. Eick [2]

[1]University of Houston
[2]Affiliation not available

December 7, 2023

Abstract

This study introduces Norma, a novel association-mining framework tailored for continuous spatial variables analysis. Norma introduces the unique Continuous Variable Threshold (CVT) pattern, aiming to identify a pair of thresholds within the value domain of two continuous variables, revealing strong associations within a specified geographic area. For example, it may unveil a strong association between COVID-19 infection rates above 2% and poverty rates above 15% in New Mexico. Norma associates pointwise functions with each variable-e.g., a function that returns poverty rates for each location in New Mexico. It employs a novel interestingness function, which measures agreement with respect to hotspots where variable pointwise functions exceed associated thresholds. Norma also employs a grid-based spatial hotspot-growing algorithm to discover high-interestingness regions and pairs of thresholds that generate interestingness surpassing a predefined threshold. Furthermore, the framework introduces measures for assessing variable relatedness based on CVT associations. A comparative case study against traditional correlation methods are presented using county-level COVID-19 infection rates and nineteen other socio-economic variables from the continuous United States, and demonstrate how Norma can be used to explore association among subset of values related to spatial continuous variables.

Appendix for paper 'Norma: A Framework for Finding Threshold Associations Between Continuous Variables Using Point-wise Functions'

Md Mahin[1] and Christoph F. Eick[1]

[1]Department of Computer Science, University of Houston, Houston, Texas, USA

1 Proofs of Interestingness Bounds

Proof of E_I: Let $P(v) = \gamma$ and $P(v') = \gamma'$, since hotspots from each variable cover γ and γ' fraction of the observation area, respectively. By the product rule of probability, the probability of the intersection of two independent events is equal to the product of their individual probabilities, i.e., $P(F \cap F') = P(F) \times P(F') = \gamma \times \gamma'$.

Proof of LB_I and UB_I: Let two variables v, v' are extremely negatively correlated. If area coverage from both variables are γ and γ' respectively, where $\gamma + \gamma' > 1$, lower bound of the interestingness function is $1 - ((1-\gamma) + (1-\gamma'))$. However, when $\gamma + \gamma' \leq 1$, lower bound of interestingness will be 0. So the lower bound of I can simply defined as: $LB = max((\gamma + \gamma' - 1), 0)$.

Let the, two variables v, v' are extremely positively correlated. If area coverage from both variables are γ and γ' respectively, upper bound of I will be $UB = \frac{min(\gamma,\gamma')}{max(\gamma,\gamma')}$, where $min(\gamma, \gamma')$ represents the intersection and $max(\gamma, \gamma')$ represents the union.

Similarly, we can define expected interestingness $E_{I'}$, Upper Bound $UB_{I'}$ and Lower Bound Upper Bound $LB_{I'}$ for alternate interestingness I' as follows:

Proof of $E_{I'}$: From multiplication rule of probability, if $P(v \geq t)$ and $P(v' \geq t')$ are c, c', than $P(X \geq t \wedge Y \geq t') = P(X \geq t) \times P(Y \geq t') = c \times c'$.

Proof of $UB_{I'}$ and $LB_{I'}$: If v and v' are extremely positively correlated $P(v \geq t) = c$ and $P(v' \geq t') = c'$, then $UB_{I'} = min(c, c')$.

If v and v' are extremely negatively correlated $P(v \geq t) = c$ and $P(v' \geq t') = c'$, than $LB_{I'} = c + c' - 1$ when $c + c' > 1$ and 0 when $c + c' \leq 1$. So, $LB_{I'} = max((c + c' - 1), 0)$.

Norma: A Framework for Finding Threshold Associations Between Continuous Variables Using Point-wise Functions

Md Mahin and Christoph F. Eick

Abstract—This study introduces Norma, a novel association-mining framework tailored for continuous spatial variables analysis. Norma introduces the unique Continuous Variable Threshold (CVT) pattern, aiming to identify a pair of thresholds within the value domain of two continuous variables, revealing strong associations within a specified geographic area. For example, it may unveil a strong association between COVID-19 infection rates above 2% and poverty rates above 15% in New Mexico. Norma associates pointwise functions with each variable-e.g., a function that returns poverty rates for each location in New Mexico. It employs a novel interestingness function, which measures agreement with respect to hotspots where variable pointwise functions exceed associated thresholds. Norma also employs a grid-based spatial hotspot-growing algorithm to discover high-interestingness regions and pairs of thresholds that generate interestingness surpassing a predefined threshold. Furthermore, the framework introduces measures for assessing variable relatedness based on CVT associations. A comparative case study against traditional correlation methods are presented using county-level COVID-19 infection rates and nineteen other socio-economic variables from the continuous United States, and demonstrate how Norma can be used to explore association among subset of values related to spatial continuous variables.

Index Terms—Spatial Data Analysis, Association Analysis, Interestingness Measure, Grid-Based Search, Pattern Mining

1 Introduction

MANY business enterprises accumulate large quantities of data from their day-to-day operations involving different variables. Association analysis aims to discover interesting relationships among those variables, often also called patterns, hidden in such large data sets. Association analysis frameworks have been developed to mine datasets for different types of relationships, such as association rules, frequent sequences, graphs, collocation, and correlation measure. The interestingnesses of patterns for different, specific types of associations are evaluated using interestingness measures specific to the particular type of association. For example, confidence, support, and lift serve are popular choices to measure the interestingness of a particular association rule [1]. Support is a popular interestingness measure in sequence mining, looking for subsequences that are frequent in the collected data.

The association analysis problem can be formally stated as follows:

Definition 1.1. Given a dataset D, a set of possible patterns P with respect to D, a type of association A, and an interestingness measure I_A with respect to A, find all patterns, p, such that: $p \in P$ and $I_A(p) \geq \theta$
Where θ is a user-defined threshold for the interestingness measure.

As there are usually large numbers of patterns P and as sizes of D are usually large, developing an efficient search procedure for a particular type of association A that finds all interesting patterns is a major research challenge in association analysis. Moreover, there are redundant patterns and many other challenges in association analysis.

In contrast to the prevalent association analysis frameworks that focus on discovering interesting associations between discrete variables [2], this paper concentrates on association analysis for continuous variables. Among some of the existing measures, correlation analysis is one example of finding relationships between various pairs of continuous variables; for example, we might be interested to identify pairs of continuous variables whose absolute correlation values exceed 0.7.

Moreover, the scope of this study is further restricted to the analysis of associations for continuous, spatial variables. This study posits that a pointwise function is associated with each continuous, spatial variable. Pointwise functions calculate the value of a continuous variable at a specific location, providing a convenient way to capture heterogeneity of spatial data by hiding data acquisition details. As an illustration, consider the continuous variables of COVID-19 infection rates and Poverty Rate. We assume the existence of point-wise functions $\psi_{Poverty\ Rate}$ and $\psi_{COVID-19\ Infection\ Rate}$ for the observation area, which can yield the COVID-19 infection rates and Poverty Rate for specific locations defined by their longitude-latitude pairs. Exempt from data generation concerns, an analyst can invoke the function $\psi_{COVID-19\ Infection\ Rate}(35.68, -105.94)$ and it might return 0.003, indicating that the COVID-19 infection rate at the Santa Fe Plaza (located at latitude 35.68 and longitude -105.94) is 0.3%. The main focus of this research is the development of association analysis frameworks for continuous,

• *The authors are with the Department of Computer Science, University of Houston, Houston, TX 77204-2013. E-mail: mmahin, ceick@uh.edu.*

spatial variables that operate on such pointwise functions.

Finding thresholds for continuous variables is an important problem in many applications. For example, when studying the relationship of air pollution and lung diseases it is important to obtain findings such as: *"$PM_{2.5}$ concentrations above $40\mu g/m^3$ are associated with increased occurrence of Bronchial Asthma"*. Finding such thresholds is also critical to come up with governmental regulations and laws to alleviate the health impacts of $PM_{2.5}$. This paper proposes methods to find such thresholds; it focuses on associations between different continuous spatial variables restricting its attention to associations between values of spatial continuous variables above a given threshold, and not all values of the continuous variables. For example, our framework might uncover a strong association between Bronchial Asthma Percentages (BPC) above 0.3% and $PM_{2.5}$ concentrations above $40\mu g/m^3$. In particular, the paper introduces a new form of association $\{(PM_{2.5}\ 40)\ (\text{BMC}\ 0.04)\}$ called CVT (short for "continuous variable threshold") association and introduces a framework called Norma to find such associations. The aim of Norma is to identify a pair of thresholds within the domain of two continuous variables, such that the values above these thresholds exhibit a strong association within a given geographic observation area with respect to a novel interestingness function. The interestingness function measures agreement with respect to hotspots, where the respective variable's pointwise function is above the associated threshold.

The scope of research focusing on exploring associations between continuous, spatial variables has been limited. In the spatial domain, single variable autocorrelation methods such as Moran I [3] are used to find collocation of similar values from one variable. The global methods like Moran I return a single association value between -1 to $+1$, indicating overall association among values from a variable within the observation area. For example, if high values from a variable are located in close proximity to other high values and low values are located in close proximity of other low values in an observation area, we expect a high Moran I score close to +1 and for high low collocation we will get a value close to -1. On the other hand, spatial local autocorrelation measures [4], [5], [6] find these associations on a local scale and locate regions where either high values neighbor other high values or low values neighbor other low values in form high-value cluster or low-value cluster. Nonetheless, none of these methods attempt to discover associations among all subsets of values from the two continuous variables. To the best of our knowledge, only one study, by Eick et al. [7], has explored continuous variable association for subset of values. The framework proposed by Eick et al. categorizes values into high values or low values to find association among them and does not examine all possible associations for different subsets of values. The framework Norma is developed considering such scenarios and explores association among the subset of values from different continuous variables. Overall, Norma tries to answer the following research questions:

- Can we develop a computationally efficient framework to find all valid CVT associations?
- Can the framework extract knowledge that traditional methods like correlation analysis ignore?
- Can the framework find actionable knowledge that is beneficial for the society?

To answer these research questions Norma makes the following contributions:

1) A new framework to mine associations between continuous variables in spatial datasets is introduced, which associates pointwise functions over an observation area with continuous, spatial variables.
2) A new form of spatial association called continuous variable threshold (CVT) association is introduced. A new interestingness measure to mine CVT patterns is introduced, which measures a pattern's interestingness as the agreement with respect to regions of hotspots in the observation area, where the value of the continuous variable is above the respective threshold. A rectangular hotspot discovery algorithm is also proposed to obtain such hotspots.
3) Three measures to asses the relatedness of variables based on the characteristics of observed CVT association are proposed.
4) A case study is presented, which assesses the merit of CVT associations to understand the factors that are associated with high COVID-19 infection rates for a socio-economic dataset from USA.
5) We also briefly discuss how CVT patterns could be generalized to find more general associations between sets of continuous variables and explain how our work differs from other methods while finding association between spatial continuous variables.

The remainder of this paper is structured as follows. In Section 2, we introduce the concepts of CVT patterns and interestingness measures. In Section 3, we detail the computational methods used to solve the CVT association mining problem. Section 4 presents case studies and comparative analyses between CVT association mining and traditional association mining methods. Finally, Section 5 concludes the paper.

2 Problem Statement: Continuous Variable Threshold Associations

2.1 Pointwise Functions

Definition 2.1. **Observation Area**: An observation area is a geographic region where values of the continuous variables are collected. We assume that an observation area *OA* is represented using a bounding polygon.

Definition 2.2. **Pointwise Functions**: A pointwise function ψ_v is a function that maps a geographic location $l \in OA$ to its corresponding value for a given variable v.

A pointwise function $\psi_{COVID-19\ Infection\ Rate}$ related to the variable *COVID-19 Infection Rate* returns the COVID-19 infection rate for any location l in the observation area. Three popular types of pointwise functions are: polygonal function, spatial density functions and spatial interpolation functions.

A **polygonal function** operates on a set of polygon-value pair $\{(f_1, value_1), (f_2, value_2), ..., (f_n, value_n)\}$

Notations

OA	Observation Area or observation polygon	V	A set of continuous variables
l	Any location within OA with format $(latitude, longitude)$	v, v'	Continuous variables
L	A set of l	t, t'	Threshold values related to variable v and v'
F	A set of polygons. In this paper, these polygons are over the geo-spatial observation area.	X, X'	Set of all possible values related to variable v, v', where $X = \{t_1, ..., t_n\} \rightarrow v$ and $X' = \{t'_1, ..., t'_m\} \rightarrow v'$.
f	A polygon. In this paper, these polygons are over the geo-spatial observation area.	$\delta()$	agreement function, measure agreement between two set of hotspot polygons F.
		$\delta'()$	Alternate agreement function.
Y	A set of sets of hotspots. $Y = \{F_1, ..., F_n\}$.	$\psi_v, \psi_{v'}$	Pointwise functions.
R	A set of Rectangualr Hotspot polygons. In this paper, they are polygon set on the Cartesian space formed by the values of v, v'	p	A continuous variable threshold pattern with format $\{(v\ t), (v'\ t')\}$
r	A Hotspot polygon. In this paper, they are polygons on the Cartesian space formed by the values of v, v'	I	Interestingness Measure.
$H()$	Hotspot function. Generate a set of hotspot polygon F for a pointwise function ψ_v based on a threshold t	$\cap_G()$	Global Intersection function. Calculate a scalar intersection polygon area from two set of hotspot polygons F.
$\cap_i()$	Intersection function. Generate an intersection polygon from two hotspot polygon f	$\cup_G()$	Global union function. Calculate a scalar union polygon area from two set of hotspot polygones F.
$area()$	Area function. Calculate the scalar area of a polygon f.	$C()$	Calculate total number of point are above a threshold t when a set of points L applied on a function ψ.
$C_\cap()$	Calculate total number of common points that are above both threshold t and t' when a set of points L applied to function ψ and ψ'.	$range()$	Function to find set of values from two value limit. $range(value_1, value_2)$ indicates all values in between $[value_1, value_2]$.
α	Threshold that restricts the maximum area of the observation OA that can be covered by a hotspot polygon set F.	t_{low}	Lowest threshold from the variable v that generates hotspots with area lower than $\alpha \times area(OA)$.
β	Threshold that restricts the minimum area of the observation area OA that must be covered by a hotspot polygon set F.	t_{high}	Highest threshold from the variable v that generates hotspots with area higher than $\beta \times area(OA)$.
θ	Interestingness threshold	$domain()$	$domain(v)$ indicates set of values a variable v can have.
TS	Two dimension threshold space created using the values of domain(v) and domain(v')	MA	Maximum Interestingness within all patterns from a variable pair.
E_I	Expected interestingness for a variable pair under the assumption of variable independence	$Lift$	Deviation of interestingness from the E_I.
UB	Maximum interestingness possible for a pattern.	LB	Minimum interestingness possible for a pattern.
AUC	Area Under the Interestingness Curve	Max_{Tcorr}	Maximum threshold correlation pattern in a variable pair.

where $\{f_1, f_2, ..., f_n\}$ are polygons inside the OA and $\{value_1, value_2, ..., value_n\}$ are values of a continuous variable v, where each value is associated to a polygon. The polygonal pointwise function $\psi_v(l)$ then return $value_i$ if l is inside polygon f_i.

For example, suppose we have a location *l=(29.76, −95.37)* within the Harris county polygon. The polygonal function $\psi_{COVID-19InfectionRate}(l)$ might return 0.3, which is the COVID-19 infection rate of county.

Another form of pointwise functions are **spatial density functions**. When using this approach a density estimation technique, such as non-parametric density estimation [8], is used to obtain a spatial density function which measures an event's density in an observation area based on the influence of a set of locations where that event occurred. For example, based on locations of burglaries, occurred in an observation area such as Harris County, Texas, a kernel density function $\psi_{Burglary}$ can be used as a pointwise function that will return the density of burglary events on a location based on the influence of existing burglary events from the nearby locations.

Moreover, **spatial interpolation functions** [9] can serve as pointwise functions; spatial interpolation functions compute the value for a query location l based on a set of location-value pairs that were observed inside the observation area; for example, we might calculate the amount of rainfall in a location l by interpolating the observed rainfall quantities from five weather stations in the observation area.

2.2 Continuous Variable Threshold Pattern

Let V a set of m distinct continuous spatial variables with associated pointwise functions $\{\psi_{v_1}, ..., \psi_{v_m}\}$. A continuous variable threshold pattern p for a dataset D with variables V is defined as a set S of variable-threshold pairs of the form $(v\ t)$, where $v \in V$ and $t \in \Re$, subject to the following constraints:

1) the cardinality of S is at least 2
2) each variable $v \in V$ occurs at most once in S

Initially, we concentrate attention on binary CVT patterns. For two continuous variables $v, v' \in V$ and associated thresholds $t \in domain(v), t' \in domain(v')$, we define pattern p as follows:

$$p = \{(v\ t), (v'\ t')\} \tag{1}$$

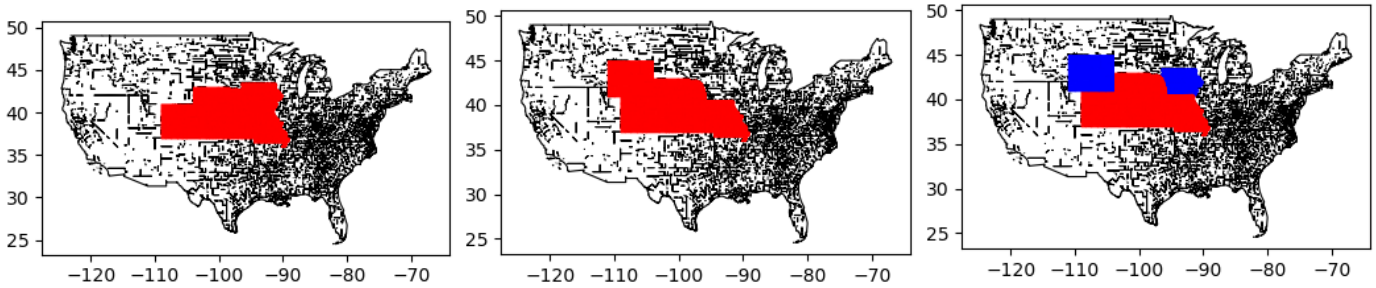
A pattern such as *p={(COVID-19 Infection Rate 0.25),(Median Income 50000)}* is interesting if the interestingness is above a user defined threshold e.g. 0.70.

2.3 Interestingness Measure

We will introduce an interestingness measure in this section. Let us consider a set of variables $\{v, v'\}$ with a set of associated thresholds $\{t, t'\}$, and associated pointwise functions $\{\psi_v, \psi_{v'}\}$; we define interestingness measure I for pattern $p = \{(v\ t), (v'\ t')\}$ as follows:

$$I(\{(v\ t), (v'\ t')\}) = \delta(H(\psi_v, t), H(\psi_{v'}, t')) \tag{2}$$

The interestingness measure in equation 2, assesses the regions where the respective pointwise functions surpass the associated threshold, and calculates the degree of agreement between them. The function δ determines the agreement of the regions where pointwise functions are above



(a) $F = H(\psi_v, t)$, Area Cover of OA = 13%

(b) $F' = H(\psi_{v'}, t')$, Area Cover of OA = 14%

(c) Intersection of F,F′. Area Cover of OA = 11% and interestingness = 0.69

Fig. 1: Behaviour of Interestingness measure based on Hotspot coverage for v and v' with percent of the area covered by each hotspot, their intersection area with area coverage, and measured interestingness value for the pattern. (c) shows intersection of F, F' in red color (69% of union of F, F') and blue color represents where F, F' does not agree (31% of union of F, F'), resulting total agreement of 0.69.

the associated thresholds. If I value is high, it indicates a strong agreement, otherwise, a weak agreement.

The hotspot function $H(\psi, t)$ in equation 3, returns the set of k polygons, denoted as $F = \{f_1, f_2 \ldots, f_k\}$ from OA. Each polygon in the set, $f \in F$, satisfies the condition that the pointwise function ψ is greater than or equal to the threshold t at all points $l \in f$:

$$\forall f \in F(\forall l \in f(\psi(l) \geq t)) \tag{3}$$

Given a set of k hotspot polygons generated for the variable v, represented as $F = \{f_1, f_2 \ldots, f_k\}$, and a set of j hotspot polygons generated for the variable v', represented as $F' = \{f'_1, f'_2 \ldots, f'_j\}$, the agreement between the two sets of polygonal hotspots is computed using Equation 4. This equation finds the ratio between the area of the overlapping space (calculated by $\cap_G$ function) and area of the union space occupied by the two sets of hotspots (calculated by $\cup_G$ function).

$$\delta(F, F') = \frac{\cap_G(F, F')}{\cup_G(F, F')} \tag{4}$$

The $\cap_G$ function from Equation 5 calculates the total sum of intersection polygon areas for two sets of polygons. The function performs an iteration over all pairs of hotspot polygons. It selects a pair of polygons (f, f') from the Cartesian product of the two sets of hotspots $F \times F'$. Next for each (f, f'), the intersection polygon is found using the $\cap_i$ function and the areas of all intersection polygons are summed to produce a scalar value.

$\cap_G$ is defined by the following formula:

$$\cap_G(F, F') = \sum_{x=1}^{k} \sum_{y=1}^{j} area(\cap_i(f_x, f'_y)) \tag{5}$$

The $\cup_G$ function in Equation 6, returns a scalar value, which is the union area covered by the hotspots from the two sets F, F'. It is found by taking the sum of the area of all polygons in both sets, and subtracting the intersection area as calculated by $\cap_G$.

$$\cup_G(F, F') = (\sum_{x=1}^{k} area(f_x) + \sum_{y=1}^{j} area(f_y)) - \cap_G(F, F') \tag{6}$$

Figure 1 depicts hotspots for an imaginary pattern $p = (\{(v\ t), (v\ t')\})$, where Figure 1.a and Figure 1.b display the hotspots of variable v, $F = H(\psi_v, t)$ and variable v′, $F' = H(\psi_{v'}, t')$. F and F' cover 13% and 14% of the observation area, respectively. The intersection hotspots of F and F', obtained by applying $\cap_i$ over all (f, f') from $\{F \times F'\}$, are presented in Figure 1.c in red color. Blue color in the figure represents the regions from the two pointwise functions where F and F' do not intersect. The intersection hotspots cover 11% of the given observation area. The red portion from Figure 1.c covers 69% of the union area covered by all hotspots F, F' from column 1 and 2 (red and blue portion combined), resulting in the interestingness value 0.69. In Figure 1.c, blue color represents the 31% of the area from the union of F, F' where they do not agree.

2.4 Approximation of the Interestingness Measure

This section presents a computationally efficient method for measuring agreement between variable threshold pairs and define an approximate interestingness measure which is quite economical to compute. Instead of relying on computing hotspots overlap, we utilize a sampling technique to measure interestingness. Specifically, we apply pointwise functions ψ_v and $\psi_{v'}$ to a set of n sample locations $L = \{l_1, ..., l_n\}$ within the observation area. Interestingness measure I', for a pattern $p = \{(v\ t), (v'\ t')\}$ in the following equation:

$$I'(\{(v\ t), (v'\ t')\}) = \delta'(C(\psi_v, t, L), C(\psi_{v'}, t', L), C_\cap(\psi_v, t, \psi_{v'}, t', L)) \tag{7}$$

In Equation 7, δ' is the approximate agreement function. The C function defined in Equation 8, applies either the ψ_v or $\psi_{v'}$ function to a set of locations L and counts the total number of locations c or c' that are above the respective threshold t or t'.

$$C(\psi_v, t, L) = count(\{l \in L | \psi_v(l) \geq t\}) \tag{8}$$

The $C_\cap$ function defined in Equation 9, also operates on the same set of locations L and counts the total number of locations $c_\cap$, where in each location, both ψ_v and $\psi_{v'}$ are above their respective thresholds t and t'.

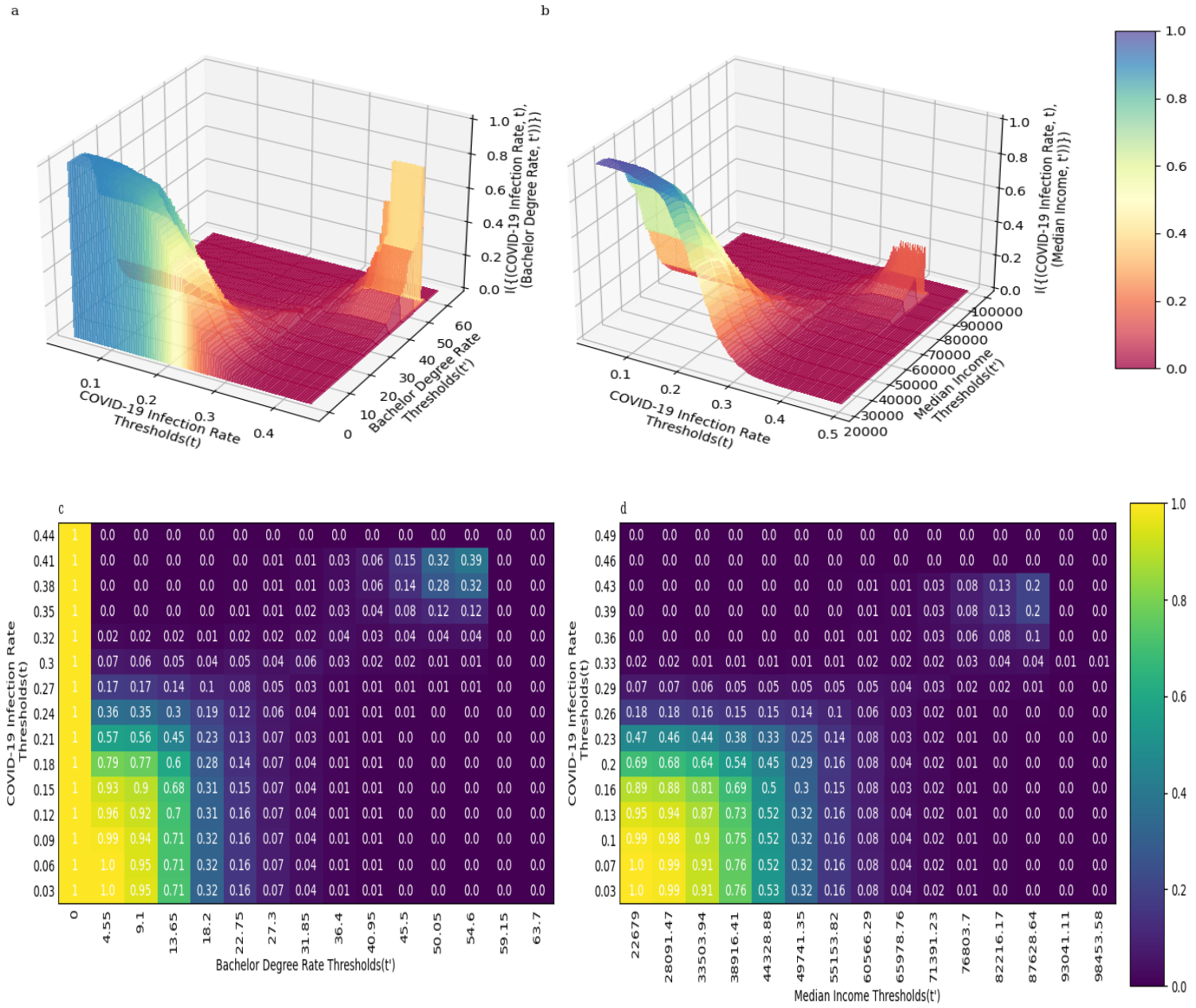


Fig. 2: A depiction of interestingness function surface for patterns $\{(Covid-19\ Infection\ Rate, t)(Bachelor\ Degree\ Rate, t')\}$ (a) and $\{(Covid-19\ Infection\ Rate, t)(Median\ Income, t')\}$ (b), with 100×100 sample threshold pairs from two variables. (c) and (d) depicts interestingness value for 15×15 pair of thresholds for $\{(Covid-19\ Infection\ Rate, t)(Bachelor\ Degree\ Rate, t')\}$ and $\{(Covid-19\ Infection\ Rate, t)(Median\ Income, t')\}$.

$$C_{\cap}(\psi_v, t, \psi_{v'}, t', L) = count(\{l \in L | \psi_v(l) \geq t \wedge \psi'_v(l) \geq t'\}) \quad (9)$$

The approximate agreement function δ' defined in Equation 10, determines the level of agreement between two patterns by computing the ratio of the number of locations where both ψ_v and $\psi_{v'}$ exceed their respective thresholds ($c_{\cap}$) to the number of locations above the threshold for either ψ_v or $\psi_{v'}$ (i.e., the union of points), which is measured as $(c + c' - c_{\cap})$.

$$\delta'(c, c', c_{\cap}) = \frac{c_{\cap}}{c + c' - c_{\cap}} \quad (10)$$

2.5 Constrained CVT patterns and Expected Interestingness

2.5.1 Constrained CVT patterns

The Interestingness measure defined in Equation 2 has the property that it returns high interestingness values if very low thresholds (t, t') are chosen for the variable pair (v, v'), as in this case the respective hotspots cover large percentages (γ, γ') of the observation area; that is, γ and γ' are close to 1. For example if $\gamma = \gamma' = 0.99$, the interestingness for the associated CVT pattern is in the interval of (0.98, 1.0) and can be viewed a trivial patterns. On the other side of the spectrum, if γ and γ are close to 0; e.g. 0.001; even if a patterns has a high agreement, we do not like to report it as it lacks support. One example can be seen in Figure 2.a

(upper right corner), where the agreement value of 0.8 is generated by two hotspots sets that cover only 0.07% of the observation area.

To address the issue of trivial and low support CVT patterns, we introduce two user-defined parameters α and β to narrow down the range of thresholds used in CVT pattern mining, where α restricts lower thresholds and β restricts upper thresholds. These parameters are within the range $[0, 1]$. For instance, $\alpha = 0.5$ means that only the thresholds which cover hotspot areas lower than 50% of the observation area are considered interesting. On the other hand, if $\beta = 0.01$ is chosen, hotspot areas associated with the threshold must cover at least 1% of the observation area to be considered interesting. Thresholds space restricted by α and β can be represented by $\{[t_{low}, t_{high}] \times [t'_{low}, t'_{high}]\}$, where $\{t_{low}, t_{high}\} \in domain(v)$ and $\{t'_{low}, t'_{high}\} \in domain(v')$. $\{t_{low}, t'_{low}\}$ are the smallest set of thresholds that satisfies α and $\{t_{high}, t'_{high}\}$ are the largest set of thresholds that satisfies β from the two variables.

2.5.2 Expected Interestingness and Upper and Lower Interestingness Bounds

Lemma 2.1 define expected interestingness value for two sets of hotspots for Interestingness defined in Equation 2 based on their area cover (γ, γ'). According to Lemma 2.1, if hotspots from v, v' covers 0.75 or 75% of the observation area, the expected interestingness value is 0.563.

Lemma 2.1. **Overlap Probability of Hotspot Sets from Independent Variables.** Let F and F' be two sets of hotspots from two independent variables v and v', respectively, which cover γ and γ' fraction of an observation area, where $\gamma, \gamma' \in [0, 1]$. Then assuming independence, the probability of hotspots being overlap is $E_I = \gamma \times \gamma'$.

On the other hand, lower and upper bound can be determined by considering two variables as extremely negatively correlated and extremely positively correlated. Lemma 2.2 define lower and upper bound of interestingness for two sets of hotspots.

Lemma 2.2. **For two sets of hotspots F and F', with area coverage γ and γ' fraction of an observation area the Lower Bound LB of their interestingness will be $max((\gamma + \gamma' - 1), 0)$ and Upper Bound UB of the interestingness value will be $\frac{min(\gamma,\gamma')}{max(\gamma,\gamma')}$.**

According to Lemma 2.2,in case of lower bound, When both variable are in extreme negative correlation, interestingness is 1 when both hotspot sets cover 100% of the observation area, but for every 1% less area coverage from any set of hotspots, agreement will be reduced by 0.1, until its reaches zero. On the other hand, in case of upper bound, When both variable are in extreme positive correlation, if one hotspot set covers 0.25 fraction of the observation area and another covers 0.5 fraction of the observation area, agreement value is expected to be 0.25, ensuring maximum overlap with smaller set of hotspots.

For a pattern $p = \{(v\ t), (v'\ t')\}$ with interstingness I, if $LB < I < E_I$, we can say the pattern has no association and if $E_I < I < UB$, we can say the pattern has some positive association.

We further define a Lift measure to reflect the strength of relationship between two sets of hotspots in the Definition 2.3.

Definition 2.3. **Lift** of a CVT pattern $p = \{(v\ t), (v'\ t')\}$ is $\frac{I}{E_I}$, where I is obtained interestingness and E_I is expected interestingnness under the assumption that both variables are independent.

According to Definition 2.3, Lift is the fraction of interestingness we get from a pattern p, divided by the expected interestingness E_I if both variables were completely independent. Lift value greater than 1 indicates positive association and Lift value between 0 to 1 indicates lack of association.

2.6 Demo of CVT Associations and the Interestingness Function I

As CVT associations are new, we try to discuss some example CVT associations in this subsection, to provide the reader with:

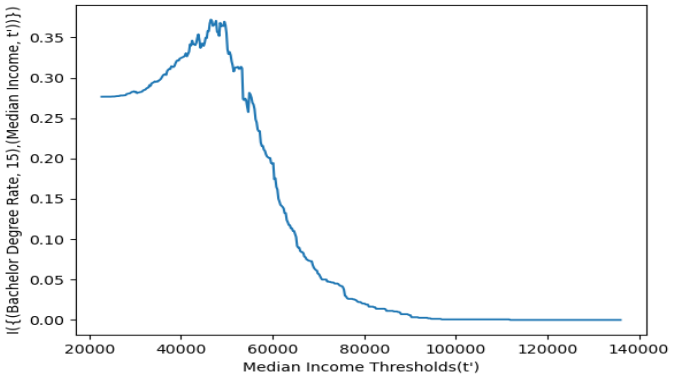
1) A better understanding of the semantics of CVT associations.
2) To illustrate how the introduced interestingness measure I works.
3) To discuss some examples to illustrate the computational challenges of mining CVT associations.

In our analysis, we have observed the introduced bivariate interestingness measure I usually has multiple peaks. Figure 2 depicts the nature of interestingness measure I for the patterns {*(Covid-19 Infection Rate t),(Bachelor Degree Rate t')*} and {*(Covid-19 Infection Rate t),(Median Income t')*}. Figure 2.a and Figure 2.b depicts the interestingness functions which were constructed based on (100×100) sample threshold pairs where trivial patterns of interestingness almost 1 can be seen for low thresholds (lower left corner). Figure 2.c and Figure 2.d shows the results in tabular form for (15×15) thresholds pairs from the same patterns, where Figure 2.c interestingness 1 observable for Bachelor Degree threshold 0, which cover whole observations area. Similarly, Bachelor Degree Rate threshold 54.6 and COVID-19 Infection Rate threshold 0.41 has interestingness 0.39.

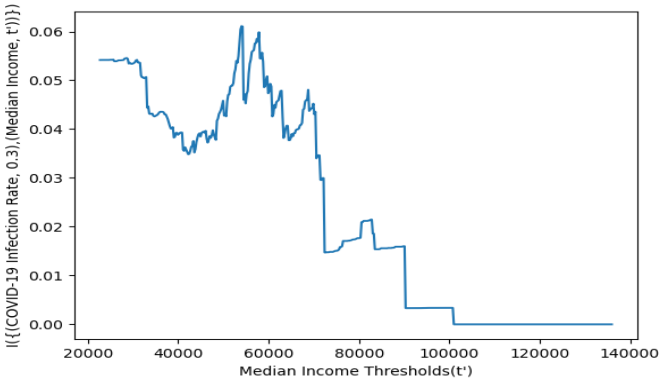
From Figure 2 we can deduce the interestingness measure I can have multiple local maxima.For example, apart from the maxima on the lower thresholds from both patterns, we can see another local maxima 0.39 on Figure 2.c for the pattern {*(Covid-19 Infection Rate t),(Bachelor Degree Rate t')*} on thresholds (0.41,59.15) and 0.2 for the pattern {*(Covid-19 Infection Rate t),(Median Income t')*} on thresholds (0.43,93041.11). Figure 2.a suggest existence of another higher maxima around 0.8 for the pattern {*(Covid-19 Infection Rate t),(Bachelor Degree Rate t')*}.

To visualize nature of CVT Associations and the Interestingness function more closely, we analyzed patterns with a constant threshold value c for one variable. Figure 3 visualizes two such interestingness measure. In Figure 3.a the maximum interestingness can be observed to be around the threshold $t' = 50000$ for the pattern $p = (Bachelor\ Degree\ Rate\ 20), (Median\ Income\ t')$.

In some cases, the function can exhibit significant complexity with several peaks, as



(a) $I(\{(Bachelor\ Degree\ Rate\ 15)(Median\ Income\ t')\})$



(b) $I(\{(COVID-19\ Infection\ Rate\ 0.3)(Median\ Income\ t')\})$

Fig. 3: A depiction of interestingness function plot for the pattern $\{(Bachelor\ Degree\ Rate\ 20)(Median\ Income\ t')\}$ and $\{(COVID-19\ Infection\ Rate\ 0.3)(Median\ Income\ t')\}$ based on 500 sample thresholds from the variable $Median\ Income$

demonstrated in Figure 3.b for the pattern $\{(COVID-19\ Infection\ Rate\ 0.3),(Median\ Income\ t')\}$. In this case, function I has multiple modes with numerous local maxima. Due to the function's continuous nature, pinpointing the exact location of the global maxima poses a significant challenge, and search algorithms may get trapped in local maxima.

2.7 Three Measures for the Relatedness of Variables Based on Mined CVT Associations

After we analyzed the CVT associations between two variables v and v' for threshold pairs in the $[t_{low}, t_{high}]$ $[t'_{low}, t'_{high}]$ rectangle, what does this tell us about the relatedness of the two variables v and v'? To address this question, we propose three different evaluation measures in this section, which measure the relatedness of v and v', based on observed CVT-association as follows:

1) The maximum observed interestingness in $[t_{low}, t_{high}] \times [t'_{low}, t'_{high}]$, called *MA*
2) The area under the interesting curve over $[t_{low}, t_{high}] \times [t'_{low}, t'_{high}]$, called *AUC*
3) The percentage of the area in $[t_{low}, t_{high}] \times [t'_{low}, t'_{high}]$ where the interestingness is above a user-defined interestingness threshold ; e.g. $\theta = 0.55$; this evaluation measure is called PIT.

For example, if the maximum observed interestingness of variables v1 and v2 is 0.42 and the maximum observed interestingness of variables v1 and v3 is 0.79, we would conclude that v3 is more related to v1 than v2. The three measures will be explained in more detail next!

2.7.1 Maximum CVT Association (MA)

The objective of *MA* is to identify maxima for the function $g(t, t')$ defined in Equation 11, which is a two-dimensional and continuous function as illustrated in Figure 2.a,b. Here, the maxima will identify the threshold pattern with maximum interestingness.

$$g(t, t') = I(\{(v\ t), (v'\ t')\}) \quad where, t_{low} \leq t \leq t_{high}\ and\ t'_{low} \leq t' \leq t'_{high} \tag{11}$$

Locating the maximum value within the rectangular space of $g(t, t')$ can pose a considerable challenge due to the extensive search space arising from a high number of potential thresholds across two variables. Additionally, the proposed measure of interestingness may exhibit multiple local maxima as depicted in Figure 2.a,b, further complicating the task. These issues further necessitates for sophisticated search procedures to find maxima.

2.7.2 Area Under the Curve (AUC)

The strength of CVT association for two variable v, v' can be quantified as the Area Under the Interestingness Curve(AUC). The higher the AUC, more strongly variable v and v' are associated threshold its threshold space. For $I(\{(v\ t), (v'\ t')\})$ over the rectangle formed by $(t_{low}, t_{high}) \times (t'_{low}, t'_{high})$, where $(t_{low}, t_{high}) \in domain(v)$ and $(t'_{low}, t'_{high}) \in domain(v')$, AUC is formally defined as follows:

$$AUC(v, t, v', t') = \int_{t_{low}}^{t_{high}} \int_{t'_{low}}^{t'_{high}} I(\{(v\ t), (v'\ t')\})\ dt\ dt' \tag{12}$$

2.7.3 Percentage of Interestingness above Threshold θ (PIT)

We define another measure Percentage of Interestingness above Threshold θ (PIT) because of the similar purpose discussed in Section 2.7.2. PIT measures total percentage of thresholds pairs whose Interestingness above an user-defined threshold θ.

Considering the whole two dimensional threshold space TS created from $domain(v)$ and $domain(v')$ as a continuous space, we can define PIT as follows:

$$PIT(\theta) = \frac{area(TS \geq \theta)}{area(TS)} \tag{13}$$

In Equation 13 defines PIT as function of the parameter θ that measures the fraction of threshold space TS above θ.

Computational procedures to implement *MA*, AUC and PIT are discussed in Section 3.3.

2.8 CVT Patterns of Arity Three or More

The interestingness measure can further be extented for a set of n continuous variables $\{v_1, ..., v_n\}$. For a set of n threshold $\{t_1, ..., t_n\}$ associated with the n variables, where $t_i \in domain(v_i)$, we define a CVT pattern with arity n, p_n as follows:

$$p_n = \{(v_1\ t_1), ..., (v_n\ t_n)\} \tag{14}$$

The interestingness measure for arity n, I_n can be defined as follows:

$$I_n(\{(v_1\ t_1), ..., (v_n\ t_n)\}) = \delta_n(H(\psi_{v_1}, t_1), ..., H(\psi_{v_n}, t_n)) \tag{15}$$

Each $H()$ function from Equation 15 generates a set of hotspot. For n set of hotspots $\{F_1, F_2, ..., F_n\}$ from n number of $H()$ function, δ_n can be defined as follows,

$$\delta_n(F_1, ..., F_n) = \frac{\cap_G(F_1, ..., F_n)}{\cup_G(F_1, ..., F_n)} \tag{16}$$

The $\cap_G$ in Equation 16 returns the scalar intersection area within all n set of hotspots and $\cup_G$ returns scalar total union area covered by all hotspot sets. The δ_n function returns a scalar fraction values within $(0, 1)$ indicating the fraction covered by the overlap compared to the total area coverage by the n set of hotspots as explained in Section 2.3

2.9 CVT Patterns in Non-spatial Environments

Analyzing threshold associations can also benefit non-spatial associations, such as correlation. In general setting, classical correlation gives an overall association value among all values from two variables. However, often relative strong correlation can be observed for high values of two variables, instead of the fact that the correlation between the two variables over the full domain of values is quite weak. For example, we created a dataset $D = TH_{Houston}$ using hourly measurements of relative humidity and temperature over a year for a location in Houston (latitude= 29.75225, longitude = -95.3689) and found that there is almost no correlation between humidity and temperature (the correlation is -0.11), but there is a strong negative correlation of -0.80 between humidity above 53% and temperatures above 80.98; similarly, there is a strong correlation of -0.77 between humidity above 48% and temperatures above 85.01—as it is reported in figure 4. Again it is important to determine thresholds above which such associations become strong. In this section we propose that instead of looking for correlations between two variables, look for correlation between subset of values above two thresholds within the two variables.

For a dataset D with a set of continuous attributes $\{v, v'\}$, and a set of associated thresholds $\{t, t'\}$ for each attribute, we define threshold correlation as TCorr as follows:

$$TCorr(D, v, v', t, t') = Correlation(D', v, v') \tag{17}$$

In Equation 17, the TCorr function measure correlation of subset of values D' from dataset D, extracted based on the threshold values $t \in v$ and $t' \in v'$. D' is formally defined in Equation 18.

$$D' = \{(a, a') | (a, a') \in D \wedge a \geq t \wedge a' \geq t'\} \tag{18}$$

Equation 18, defines D' as the set of all value pairs (a, a'), where $a \geq t$ and $a' \geq t'$. Here, $a \in domain(v)$ and $a' \in domain(v')$.

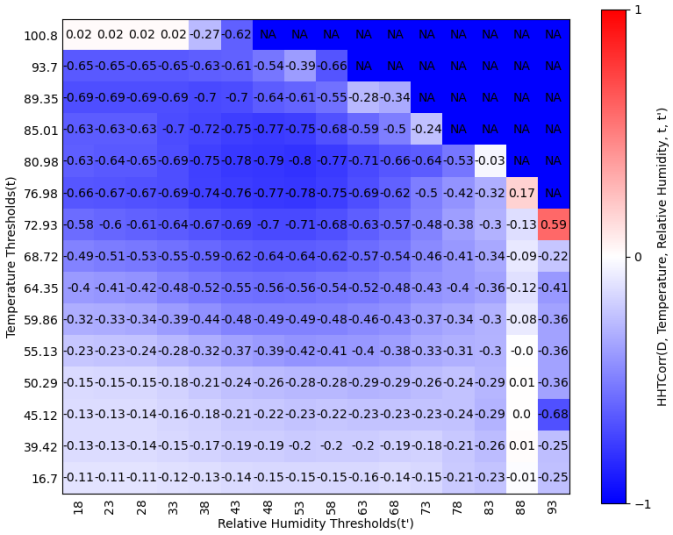


Fig. 4: Threshold Correlations for 15×15 samples for $TCorr(D, Temperature, Relative\ Humidity, t, t')$ and Support.

Figure 4 shows the behaviour of the Threshold Correlation for the dataset $D = TH_{Houston}$. Temperature variable in $TH_{Houston}$ was measured in Fahrenheit and relative humidity percentages are reported as an integer value between 0% to 100%. Open Weather Map data were used to create the dataset. $TH_{Houston}$ contains 8760 temperature-humidity pairs. The number of unique temperature values is 3104 and the number of unique humidity values is 80. The correlation between the two variables in D is -0.11. Figure 4 shows the behaviour of TCorr function for 15 sample thresholds from each attribute.

Similar to interesting CVT patterns, interesting threshold correlation indicates variable threshold pairs where the interestingness measure is above a user defined threshold; e.g. the absolute value of the threshold correlation is above 0.7. We claim that a lot of the algorithms we introduce in Section 3 can be used to mine interesting threshold correlation. More on threshold correlation analysis has been reported in a paper published in GeoAI'23 [10].

3 Computational Methods to Find Strong CVT Associations

This section discusses different key algorithms to generate CVT associations.

3.1 Grid Based Hotspot Growing Algorithm

In this paper, we propose a grid-based hotspot growing algorithm for identifying hotspots above a continuous variable threshold t. Our algorithm bears resemblance to the DBSCAN clustering algorithm [11] and utilizes a common graph traversal strategy to discover connected sets of locations above the threshold t for hotspot conversion.

In our approach, we relies on sampling and create a $(n \times n)$ equidistant sample location matrix G within OA, where

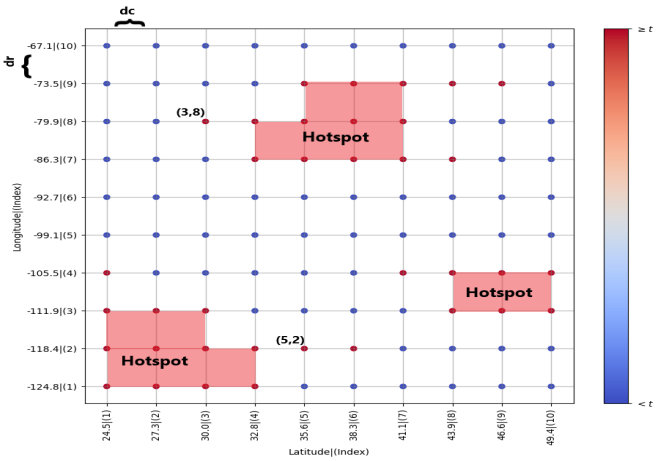


Fig. 5: An illustration of sample grid points over an OA. Red points indicates $value \geq t$, blue points indicates $value < t$. Red cells indicates hotspots. dr is the distance between two row and dc is distance between two column. The red cells indicates hotspots and red points not in red cell are points above t that are not part of any hotspot.

Algorithm 1: Grid Based Hotspot Discovery

Input : grid coordinate matrix G,threshold t, point-wise function ψ_v
Output: A set of hotspot polygons F

1 Create empty Queue Q and empty hotspot polygon list F;
2 **for** *location* $l \in G$ **do**
3 Initialize Connected Location list CP;
4 **if** $\psi(l) \geq t$ *and* l *is not visited* **then**
5 Add l to Q;
6 **end**
7 **while** Q *is not empty* **do**
8 Target Location Tp = pop(Q);
9 Find four Neighbour Locations Np of Tp;
10 **for** *each neighbour* $N \in N_p$ **do**
11 **if** $N \in G$ *and* $\psi(l) \geq t$ *and* N *not visited* **then**
12 Add N to Q;
13 **end**
14 **end**
15 Mark Tp visited;
16 Add Tp to CP
17 **end**
18 Mark p visited;
19 Find list of cell Lc of G from CP;
20 Calculate union polygon f from Lc;
21 Add f to F;
22 **end**
23 **return** F

each location maintains a distance of dr in the row direction and dc in the column direction with neighboring locations (defined in the Definition 3.1). Next, for a variable v with respective pointwise function ψ_v, we find all disjoint sets of connected locations (defined in the Definition 3.2) within G, where each location have pointwise function value $\psi_v \geq t$. Finally, from each connected set of locations (defined in the Definition 3.3), we find all grid cells and return their union polygon as hotspot polygon.

Definition 3.1. **Neighbours**: Two locations l_1 and l_2 in a matrix of sample locations G are neighbours if, for a distance threshold τ: $distance(l_1, l_2) \leq \tau$.

Definition 3.2. **Connected Locations**: For a variable v with respective pointwise function ψ_v and a continuous variable threshold t, two locations l_1 and l_2 in a matrix of sample locations G are connected if:

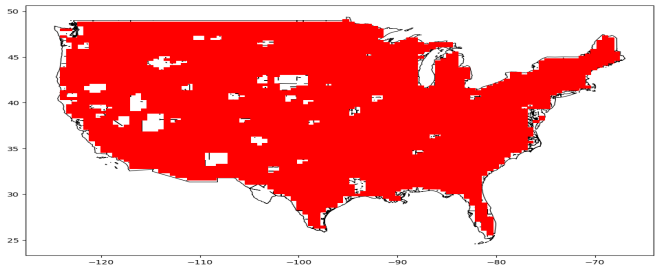
- l_1 and l_2 are neighbours.
- $\psi_v(l_1) \geq t$ and $\psi_v(l_2) \geq t$.

Definition 3.3. **Connected Set of Locations** are the transitive closure of connected locations where for any two locations l_1 and l_n within the set there exist a chain of locations $\{l_1, l_2, ..., l_n\}$, where l_i and l_{i+1} are Connected locations.

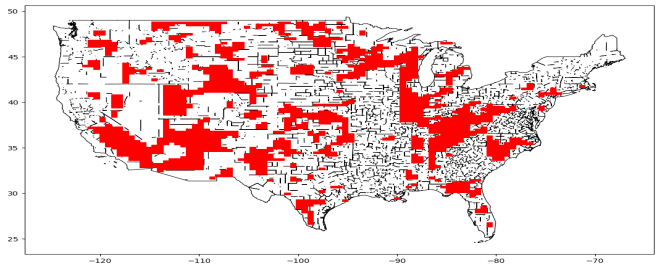
In this study, within the matrix G, for an index $[i, j]$, the connected locations (defined in the Definition 3.2) are located at indexes $[i+1, j]$, $[i-1, j]$, $[i, j+1]$, and $[i, j-1]$.

Figure 5 illustrates our complete approach. In this figure we have created a (10×10) sample location matrix from the continuous part of USA, where two neighbouring locations maintains dr distance for longitude and dc distance for latitude. The red points indicates value of the point-wise function ψ_v on that location is $\geq t$ and blue point indicates $\psi_v < t$. Each red rectangles within the figure represents a hotspot. In our approach, if all locations of a grid cell is present in a connected set of locations it is included in a hotspot. As a result some locations, such as location on index (3,8) in Figure 5 is not part of any hotspot. This procedure maintains the shape of the hotspot to some extent. Moreover, retaining solely the cells simplifies the generation of hotspot polygons by obviating the need for employing additional, intricate polygon generation mechanisms, such as the convex hull. The detailed procedure of extracting such hotspots are explained in the algorithm 1.

Algorithm 1 outlines the whole grid based hotspot discovery procedure. From step 1 to 17, the algorithm finds a set of connected locations above t using a queue Q. For each connected locations list CP, the algorithm iterates through all locations from the matrix G, adds a location to the Q if the location is not visited and corresponding pointwise function returns a value above t. Next, from line 7 to 14, it iterates through all the neighbors of the location appended in Q. In this continuous process, the algorithm extracts a location from the Q, finds its neighbors, and appends each neighbor to the Q if they are not previously visited and pointwise function value is above t. As a result, after each iteration, the Q grows if more neighbor locations above t are available. When the Q is empty, a set of connected locations is already added to CP. The neighbor locations indicate the locations dr or dc distance away from the location. In line 18 it marks the current traversed location visited to avoid repeated checking, in case it is below t. Then, the algorithm converts each set of connected location CP into a polygon from lines 19 to 20 using a cell-to-polygon conversion technique. The algorithm first finds all cells list



(a) $H(\psi_{Covid-19\ Infection\ Rate}, 0.14)$. Area coverage is 90%.



(b) $H(\psi_{Covid-19\ Infection\ Rate}, 0.25)$.Area coverage is 25%.

Fig. 6: Hotspots using algorithm 1 for the variable Covid-19 Infection Rate with two different thresholds.

Lc, based on the condition that all four of the cell corners need to be within the CP. A polygon f is then generated using the cells within Lc at line 20. Finally, at line 21, the generated polygon f is added to the hotspot polygon set F, and the set is returned as the output of the algorithm at line 23.

The figure 6 represents the hotspots for contiguous USA for the COVID-19 Infection Rate variable. In Figure 6.a, the threshold $t = 0.14$ is set too low, which results in the hotspot covering almost the entire observation area. However, when the threshold is increased to 0.25 as shown in Figure 6.b, the area coverage decreases significantly (90% to 25%).

Numerous grid-based hotspot growing algorithms have been proposed in the literature, including methods by Akdag et al. [12], Deng et al. [13], Wang et al. [14], and Darong et al. [15]. These methods typically operate on density-based principles, expanding hotspots based on cells with densities above a threshold. These cell density are typically depends on counting the number of event locations such as accident within each cell. It's important to note that these methods calculates cell density based on points that lack associated values. In contrast, our approach centers around pointwise functions that provide an associated value for each location, which can itself be a representative of density for a location. To better align with this characteristics of our pointwise functions, we adopt a representative-based approach. In this approach, we use only the four corner points of a cell to determine whether it is dense, and then we grow the hotspot accordingly. This approach is particularly well-suited for our current study, as the point-wise functions operate over polygons, and points within the same polygon yield the same associated value.

3.1.1 Computational Complexity of Hotspot Discovery Procedure

During each iteration, the algorithm evaluates four neighboring locations from N sample locations, where neighbours are defined within a given matrix. Consequently, it exhibits a worst-case time complexity of $\mathcal{O}(4 \times N)$, accompanied by an additional space complexity of $\mathcal{O}(N)$.

The polygon generation process in Norma undertakes a linear search to identify all hotspot cells, for N sample locations within a connected set of locations. The total number of possible disconnected cells are $(\frac{\lfloor\sqrt{N}\rfloor}{2})^2$, which, in the worst-case scenario, simplifies to $\frac{N}{5}$. This can be understood using the Figure 5, if we consider a hotspot cell, surrounded by not hotspot cells. This results in the complexity of $\mathcal{O}(N)$.

3.2 Hotspot Discovery in the Threshold Space

The hotspot discovery algorithm, delineated in Section 3.1, can be further utilized to identify hotspots exceeding an interestingness threshold of θ. In this revised search procedure, the geographic location matrix G will be substituted with a sample threshold pair matrix, the location l will be replaced by a threshold pair (t, t') from that matrix, and the pointwise function will be replaced by the interestingness measure I. The threshold pair (t, t') indicates a location within the threshold space formed from domain(v), domain(v'). The definition of new neighbourhood is defined based on the distance threshold pairs $(\Delta t, \Delta t')$, where any point falls within the radius drawn by $(\Delta t, \Delta t')$ from the point (t, t') is the neighbour of (t, t'). The search procedure has been formally defined in Equation 19.

$$Z = \{(t, t') | g(t, t') \geq \theta\} \tag{19}$$

We can discover interesting hotspots above a certain interestingness threshold-value θ from the interestingness space abstracted by the function $g(t, t')$ in form of smaller, disjoint polygons. Specifically, our objective is to identify a set of n non-overlapping hotspot polygons $R = \{r_1, r_2, ..., r_n\}$ from the threshold space, where all threshold pairs (t, t') within a given hotspot polygon $r \in R$ are expected to generate interestingness values above a predefined threshold θ, as formally defined by Equation 20.

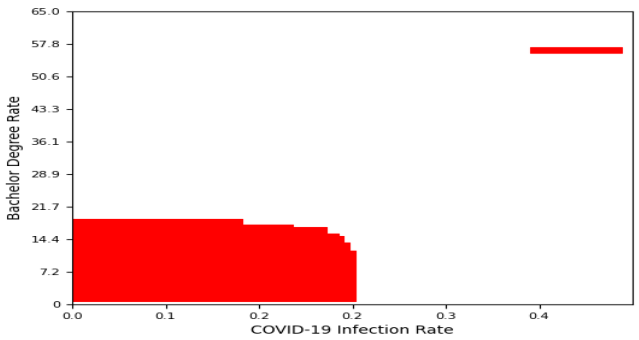
$$\forall t \forall t' ((inside((t, t'), r) \rightarrow I((v\ t), (v'\ t')) \geq \theta)) \tag{20}$$

In Equation 20, the $inside$ function checks if a threshold pair (t, t') is inside polygon r.

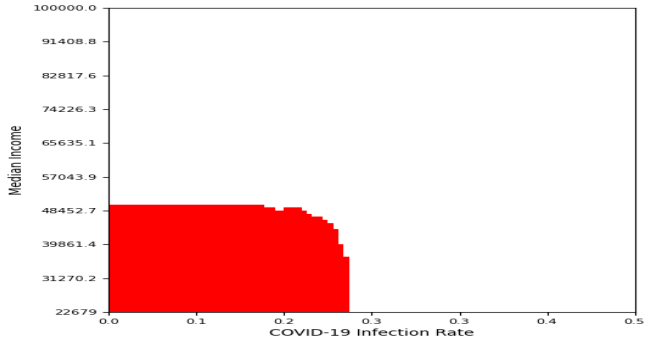
Figure 7 depicts two examples of hotspot polygons observed on a grid intersection points for CVT patterns {*(COVID-19 Infection Rate t),(Bachelor Degree Rate t')*} shown in Figure 2.a.1 and {*(COVID-19 Infection Rate t),(Median Income t')*} shown in Figure 2.b.1, with $\theta \geq 0.3$. Figure 7.a shows two separate polygons, while Figure 7.b shows only one polygon above the interestingness threshold of 0.3, which is consistent with the patterns shown in Figure 2.a.1 and Figure 2.b.1.

3.3 Implementation of Computational Methods to Measure CVT Association Relatedness

In this study, the computational approach for related metrics has been deliberately kept straightforward, solely employ-



(a) $I(\{(COVID-19\ Infection\ Rate, t)(Bachelor\ Degree\ Rate, t')\})$.



(b) $I(\{(COVID-19\ Infection\ Rate, t)(Median\ Income, t')\})$.

Fig. 7: Hotspots polygons on the search space for patterns illustrated in Figure 2.a.1 and Figure 2.b.1, with $\theta \geq 0.3$

ing a sampling technique. The implementation procedure applied for each relatedness measure is discussed below:

Maximum CVT Association (MA) is identified based on the peak value among N sample locations.

To compute **AUC of CVT Association for a Variable Pair**, for computational efficiency, we approximated AUC by calculating the average interestingness for $n \times n$ sample thresholds, where n equidistant thresholds are sampled from the interval $[t_{low}, t_{high}]$ and another n equidistant thresholds are sampled from the interval $[t'_{low}, t'_{high}]$.

PIT can be implemented using the hotspot polygons discussed in the Section 3.2 using the following equation:

$$PIT(\theta) = \frac{\sum_{r \in R} area(r)}{area(TS)} \tag{21}$$

In Equation 21 PIT is defined as a function of θ that measures what fraction hotspot set R covers within the whole threshold space TS. Every hotspot in hotspot set R is always above θ.

However we can further simplify the computation using the following measure. Let, we have a set S of n thresholds pair from variables v and v'. We define PIT as follows:

$$PIT(\theta) = \frac{Count(\{y | y = I(\{(v\ t), (v'\ t')\}) \geq \theta \wedge (t, t') \in S\})}{n} \tag{22}$$

In Equation 22 we measure fraction of sample threshold pairs that are above the threshold θ.

3.4 Computational Complexity of the Norma Framework

Complexity of Agreement Calculation: Considering two sets of hotspots, each containing n hotspots, where $n << N$, the intersection operation exhibits a worst-case complexity of $\mathcal{O}(n^2)$, while the union operation demonstrates a time complexity of $\mathcal{O}(2n)$. Consequently, the comprehensive complexity for agreement calculation aligns with $\mathcal{O}(n^2)$. It is pertinent to clarify that the inherent computational complexity associated with polygonal intersection and union operations is beyond the scope of this discussion and thus is not explored in detail within this paper.

Complexity for CVT Patterns from a Variable Pair: The collective complexity for a single pattern is encapsulated as $\mathcal{O}(N) + \mathcal{O}(N) + \mathcal{O}(n^2))$ based on the complexity of hotspot and agreement calculation. Considering $n << N$, the overarching computational complexity attains a linear time order of $\mathcal{O}(N)$ and, while the space complexity is represented as $\mathcal{O}(N)$.

When incorporating an additional N sample thresholds for each variable pair, the computational complexity escalates to $\mathcal{O}(N^2)$.

4 Case Study

In this section, we conduct experiments to investigate the characteristics of *CVT* association patterns within variable pairs and relatedness across variable pairs. Our objective is to address the following questions:

- What knowledge CVT association analysis can extract that traditional methods like correlation analysis ignores?
- How similar are the CVT association analysis measures to the traditional frameworks such as correlation analysis?
- How CVT association analysis can be used to rank associations among different variable pairs?

4.1 Dataset Description

This study gathered twenty county-level variables from the John Hopkins Covid-19 data repository [16] and the New York Times COVID-19 data repository [17]. The variables from the John Hopkins dataset are, the percentage of the population with a Bachelor's Degree until 2018 ($Bachelor\ Degree\ Rate$), daily precipitation and temperature in 2019, population and household density compared to land area in 2010, unemployment rate in 2018 ($Unemployment\ Rate$), employment rates in various industries (agriculture, mining, construction, manufacturing, trade, transportation, information technology, fire services, and service providing) in 2018, median household income in 2018 ($Median\ Income$), and poverty rate in 2018 ($Poverty\ Rate$). Yearly average temperature and precipitation were calculated from the daily data. From the New York Times dataset, total COVID-19 cases and deaths until April 2022 were obtained. The variables $COVID-19\ Infection\ Rate$ and $COVID-19\ Death\ Rate$ were computed by dividing the respective variables using county-wise population. Despite the variables being collected at different time-frames, they are assumed to represent the socio-economic and climatic conditions of specific

Variable	Mean	Standard Deviation	Minimum	Maximum	25th Percentile	50th Percentile	75th Percentile
$COVID-19\ Infection\ Rate$	0.25	0.06	0.03	1	0.215	0.25	0.28
$COVID-19\ Death\ Rate$	0.004	0.002	0	0.012	0.0026	0.0037	0.005
$Average\ Temperature$	54.35	9.22	34	78	47	55	62
$Average\ Precipitation$	3.18	1.32	0	6	2	3	4
$Bachelor\ Degree\ Rate$	21.08	9.42	0	78	15	19	25
$Unemployment\ Rate$	3.64	1.43	1	18	3	3	4
$Poverty\ Rate$	14.95	6.27	3	56	10.5	14	18
$Median\ Income$	50942.99	13416.07	22679	136191	42232	48711	56521.5
$Government\ Employment\ Rate$	5.02	3.07	0	32	3	4	6
$Fire-service\ Employment\ Rate$	4.1	1.9	0	21	3	4	5
$Transportation\ Employment\ Rate$	4.96	2	0	24	4	5	6
$Technology\ Employment\ Rate$	0.91	0.87	0	17	0	1	1
$Service-provider\ Employment\ Rate$	42.38	6.97	5	82	38	42	47

TABLE 1: County level variable value distribution

geographic regions and are unlikely to vary greatly over a few years. Results from twelve variables among twenty is presented in this paper, along with their higher-level value distribution presented in Table 1. This will aid in understanding the stands of different thresholds from each CVT association in their respective value ranges.

4.2 Tools

Python is used for all implementation aspects related to the Norma framework. The Shapely polygon library is employed to manage spatial objects and compute spatial area coverage, while the Geopandas library is used for the visualization of shape objects. Additionally, the Scipy library is leveraged to measure correlations. The entire framework, encompassing both code and data, is publicly available at https://github.com/mmahin/ThresholdOptimization.git.

4.3 Comparative Study of Binary Correlation and CVT Association

4.3.1 Experimental Procedure

In this study, we conducted an analysis involving 190 variable pairs derived from twenty continuous variables. To analyze patterns from each variable pair, we created a maximum of 100 sample thresholds from each variable's domain, resulting in maximum (100×100) number of sample thresholds for each variable pair. However, due to many variables not having 100 unique thresholds, we restricted the number of thresholds to the actual number of unique values for those variables. Consequently, from 190 variable pairs, we analyzed a total of 234,694 patterns.

Furthermore, to ensure the selection of interesting patterns, we constrained the observation area coverage for each variable to a range of at most 50% down to a minimum of 1% using parameters $\alpha = 0.5$ and $\beta = 0.01$ as discussed in Section 2.5.

In this study, we conducted analysis using both Equation 2, Equation 17 and reported a subset of the results in Table 2. For each variable pair, we reported the Pearson correlation in column 2, maximum CVT association *MA* along with the respective thresholds in column 3 , to better understand the support for each *MA*, we additionally reported expected CVT association E_I for the thresholds on the maxima under the assumption of independence, along with the area coverage for the hotspots in column 4 and Lift for the maxima on column 5; additionally we reported area under the curve (AUC) for CVT association in column 6 and maximum threshold correlation Max_{TCorr} along with respective thresholds in column 7. Max_{TCorr} represents the maximum threshold correlation among all threshold correlation pattern within a variable pair. Furthermore, to compare the measures, we ranked each of the four measures: Pearson correlation, *MA*, AUC, and Max_{TCorr} based on their respective values among the 190 variable pairs, and the rankings are presented along with the CVT pattern in column 1.

4.3.2 Result Interpretation:

From Table 2, it is observable that CVT patterns provides a more detailed insights about relations among values within a variable pair. For example, based on Pearson correlation it can be interpreted that the variable pair $(Population\ Density, Household\ Density)$ has a very strong positive linear relationship within the value domain. However using CVT association we can further deduce relations specific to spatial domain, such as amount of spatial overlap for certain group of values. For example, based on *MA*, this variable pair has maximum spatial overlap of 100% for values above thresholds (267,111). These relation can be found even when correlation measure does not find any apparent linear relation. For instance, for the variable pairs $(Average\ Precipitation, Household\ Density)$, $(Trade\ Employee\ Rate, Household\ Density)$, the correlation measure is showing almost no apparent linear relation, but CVT association measure have found strong overlap of 71% and 56% on the thresholds that generates maximum CVT association.

We can further deduce how much these maxima deviates from any random settings using E_I and $Lift$. For example, {*(Mining Employee Rate t),(COVID-19 InfectionRate t)*} has an expected E_I of 0.17 resulting in a Lift of 1.8, almost twice than the random settings.The most high Lift of 1189.8 is observed for the maxima of pattern $\{(Population\ Density, Household\ Density)\}$, indicating very high difference from a totally non-related, random event. Additionally, we can say values above all presented maxima's are strongly associated as per discussion from section 2.6.2 ($I > E_I$ and $Lift > 1$).

The AUC measure gives average association strength based on all CVT association patterns

Pattern and Ranking within 190 Variable Pairs Rank Order:{Pearson Correlation, MA, AUC, Max_{TCorr}}	**Pearson Correlation**	***MA* on (t,t')**	E_I **on *MA* (area(H(v,t)), area(H(v',t')))**	**Lift on *MA***	AUC	Max_{TCorr}
{($Population\ Density\ t$),($Household\ Density\ t'$)} **Ranks:**{190,190,190,190}	0.99	1 on (267, 111)	0.0008 (0.03, 0.03)	1189.8	0.38	0.99 on (34, 21.3)
{($Average\ Precipitation\ t$),($Household\ Density\ t'$)} **Ranks:**{122,189,172,9}	0.04	0.72 on (3, 4.3)	0.21 (0.41, 0.5)	3.5	0.14	0.04 on (3, 112)
{($Average\ Precipitation\ t$),($Population\ Density\ t'$)} **Ranks:**{125,188,166,12}	0.04	0.71 on (3, 10.1)	0.21 (0.41, 0.5)	3.45	0.13	0.05 on (3, 237)
{($Manufacturing\ Employee\ Rate\ t$),($Population\ Density\ t'$)} **Ranks:**{67,187,113,42}	-0.08	0.64 on (6, 10.1)	0.25 (0.49, 0.5)	2.6	0.08	0.13 on (21, 94)
{($Manufacturing\ Employee\ Rate\ t$),($Household\ Density\ t'$)} **Ranks:**{69,186,121,30}	-0.07	0.64 on (6, 4.3)	0.25 (0.49, 0.5)	2.6	0.09	0.1 on (21, 108)
{($Trade\ Employee\ Rate\ t$),($Household\ Density\ t'$)} **Ranks:**{81,183,138,65}	-0.04	0.56 on (12, 4.3)	0.25 (0.5, 0.5)	2.22	0.1	0.18 on (16, 77)
{($Poverty\ Rate\ t$),{($COVID-19\ Death\ Rate\ t$) **Ranks:**{184,182,181,93}	0.44	0.55 on (12, 0.003)	0.27 (0.54, 0.5)	2.06	0.16	0.25 on (12, 0.006)
{($Mining\ Employee\ Rate\ t$),($COVID-19\ Infection\ Rate\ t$)} **Ranks:**{130,11,51,189}	0.06	0.17 on (2, 0.22)	0.09 (0.2, 0.48)	1.8	0.04	0.73 on (6, 0.31)
{($COVID-19\ Death\ Rate\ t$),($Population\ Density\ t$)} **Ranks:**{49,107,39,188}	-0.13	0.39 on (0.003, 0.2)	0.25 (0.5, 0.5)	1.6	0.04	0.69 on (0.005, 289.8)

TABLE 2: Two Threshold CVT association patterns $\{(v\ t), (v'\ t')\}$ in normal and spatial settings, while $area(H(v,t))$ and $area(H(v',t'))$ is restricted between $\alpha = 0.5$ and $\beta = 0.01$. Results are presented with Pearson correlation. Column 1 presents pattern name along with pattern ranking when all patterns are ranked based on the high to low values.

within a variable pair. For instance, the variable pair $(Population\ Density, Household\ Density)$ has an AUC of 0.38, substantially higher than the 0.04 from the variable pair $(COVID-19\ Death\ Rate, Population\ Density)$, indicating stronger spatial association among all values from the first pair.

Even in non-spatial environments, threshold correlation is able to find strong linear relation within the subset of values from the domain of two variables, which is overlooked when correlation is computed using the whole domains. For instance, for variable pairs $(Mining\ Employee\ Rate, COVID-19\ Infection\ Rate)$ and $(COVID-19\ Death\ Rate, Population\ Density)$, there is no apparent strong linear relationship. However, threshold correlation have found strong subset level positive linear relationship of 0.73 for values above the thresholds 6 for the variable $Mining\ Employee\ Rate$ and 0.31 for the variable $COVID-19\ Infection\ Rate$ from variable pair $(Mining\ Employee\ Rate, COVID-19\ Infection\ Rate)$. Similarly another strong subset level positive linear relationship of 0.69 is found for values above the thresholds 0.005 for the variable $COVID-19\ Death\ Rate$ and 289.8 for the variable $Population\ Density$ from variable pair $(COVID-19\ Death\ Rate, Population\ Density)$.

Measure1, Measure 2	Spearman rank correlation	Significance
Pearson Correlation, *MA*	0.45	$1.15e^{-10}$
Pearson Correlation, AUC	0.74	$7.7e^{-35}$
Pearson Correlation, Max_{TCorr}	0.22	$2e^{-3}$
MA, AUC	0.65	$6.4e^{-24}$
MA, Max_{TCorr}	0.12	0.09
AUC, Max_{TCorr}	0.22	0.003

TABLE 3: Spearman Correlation Test among the four measures.

4.3.3 Rank Correlation:

To compare different measures we ranked 190 variables pairs based on four measures and found dissimilarity in ranking. For example, the pattern $\{(Mining\ Employee\ Rate\ t)\{(COVID-19\ Infection\ Rate\ t)$ ranked 130^{th} for Pearson Correlation, 11^{th} for *MA*, 51^{th} for AUC , 189^{th} Max_{TCorr}, as shown in column 1 of the Table 2.

To further analyze relation among the association results we have applied Spearman rank correlation test on the ranks for 190 associated patterns.

Table 3 delineates the outcomes of the Spearman rank correlation test among the four measures. Observably, all measures exhibit positive correlations, with the majority also demonstrating strong linear relationship. The correlations range from a low of 0.12 to a high of 0.74, underscoring that while there is some agreement in the rankings provided by the measures, but they are not absolute. For example, correlation of 0.74 among Pearson correlation and AUC indicates both measures finds identical relationship with some low disagreements. Given that even the highest correlation does not approach perfect agreement ($r = 1$), and considering the variations across different measure pairings, we can infer that none of the measures render the others redundant.

5 Conclusion

This research introduced "Norma," a novel framework for association mining of continuous spatial variables. The framework introduces a new type of spatial pattern, Continuous Variable Threshold (CVT) pattern, which utilizes point-wise functions and measures spatial association among the values above two thresholds from two spatial continuous variables. The framework proposes a grid-based hotspot growing algorithm to find such association and introduces three measures MA, AUC, PIT to measure relatedness of variables based on mined CVT associations.

Additionally, it provides extension of CVT association for non-spatial environment. Furthermore, a case study demonstrated the knowledge discovery capacity of CVT associations by utilizing twenty county-level variables (spanning socio-economic, meteorological, and disease infection parameters) across the contiguous United States. Findings from the study reveal that CVT associations not only extracts more detailed summary of the relation among the values of two continuous variables when contrasted with traditional correlation measures but also offer a panoramic view of spatial associations among all values from two spatially continuous variables. This posits CVT association analysis as a potent alternative to correlation within the geospatial domain. Furthermore, another study involving rank correlation indicates that while the measures derived from CVT association analysis bear some similarities to those of the correlation measure, they are not redundant.

6 Future Work

The current CVT association framework is analyzed based on a single dataset with twenty variables, where the variables has values from polygons. On the other hand, most variables do not have high number of unique values, limiting the number of possible thresholds for the analysis. As a result, in future to find effectiveness of the CVT association framework, we need to analyze datasets that requires different pointwise functions other than polygons and where variables have large domain of values. Additionally, in this study we have employed grid-based sampling based techniques to compute CVT associations. However this method might not be very accurate for variables with very high number of thresholds. Consequently, we would need to develop optimization procedure that can handles very large number of thresholds. Moreover, we need to apply CVT pattern analysis to more dataset to establish its merit.

References

[1] K. McGarry, "A survey of interestingness measures for knowledge discovery," *The knowledge engineering review*, vol. 20, no. 1, pp. 39–61, 2005.

[2] J. Hipp, U. Güntzer, and G. Nakhaeizadeh, "Algorithms for association rule mining—a general survey and comparison," *ACM sigkdd explorations newsletter*, vol. 2, no. 1, pp. 58–64, 2000.

[3] P. A. Moran, "Notes on continuous stochastic phenomena," *Biometrika*, vol. 37, no. 1/2, pp. 17–23, 1950.

[4] N. Sauber, H. Theisel, and H.-P. Seidel, "Multifield-graphs: An approach to visualizing correlations in multifield scalar data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 917–924, 2006.

[5] L. Anselin, I. Syabri, O. Smirnov *et al.*, "Visualizing multivariate spatial correlation with dynamically linked windows," in *Proceedings, CSISS Workshop on New Tools for Spatial Data Analysis, Santa Barbara, CA*, 2002.

[6] L. Anselin, "Local indicators of spatial association—lisa," *Geographical analysis*, vol. 27, no. 2, pp. 93–115, 1995.

[7] C. F. Eick, R. Parmar, W. Ding, T. F. Stepinski, and J.-P. Nicot, "Finding regional co-location patterns for sets of continuous variables in spatial datasets," in *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, 2008, pp. 1–10.

[8] A. J. Izenman, "Review papers: Recent developments in nonparametric density estimation," *Journal of the american statistical association*, vol. 86, no. 413, pp. 205–224, 1991.

[9] Y. Xie, T.-b. Chen, M. Lei, J. Yang, Q.-j. Guo, B. Song, and X.-y. Zhou, "Spatial distribution of soil heavy metal pollution estimated by different interpolation methods: Accuracy and uncertainty analysis," *Chemosphere*, vol. 82, no. 3, pp. 468–476, 2011.

[10] C. F. Eick, M. Mahin, G. Chen, and H. Zhang, "On threshold correlation with application to studying the relationship of temperature and relative humidity," in *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, ser. GeoAI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 53–62. [Online]. Available: https://doi.org/10.1145/3615886.3627741

[11] A. Ram, S. Jalal, A. S. Jalal, and M. Kumar, "A density based algorithm for discovering density varied clusters in large spatial databases," *International Journal of Computer Applications*, vol. 3, no. 6, pp. 1–4, 2010.

[12] F. Akdag and C. F. Eick, "An optimized interestingness hotspot discovery framework for large gridded spatio-temporal datasets," in *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, 2015, pp. 2010–2019.

[13] C. Deng, J. Song, R. Sun, S. Cai, and Y. Shi, "Griden: An effective grid-based and density-based spatial clustering algorithm to support parallel computing," *Pattern Recognition Letters*, vol. 109, pp. 81–88, 2018.

[14] M. Wang, A. Wang, and A. Li, "Mining spatial-temporal clusters from geo-databases," in *Advanced Data Mining and Applications: Second International Conference, ADMA 2006, Xi'an, China, August 14-16, 2006 Proceedings 2*. Springer, 2006, pp. 263–270.

[15] H. Darong and W. Peng, "Grid-based dbscan algorithm with referential parameters," *Physics Procedia*, vol. 24, pp. 1166–1170, 2012.

[16] B. D. Killeen, J. Y. Wu, K. Shah, A. Zapaishchykova, P. Nikutta, A. Tamhane, S. Chakraborty, J. Wei, T. Gao, M. Thies, and M. Unberath, "A County-level Dataset for Informing the United States' Response to COVID-19," Apr. 2020.

[17] "The New York Times covid-19 data," https://github.com/nytimes/covid-19-data, 2023, accessed on June 30, 2022.

Md Mahin received the B.Sc. degree in Computer Science and Telecommunication Engineering from Noakhali Science and Technology University, Bangladesh, in 2014. He is an ongoing PhD student in the Department of Computer Science at University of Houston, Texas. His research interest includes unsupervised pattern mining, association mining, clustering, geospatial data mining. Additionally he is enthusiastic towards deep learning, generative models, language models and Reinforcement Learning.

Christoph F. Eick received the MS and PhD degrees in computer science from the University of Karlsruhe in Germany, in 1979 and 1984, respectively. Currently, he is an associate professor in the Department of Computer Science at the University of Houston. His research interests include clustering, spatial data mining, and association analysis.