Towards Geographically Robust Statistically Significant Regional Colocation Pattern Detection

Subhankar Ghosh, Jayant Gupta, Arun Sharma, Shuai An, Shashi Shekhar
{ghosh117,gupta423,sharm485,an000033,shekhar}@umn.edu
University of Minnesota
Minneapolis, Minnesota, USA

ABSTRACT

Given a set S of spatial feature-types, its feature-instances, a study area, and a neighbor relationship, the goal is to find pairs <a region (r_g), a subset C of S> such that C is a statistically significant regional colocation pattern in region r_g. For example Caribou Coffee and Starbucks are significantly co-located in Minneapolis but not in Dallas at present. This problem has applications in a wide variety of domains including ecology, economics, and sociology. The problem is computationally challenging due to the exponential number of regional colocation patterns and candidate regions. The current literature on regional colocation pattern detection has not addressed statistical significance which can result in spurious (chance) pattern instances. In this paper, we propose a novel technique for mining statistically significant regional colocation patterns. Our approach determines regions based on geographically defined boundaries (e.g., counties) unlike previous works which employed clustering, or regular polygons to enumerate candidate regions. To reduce spurious patterns, we perform a statistical significance test by modeling the observed data points with multiple Monte Carlo simulations within the corresponding regions. Using Safegraph POI dataset, this paper provides a case study on retail establishments in Minnesota for validation of proposed ideas. The paper also provides a detailed interpretation of discovered patterns using game theory and regional economics.

CCS CONCEPTS

• **Information systems**; • **Geographic information systems**; • **Statistical Significance**; • **Regional Economics**;

KEYWORDS

Regional Colocation pattern, Statistical Significance, Neighborhood Graph, Spatial Heterogeneity, Game Theory

ACM Reference Format:
Subhankar Ghosh, Jayant Gupta, Arun Sharma, Shuai An, Shashi Shekhar. 2022. Towards Geographically Robust Statistically Significant Regional Colocation Pattern Detection. In *The 5th ACM SIGSPATIAL International Workshop on GeoSpatial Simulation (GeoSim '22) (GeoSim '22), November 1, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3557989.3566158

GeoSim'22, Seattle, WA, USA,

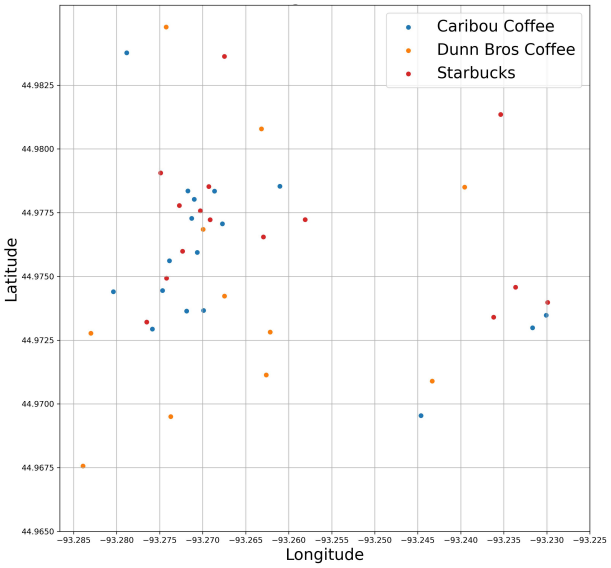
ACM ISBN 978-1-4503-9537-3/22/11…$15.00
https://doi.org/10.1145/3557989.3566158

1 INTRODUCTION

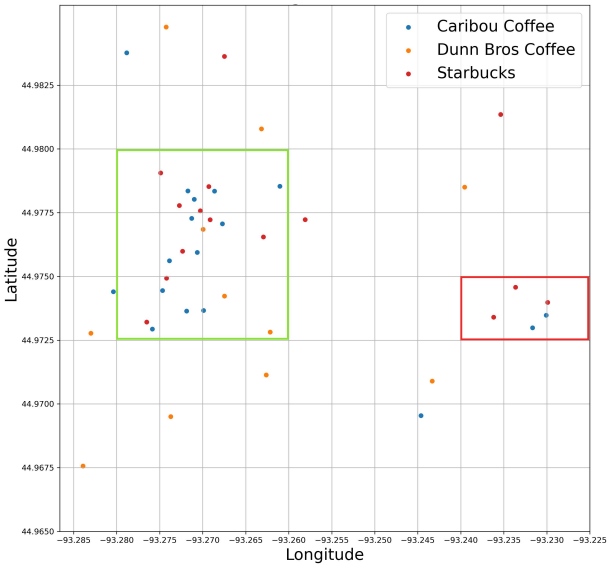
Given instances of a set S of spatial features (e.g., coffee shops, restaurants), a study area, and a neighbor relationship (e.g., geographic proximity), the goal is to identify pairs <region r_g, subset C of S> such that regional colocation instances of C are statistically significant in that region r_g. Figure 1(a) shows a set of instances input into a regional colocation miner, consisting of three different spatial feature-types (Caribou Coffee, Dunn Brothers and Starbucks) in the Twin Cities region (i.e., Minneapolis and St. Paul, MN). As shown in Figure 1(b), the output is a set of statistically significant regions of interest where the features are regionally co-located. The patterns are statistically significant at a confidence level of 95% (p-value ≤ 0.05). The region within the green polygon lies in Minneapolis and shows a strong regional colocation between all the three features. Whereas, the region within the red rectangular polygon lies in St. Paul and shows regional colocation between two features (i.e., Caribou Coffee and Starbucks). Rest of the area within the map shows very little spatial interaction between these features.

The problem of mining statistically significant regional colocation patterns has societal importance with applications in retail, public health, ecology, public security, transportation, etc. For economic reasons (e.g., customer reach) retail establishments like fast food chains or coffee shops often co-locate. These retail brands are seen as substitutes for each other and are direct competitors. It is counter-intuitive to see direct competitors locate close to each other. Thus, empirically finding significant regional colocation patterns among competing retail stores within certain economic boundaries has tremendous value for retail analysis. Besides retail, regional colocation patterns are important in public health, for example, the 1894 plague outbreak in Hong Kong where the infected cases co-located with the rats near the port helped inform local public health policies. In addition, finding regional colocation patterns across different biological species helps to identify new interdependence relationships governed by symbiosis. Table 1 provides a few additional application domains and their example use-cases.

The problem of regional colocation pattern detection is computationally challenging due to an exponential number of candidate patterns as well as a large number of candidate regions. For example, our dataset consists of 1473 different retail brands and their locations in Minnesota resulting in 2^{1473} different candidate patterns. In addition, since our space is continuous, any space partitioning based approach would lead to an infinite number of candidate region subsets. Significance testing also adds to the computational complexity as we compare the value of our prevalence measure



(a) Input: Instances of Caribou Coffee, Dunn Bros and Starbucks in a region of Twin Cities



(b) Output: Subregions within which all or subsets of Caribou Coffee, Dunn Bros, and Starbucks co-located.

Figure 1: Regions within which all or subsets of Caribou Coffee, Dunn Bros and Starbucks co-locate in Twin Cities

for a candidate pattern against the value obtained in multiple null hypotheses. For a 95% statistical confidence we need to perform 99 Monte Carlo simulations to generate the null hypotheses while for a 99% confidence we need 499 Monte Carlo simulations. Besides computational cost, challenges arise from explainability of system identified regional patterns and spatially-variable thresholds. For example, explaining an output regional pattern may prove difficult

Table 1: Regional colocation applications.

Application Domain	Example
Retail	<China, McDonald's and KFC>, <USA, McDonald's and Jimmy John's>
Public Health	<Ports, Plague and rats>, <Middle East, Middle East Respiratory Syndrome (MERS) and MERS-CoV>
Ecology	<Indian/Pacific Ocean, Anemone and Clownfish>, <Nile river delta, Nile Crocodile and Egyptian Plover>
Public	<Region around bars, Assault crimes and drunk driving>
Transportation Industry	<Near bus depots, High NO_x concentrations and buses>

without additional local information. However, such interpretations are useful to domain experts such as economists, biologists, etc. Similarly, consensus on predefined threshold is difficult as the neighbor relationship (distance between co-located features) is not consistent across different locations (e.g., retail colocations in New York vs Minneapolis).

Previous works [3, 10] on regional colocation pattern detection can be divided into two categories. The first category is data unaware space partitioning such as Quadtrees and grids. The approaches in this category do not consider spatial distribution and an inappropriate partitioning might divide potential localities, e.g. a Quadtree [3] based approach might not completely capture the Minneapolis Downtown region of the colocation pattern around the green polygon in Figure 1(b) and might break up geographic entities, such as counties, cities, states. Also, prior works [3, 6, 13] incorporating these techniques lacked statistical significance and focused on data existing on a projected or planar surface.

The second category uses clusters of colocation instances but ignores regions without clusters. This approach is susceptible to spatial auto-correlation and enumerates patterns which might not be statistically significant resulting in the enumeration of spurious (chance) regions. Such methods [5, 6] perform poorly when the feature instances are uniformly distributed (i.e., not clustered) in the study area. Finally, they are dependent on a pre-defined distance and participation index threshold which is geographically inconsistent. For example, New York being a busier city than Minneapolis, the expected distance between features of candidate patterns is much less. Paper [1] discussed statistical significance in the context of global colocation pattern detection, but does not mention colocation patterns which might be local. In this paper we focus on statistically significant regional (or local) colocation patterns which might have a positive spatial interaction only within certain regions or subsets of the study area but not within the whole study area. This is done by enumerating candidate regions which are geographically and economically explainable.

Overall, we propose a spatial graph-based approach where, we use statistical significance and a dynamic distance threshold parameter for enumerating candidate regions. Our approach preserves topological relation between regions, reduces chance patterns, and

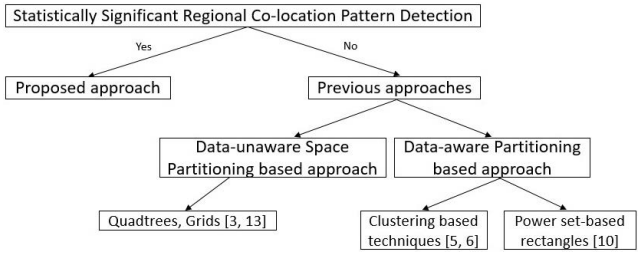


Figure 2: A decision tree comparing our work with previous approaches.

does not rely on predefined distance or participation index (pi) threshold parameters. In addition, we provide a game-theory based interpretation of the identified regional colocation using local demographic and economic information.

Contributions: The contributions of this paper are as follows:

– The paper formally defines the problem of statistically significant regional colocation pattern detection.

– The paper proposes a spatial graph-based approach to enumerate spatial aware regions of interest which are statistically significant and economically explainable.

– The paper provides a case study on retail establishments in Minnesota using the Safegraph POI dataset.

– The paper provides a domain interpretation of the discovered patterns using game theory. An additional interpretation based on county level macroeconomic indices can be found in the Appendix.

Scope: For simplicity, this paper focuses on colocation patterns consisting of two or three different features (retail brands). However, the proposed technique can apply to larger feature sets as well. In our experiments, we enumerated regions based on contiguous collection of counties. Nevertheless this work can be extended to the other types of neighboring regions (e.g., grids).

Organization: The paper is organized as follows: Section 2 introduces basic concepts and the problem statement. In Section 3 we provide an overview of the maximal sub-graph based approach to finding statistically significant regional colocation patterns. A Case study is presented in Section 4 and Section 5 provides a domain interpretation using game theory. A literature survey appears in Section 6. Finally, Section 7 concludes this work and briefly lists potential future directions.

2 BASIC CONCEPTS AND PROBLEM DEFINITION

2.1 Basic concepts for Colocation Detection

A **feature instance** is a geo-located spatial entity which is a type of boolean feature f with a geo-reference point location p (e.g., latitude, longitude), represented as $<f, p>$. Multiple instances of a feature are represented as f_i and can be related to other feature instances f_j via a **neighborhood relation** $\mathcal{R}$. For example, geographic proximity is represented as $\mathcal{R}_{f_i,f_j} \leq \theta$, where θ is the neighborhood threshold. In a **neighborhood graph**, we represent features which satisfy such relations as a *node* and this relationship between two related features as an *edge*.

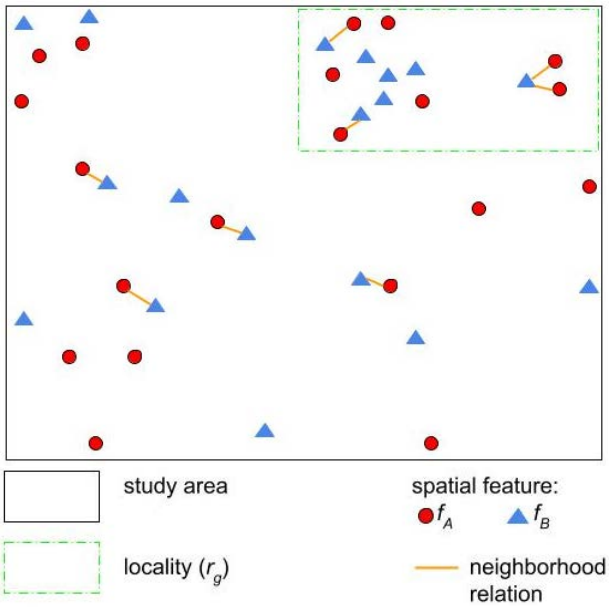


Figure 3: Colocation pattern instances and candidate locality

An instance of a **colocation** satisfies the neighborhood relation $\mathcal{R}$ and forms a **clique**. A **colocation candidate** C is a set of features defined in the given study area (S_A) or a region (r_g) where $r_g \in S_A$. For example, figure 3 shows 20 spatial objects of type f_A (circle) and 18 spatial objects of type f_b (triangle). It also shows 8 instances of colocation pattern $\{f_A, f_B\}$.

A **Participation Ratio (pr)** is the ratio of feature instances participating in a relation $\mathcal{R}$ to the total number of instances inside the study region (S_A). For a given colocation candidate C and feature f it is represented as $pr(f, C)$. Mathematically, it can be written as follows,

$$pr(f, C) = \frac{number\ of\ participating\ feature\ f\ instances}{total\ number\ feature\ f\ instances} \quad (1)$$

For example, in Figure 3, $pr(f_A, \{f_A, f_B\})$ and $pr(f_B, \{f_A, f_B\})$ are 8/20 and 7/18 respectively.

Participation ratio within a region (r_g) is defined as $pr(f, [r_g, C])$. For example, $pr(f_A, [r_g, \{f_A, f_B\}])$ and $pr(f_B, [r_g, \{f_A, f_B\}])$ defined in *locality* r_g is 4/7 and 3/7 respectively.

A **Participation Index (pi)** is the minimal participation ratio of all feature types in a colocation candidate.

$$pi(C) = \min_{f \in C}(pr(f, C)) \quad (2)$$

For example, in figure 3, $pi(\{f_A, f_B\}) = min(8/20, 7/18) = 7/18$. The participation index quantifies the spatial interaction within features.

Colocation patterns are the set of prevalent colocation candidates i.e., candidates comprised of features which have a high positive spatial interaction between them.

A **regional colocation pattern** [10] is a pair of region (r_g) and colocation pattern (C), i.e., $[r_g, C]$ where the features in pattern C have a positive spatial interaction in r_g.

Regional Participation Index is the minimal participation ratio of all feature types in the colocation candidate within region r_g.

$$pi([r_g, C]) = \min_{f \in C}(pr(f, [r_g, C])) \quad (3)$$

For example in Figure 3, $pi([r_g, \{f_A, f_B\}]) = min(3/7, 4/7) = 3/7$.

2.2 Basic concepts for Statistical Significance

Framework: A statistically significant colocation miner determines whether an assigned positive spatial interaction between features is statistically significant or could have been observed if the features were in complete spatial randomness (CSR). Figure 4 shows a flowchart of the steps involved in testing if a candidate pattern is a statistically significant global colocation pattern or not. We later expand this framework to detect statistically significant regional colocation patterns (Section 3).

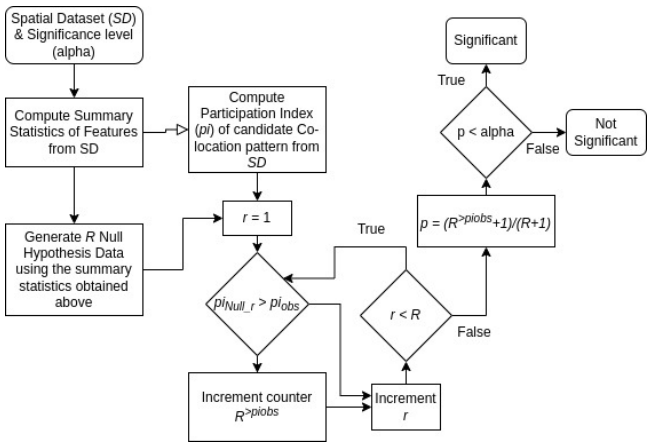


Figure 4: Flowchart depicting the steps involved in the statistical significance test of a candidate colocation pattern

Complete Spatial Randomness (CSR): In CSR:

(1) Every feature instance has an equal probability of existing at any point in the study area.
(2) The location of any instance is unaffected or independent of the location of any other instance in the study area.

As participation index (pi) is used to quantify the strength of a spatial interaction. The objective is to determine the probability of a pattern's pi in the observed data. Assuming the features were independent of each other, if the probability (i.e., *p-value*) is low we consider the pattern statistically significant. We assign an observed pi value for a candidate pattern as statistically significant at level α if the $p\text{-}value \leq \alpha$.

Null Model Design: Null Hypothesis models complete spatial randomness and our null model is designed as follows.

- For equal probability of feature instances, we have generate an equal number of instances of each feature in every candidate region.
- For feature instance independence, We sample the instances from a Poisson Point Process [9]. To check for acceptable auto-correlation, we use Pair Correlation Function (pcf) upto a distance of 2000 meters.

Figure 5(b) gives an example of our Null model.

Point distribution: A collection of geo-distributed points referring to an event (e.g., road accident) in a spatial domain.

Point Process (*PP*): It is a statistical process which governs the data generation of a point distribution. It defines the probability distribution of a point over a region. For example, a homogeneous point process such as CSR has an equal probability for each point existing at any location in the study area. Point processes are essential for defining a null or alternate hypothesis which we utilize for our statistical significance test.

Poisson Point Process (*PPP*): A point process PP defined on some underlying space S_P is a Poisson Point Process with intensity Λ if it has the following properties:

(1) The number of points in a bounded Borel set (bounded sets that can be constructed from open or closed sets by repeatedly taking countable unions and intersections) $B \subset S_P$ is a Poisson random variable with mean $\Lambda(B)$.
(2) The number of points in n disjoint Borel sets forms n independent random variables.

A PPP can be defined on any generalized mathematical spaces. An essential property of PPP is that the number of points of the point process located in two (or more) disjoint regions form independent random variables. This property results in independent scattering or complete independence.

Statistical Significance Test: Let $pi_{\emptyset}(C)$ denote participation index for pattern C in the Null Hypothesis and $pi_{obs}(C)$ represent participation index for candidate colocation C in the observed data. Then, we compute the following probability [1],

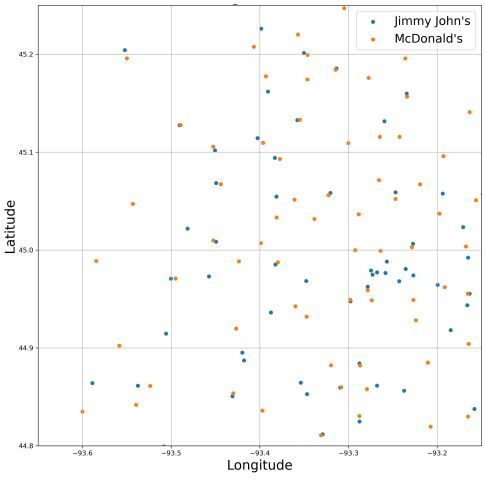
$$p = pr(pi_{\emptyset}(C) \geq pi_{obs}(C)) = \frac{R^{\geq pi_{obs}} + 1}{R + 1},$$

where $R^{\geq pi_{obs}}$ represents the number of Monte Carlo simulations within which the participation index ($pi_{\emptyset}(C)$) for pattern C is greater than that in the observed data ($pi_{obs}(C)$), R represents the total number of Null Hypotheses datasets generated. If $p \leq \alpha$, we consider $pi_{obs}(C)$ as statistically significant at level α.

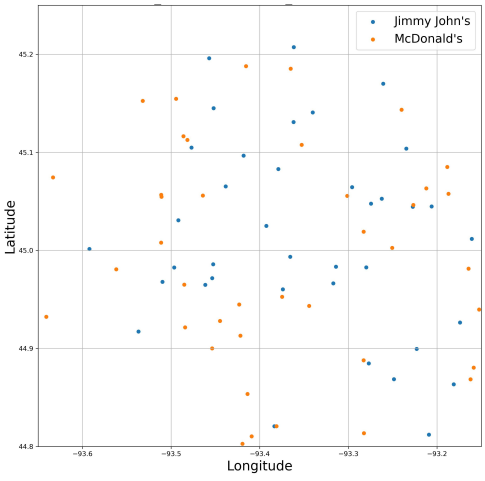
Regional Statistical Significance Test: In this test we perform the significance test as above using simulated (i.e., computer generated) candidate regions. For example, if we are trying to determine if f_A and f_B are statistically significant in locality r_g (Figure 3), we generate null hypothesis samples within its boundary and use the participation index result from each sample to information to perform the significance test for r_g.

Figure 5(a) shows a plot of different instances of Jimmy John's and McDonald's in Hennepin County and Figure 5(b) displays one of the R different Null Hypotheses which are used to compare the participation index of the colocation pattern $\{Jimmy John's, McDonald's\}$ in the observed data. At a distance of 800 meters the observed participation index (pi_{obs}(["Hennepin", $\{Jimmy John's, McDonald's\}$])) is 0.429, while for one of the generated null hypothesis ($pi_{\emptyset_{52}}$ (["Hennepin", $\{Jimmy John's, McDonald's\}$])) it is 0.0476. where $pi_{\emptyset_{52}}$ is the 52^{nd} Null Hypothesis.

For a statistical confidence of 95% the following inequality should hold:



(a) Observed data of ["Hennepin", {Jimmy John's, McDonald's}].



(b) Testing null hypothesis for ["Hennepin", {Jimmy John's, McDonald's}].

Figure 5: Significance testing of regional colocations.

$$\left[\sum_{r=1}^{R} \mathbb{1}(pi_{obs}(r_g, \{f_A, f_B\}]) \leq pi_{\emptyset_r}([r_g, \{f_A, f_B\}])\right] < 5 \quad (4)$$

where $R = 100$ refers to the total number of Monte Carlo simulations, r_g is the region of interest and $\mathbb{1}$ denotes the Indicator function. We can compute R from α using $\alpha(R+1) = 5$ [2].

2.3 Other Basic Concepts

Maximal Connected Sub-graph: For a simple graph $G = (V, E)$, the maximum connected component of G is the maximal connected sub-graph H. Sub-graph H of G is connected if $\exists$ a path for every pair of distinct vertices $u \in V_H, v \in V_H$ in H.

If $G = G_1^{n_1} \cup G_2^{n_2} \cup G_3^{n_3} \cup \ldots$ where $G_i^{n_i}$ represents the i^{th} connected component of G, then

(1) $n_{i_{max}} = max(n_i | G_i^{n_i} \subset G)$
(2) $H = \{G_j^{n_j} | G_j^{n_j} \subset G, n_j = n_{i_{max}}\}$.

2.4 Formal Problem Formulation

The problem of statistically significant regional colocation pattern detection is formulated as follows:

Input:

(1) A set of spatial-feature N geo-located spatial feature instances.
(2) A Study Area S_A composed of predefined regional boundaries (e.g., county).
(3) A statistical significance level α.
(4) Neighbor relationship ($\mathcal{R}$).

Variables: Distance between feature instances, $d \in (\theta_l : \theta_u)$, which is data driven

Output:

(1) Statistically Significant Colocation Patterns $C : \{f_A, f_B, \ldots\}$ (if they exist)
(2) Region r_g such that $r_g \subset S_A$

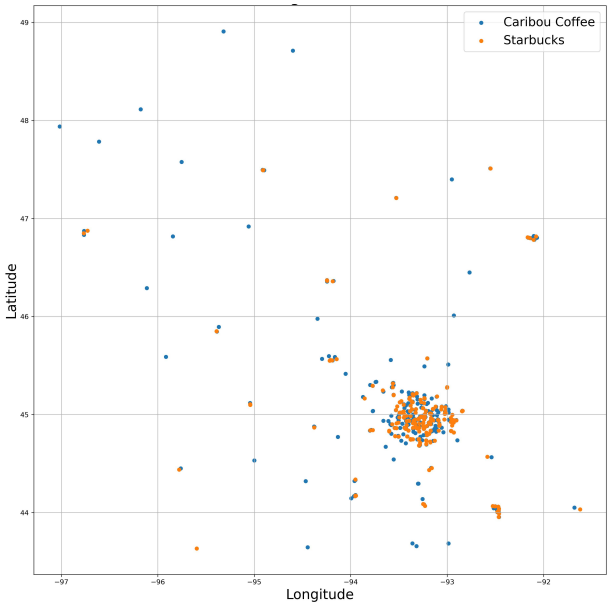
Objective: Reducing chance patterns.
Constraints: Correctness and Completeness.

Figure 6(a) shows the input with two types of features and their instances. Figures 6(b) shows the regions where the regional colocation of the two features is significant. Figure 6(c) shows the refined output with the largest contiguous region where the regional colocation is significant. Finally, Figure 6(d) shows a sub-graph representation of the contiguous output regions.

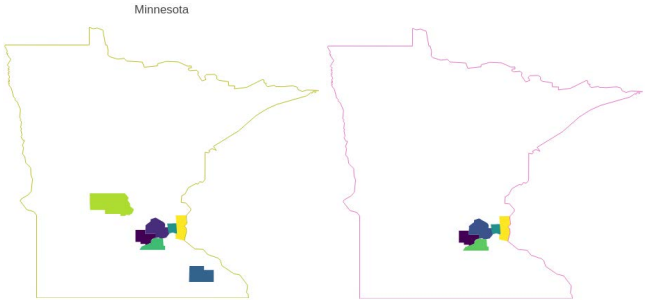
3 MAXIMAL SUB-GRAPH BASED APPROACH WITH STATISTICAL SIGNIFICANCE

We represent our dataset as a graph $G = (V, E)$. Here the vertices V represent the counties which are our candidate regions, while the edges E denote the neighborhood relationship between the counties. Our candidate regions can be composed of one or more counties. Then we find the largest connected sub-graph composed of neighboring counties which form a region within which a candidate pattern C is statistically significant.

Detecting Prevalent Regions: Our regions of interest are composed of counties or a union of neighboring counties. We start by considering counties with at least 3 instances of each of the features which comprise the colocation pattern. Our experiments primarily focus on retail establishments with the maximum distance between instances of features being 2000 meters. This is necessary since with increasing distance the participation index approaches 1. Then, we perform a statistical significance test for the candidate colocation patterns within each county. After that, we form an undirected unweighted graph *(G = V, E)* using the counties which share a geographic border between them. This graph representation makes it easier to find the largest region composed of counties within which the candidate pattern is statistically significant. For example in Figure 6(b), we find that at a distance of 400 meters between the features (i.e., Caribou Coffee and Starbucks) 7 counties, namely 'Carver', 'Hennepin', 'Olmsted', 'Ramsey', 'Scott', 'Stearns', and 'Washington' provide statistical significance for the pattern. This results in multiple sub-regions where pattern are statistically significant.

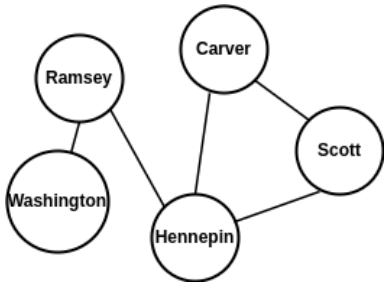


(a) Input locations of Caribou Coffee & Starbucks in Minnesota



(b) Output regions where pattern is significant

(c) Contiguous output regions where pattern is significant



(d) Sub-graph representation of Contiguous output regions where pattern is significant

Figure 6: Significance testing of regional colocation for [r_g, {Caribou Coffee, Starbucks}]

Since all counties within which a pattern is statistically significant may not be neighbors, we find all the contiguous regions composed of these counties which share geographic borders among them. By applying spatial join we can form larger regions composed of these counties within which the pattern is significant. Our final output is the **largest connected component** from the sub-graph obtained above and can be found using a **graph traversal algorithm** (e.g., breadth-first search). Figure 6(c) shows the largest contiguous sub-region composed of 5 counties 'Carver', 'Hennepin', 'Ramsey', 'Scott', 'Washington' within which the pattern is statistically significant and Figure 6(d) presents the maximal sub-graph representation of the same. We further validate our output region by performing a significance test on the area obtained from the spatial join of these significant counties.

Statistical Significance: Significance testing ensures that a pattern is not detected happen by chance. Thus, we want to find prevalent patterns which are rare if the constituent features are spatially independent of each other. To do this we compare the value of a prevalence measure (e.g., participation index *pi*) of a candidate pattern (C) in the observed data against that in the multiple Null Hypotheses data. These Null Hypotheses model the complete spatial randomness and have similar summary statistics (e.g., intensity i.e., instances per unit area, pair correlation function) as our observed data. Since we want the data points in the Null Hypotheses to have no spatial interaction we sample them from a Poisson Point Process with the same intensity as that of the observed data. To test for regional colocation patterns we need to ensure that the simulated data is being generated in a simulated region whose area is of a similar size as that of our region of interest (i.e., counties). For each candidate colocation pattern we then calculate the participation index in observed data and that in each of the Null Hypotheses. For a statistical confidence level of 95% we want the number of Null Hypotheses where the *pi* is greater than that in the observed data to be less than 5 (for 100 generated Null Hypotheses).

Algorithm 1 provides the pseudo-code of maximal sub-graph based approach to find statistically significant regional colocations.

4 CASE STUDY ON SAFEGRAPH POI DATASET

The goal of the experiment is to provide a case study of the proposed Maximal Connected Sub-graph algorithm on the real-world retail establishments based on null hypothesis and varying parameter threshold (e.g., distance) to validate our proposed approach. We provide extensive interpretation of results via validation questions related to changing parameters, geographic explainability and micro-economics. Figure 7 shows the overall validation framework with detailed comparison analysis description.

Real World Dataset: We used data from SafeGraph, a mobility data vendor who provided anonymized aggregated location data to researchers studying the effects of COVID-19 on citizen mobility patterns towards numerous Points Of Interests (POIs). The dataset consists of 1473 retail brands in Minnesota with USPS (776) and Western Union (548) having the maximum number of instances. To ensure that features in our enumerated regions had a significant presence, we only considered counties which had at least 3 instances of each feature/retail brand.

For example, experiments were performed on colocation patterns consisting of two (*e.g.*, Caribou Coffee, Starbucks) or three (*e.g.*, Caribou Coffee, Dunn Bros, Starbucks) features which can

Algorithm 1 Maximal Sub-Graph based approach

Input:
A Spatial Dataset S consisting of features $\{f_A, f_B, ...\}$
Statistical significance level α
Maximum Pattern size N
Output:
A Colocation pattern C which is a subset of features $\{f_A, f_B, ...\}$ from S
Region r_g within which the pattern C is statistically significant
Variables:
Distance between feature instances $d \in (\theta_l : \theta_u)$

procedure :
  for each: f_k in $\{f_A, f_B, ...\}$ **do**
    Generate R (dependent on α) null hypotheses from its summary statistics in the study area
  for each: candidate pattern $C_m \in \{C_1, C_2, ..., C_M\}$ **do**
    for distance $d \in (\theta_l : \theta_u)$ **do**
      for each: county ∈ county list **do**
        Reset counter $R^{\geq pi_{obs}}$ to 0
        Calculate pi_{obs} (observed data) for C_m in county at distance d
        for i ∈ $[1, R]$ **do**
          Calculate $pi_{\emptyset}$ of C_m in R^{th} null hypothesis
          if $pi_{\emptyset} \geq pi_{obs}$ **then**
            Increment $R^{\geq pi_{obs}}$
        $p - value_{C_m} = \frac{R^{\geq pi_{obs}} + 1}{R+1}$
        if $p - value_{C_m} \leq \alpha$ **then**
          Add county to Significant Counties
      Use graph traversal (*BFS*) to find connected significant components
      if number of connected significant components < 1 **then**
        continue
      else
        r_g = largest connected significant component
      Add C_m and r_g to output list
      Continue to next pattern C_{m+1}

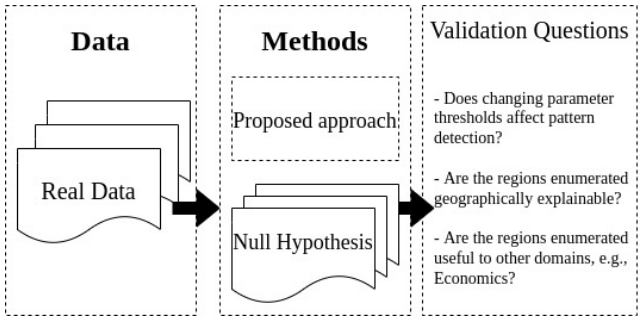


Figure 7: Overall validation framework

be considered as interchangeable. Our approach is applicable to patterns with higher numbers of constituent features as well.

Null hypothesis generation Our Null Hypothesis models the complete spatial randomness where there exists no spatial interaction between the features which constitute the candidate pattern. The main utility of this data is to ensure that our output patterns have a very low probability of existing if the features within the pattern were spatially independent of each other. To check for auto-correlation within the individual features we used the Pair Correlation Function (PCF) or $g(d)$. When $g(d) > 1$, it suggests there is clustering at distance d within the feature instances, while $g(d) = 1$ represents complete spatial randomness. Since our experiments were primarily focused on retail brands we found almost no auto-correlation within instances of the same feature till 2000 meters. For example, no two Starbucks were located very close to each other. This makes sense from an economic perspective as the operation cost of a larger-sized Starbucks shop is lower than the costs of two smaller Starbucks combined because of the scale of economy. Therefore, for these features, i.e. those without any spatial interaction within the feature instances, we generated the Null Hypothesis data by fitting a Poisson point process to the observed data with similar intensity. This ensures that within each candidate locality, each constituent feature of a pattern has the same number of instances as that in the R different Null Hypotheses generated. Figure 5(a) shows an example of our original data for Jimmy John's and McDonald's while Figure 5(b) shows one of the Monte Carlo simulations to model the corresponding Null Hypothesis.

4.1 Experimental Results

4.1.1 Effect of dynamic distance threshold on output region. For the pattern $[r_g, \{f_A, f_B\}] : [r_g, \{Caribou\ Coffee, Starbucks\}]$, where r_g:{Hennepin County, Scott County} we found $pi_{obs,d=200}(Hennepin, \{Caribou\ Coffee, Starbucks\}) = 0.34$ while $pi_{obs,d=200}(Scott, \{Caribou\ \ Coffee, Starbucks\}) = 0.25$ at a distance of 200 *meters* between the feature instances with 42 *and* 2 instances of the pattern in those counties respectively. With an increase in the distance threshold we get a larger region within which the pattern is statistically significant. For example at a distance of 400 meters between the instances of the features, we get a significant region composed of Carver, Hennepin, Ramsey, Scott and Washington Counties.
$pi_{obs,d=400}(Carver, \{Caribou\ Coffee, Starbucks\}) = 0.5$,
$pi_{obs,d=400}(Hennepin, \{Caribou\ Coffee, Starbucks\}) = 0.51$,
$pi_{obs,d=400}(Ramsey, \{Caribou\ Coffee, Starbucks\}) = 0.34375$,
$pi_{obs,d=400}(Scott, \{Caribou\ Coffee, Starbucks\}) = 0.375$ and
$pi_{obs,d=400}(Washington, \{Caribou\ Coffee, Starbucks\}) = 0.41176$
with $4, 75, 11, 3$ *and* 7 *instances* of the pattern in the counties respectively. Figures 6(b), 6(c) and 6(d) present the output significant regions, output largest contiguous region and sub-graph representing the contiguous region for the pattern at a distance of 400 meters respectively.

To ensure that the union of contiguous counties also formed a significant region we performed a statistical significance test on them as well. Our final output shown in Figure 6(c) represents the region within which the pattern is significant not only in the individual counties but also in the union of the contiguous counties. To

ensure this, we applied a spatial join to obtain a new boundary for the contiguous counties and then applied the statistical significance test to this new region. Figure 8 presents a plot of the change in the participation index in the observed data in this final output region ($pi_{obs}(Output\ region, \{Caribou\ Coffee, Starbucks\})$) against the mean of the participation index in the Null hypotheses generated for the same region ($mean(pi_{Null}(Output\ region, \{Caribou\ Coffee, Starbucks\}))$) with varying distance. In our experiments with different patterns, this contiguous region was found to be statistically significant, though it is possible to obtain subsets of this contiguous region by applying a graph traversal algorithm and check for significance within them.

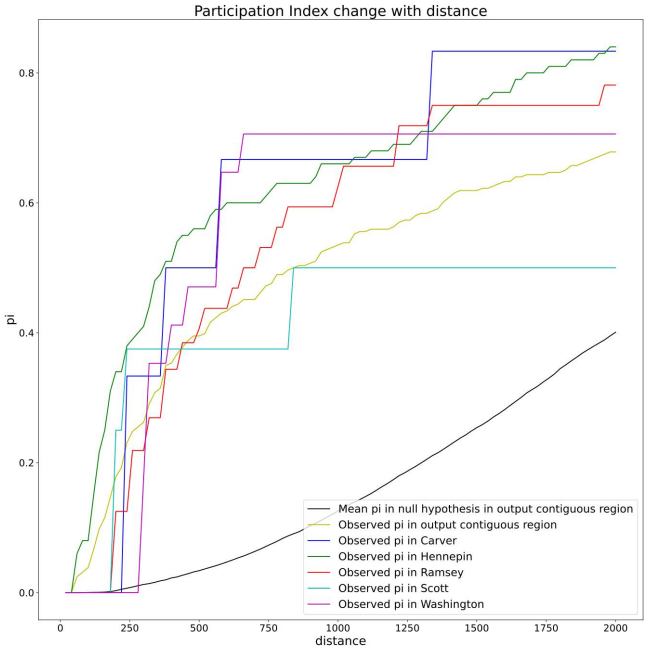
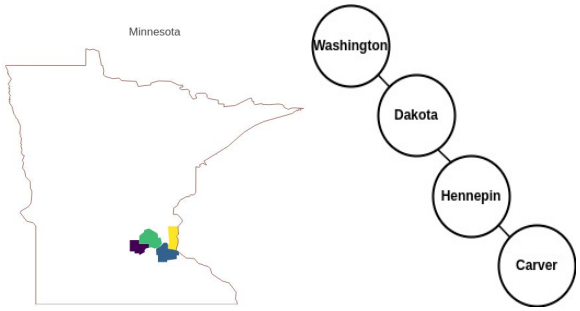


Figure 8: Change in participation index for pattern {Caribou Coffee, Starbucks} in observed data (pi_{obs}) vs participation index in Null Hypotheses ($pi_{mean\ of\ Null\ Hypotheses}$) against *distance* in meters

Our experiments were performed on patterns consisting of 2 and 3 features. This approach is applicable to patterns consisting of more features as well, although the computational cost increases with the pattern size.

For the pattern $\{f_A, f_B, f_C\} : \{Jimmy John's, McDonald's, Subway\}$, the colocation was evident in different regions across the study area. At a distance of about 1200 meters between the feature instances we observed colocations in Hennepin and Washington Counties, with 17 instances of the pattern in Hennepin and 4 instances in Washington at this distance. On increasing the distance threshold to 1600 meters we obtained richer results. Figure 9(a) shows 'Carver', 'Dakota', 'Hennepin' and 'Washington' counties across which the pattern was found to be statistically significant while Figure 9(b) shows the sub-graph representation of the region.

$pi_{obs,d=1600}(Carver, \{Jimmy John's, McD, Subway\})$ = 0.2857, $pi_{obs,d=1600}(Dakota, \{Jimmy John's, McD, Subway\})$ = 0.3428,



(a) Significant regions for the pattern {Jimmy John's, McDonald's, Subway} at 1600 meters.

(b) Sub-graph representing significant regions for the pattern {Jimmy John's, McDonald's, Subway} at 1600 meters.

Figure 9: Significant regions for the pattern {Jimmy John's, McDonald's, Subway} at 1600 meters.

$pi_{obs,d=1600}(Hennepin, \{Jimmy John's, McD, Subway\})$ = 0.4166, $pi_{obs,d=1600}(Washington, \{Jimmy John's, McD, Subway\})$ = 0.3125 with 2, 12, 49 *and* 5 *instances* of the pattern in the counties respectively.

4.1.2 Preliminary comparison with related work. : This section discusses the baseline approach for regional colocation pattern detection with Minimum Orthogonal Bounding Rectangles (MOBRs). This paper employs a space partitioning based approach to find geographic entities within which a colocation pattern is statistically significant. Our baseline methods were **Quad & QGFR** algorithms by Li et al. [10] whose data-aware space partitioning approach is based on Minimum Orthogonal Bounding Rectangle (*MOBR*). Quad uses a threshold (θ) on the participation index (pi) for a candidate pattern C and enumerated MOBRs within which the pattern satisfied the pi threshold and also had atleast γ instances of the pattern C. QGFR uses a new measure *MaxPI* bound which serves as the smallest upper-bound on the participation ratio (pr) of any feature in the colocation pattern C. These algorithms lack statistical significance and only work for data points on a projected plane. We compare our results with this work. Since our dataset is not planar we constructed Minimum Bounding Rectangles (MBRs) while keeping the algorithms unchanged. The MBR based approach with pi-threshold (θ) = 0.6 and minimum number of colocation pattern instances (γ) produced 3368 MBRs or potential localities for the pattern $\{r_g, [Caribou\ Coffee, Starbucks]\}$ in our study area. We performed a statistical significance test similar to our approach for a confidence level of 95%. This resulted in **451 non-significant** pattern regions and **2917 significant** pattern regions within which the pattern was significant. Hence, a regional colocation miner without a statistical significant test results in enumerating output regions where the colocation might have occurred by chance.

5 DOMAIN INTERPRETATION BASED ON GAME THEORY

In this section, we describe a basic two-player Hotelling model to explain why Starbucks and Caribou Coffee co-locate from a game theory perspective. We also provide the conclusions of a

three-player Hotelling model to show why Starbucks, Dunn Brothers Coffee, and Caribou Coffee co-locate. Since we used county boundaries rather than MOBRs in our analysis, we further used macroeconomic data at a county level to explain why coffee shops co-locate in some counties but not in others from a macroeconomic angle, which can be found in Appendix A.

Harold Hotelling [7] proposed a location competition model to explain why retail stores of the same kind tend to co-locate. The basic model had the following assumptions: (1) The city is a linear line from 0 to 1. (2) There are only two firms selling homogeneous products, for example Starbucks and Caribou Coffee. (3) The customers are uniformly distributed between 0 and 1. (4) The unit cost of transportation for each customer is the same, thus the total cost is linear in distance. (5) The two firms compete for location, not for prices. (6) Each consumer needs only one unit of goods. The Hotelling model shows that under these assumptions, the two sellers will simultaneously choose to co-locate in the middle point, which is known as a pure strategy Nash equilibrium in game theory.

A Nash equilibrium in non-cooperative game theory means that the chosen strategy of any player is the best response to the strategies of all other players. None of the players has an incentive to deviate from this set of strategies. In our study, Starbucks chose the middle point as the best response to Caribou Coffee if Caribou Coffee chose the middle point and vice versa. The choice of any other point will cause a decrease of profit.

Suppose the initial locations of a Starbucks (s) store and a Caribou Coffee store (c) are as shown in Figure 10(a). The middle point m is between s and c. Since Starbucks and Caribou Coffee provide homogeneous products at similar prices, consumers are indifferent to the choice of the products, but are sensitive to transportation costs. The consumer at point m is indifferent between the two stores since the distance to either store is the same. However, any consumer located to the left of m will go to Starbucks, while any consumer to the right of m will go to Caribou Coffee because the travel cost is lower. The stores compete for the best location to attract as many customers as possible. The profits of stores in any initial scenarios can be simply represented by the number of customers they attract as shown in the following functions:

If $s < c$ (Figure 10(a)):

$Profit_{Starbucks} = \frac{s+c}{2}$ $\quad Profit_{CaribouCoffee} = 1 - \frac{s+c}{2}$

If $s = c$:

$Profit_{Starbucks} = Profit_{CaribouCoffee} = 0.5$

If $s > c$:

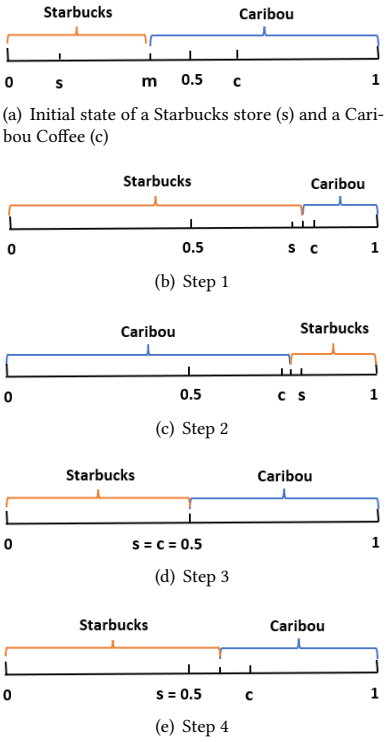
$Profit_{Starbucks} = 1 - \frac{s+c}{2}$ $\quad Profit_{CaribouCoffee} = \frac{s+c}{2}$

According to these profit functions, the competition process with the initial state in Figure 10(a) is shown below.

Step 1: Starbucks will move to the right and stay as close as possible to Caribou's left-hand side. This will allow Starbucks to receive the biggest possible market share.

Step 2: Caribou will move to the left-hand side of Starbucks and stay as close as possible to receive the biggest possible market share.

Step 3: The competition process in Step 1 and Step 2 will keep going and the two coffee shops will move to the left together until they reach the middle point. This is the pure strategy Nash equilibrium.



(a) Initial state of a Starbucks store (s) and a Caribou Coffee (c)

(b) Step 1

(c) Step 2

(d) Step 3

(e) Step 4

Figure 10: Competition process of Starbucks and Caribou

Step 4: Any shop that deviates from the middle point will receive a smaller market share.

The result with this competition process is that the Starbucks store and the Caribou Coffee store co-locate at the middle point, and no one wants to deviate. A similar competition process can be applied for any other initial states, where the same colocation patterns will appear at the end.

For the Hotelling model with three brands, there is no pure strategy Nash equilibrium found. However, the three brands will stay near the middle point (not exactly in the middle) without settled locations. This situation is not stable, as every brand wants to stay near the others while still changing positions to undercut the others. However, shops cannot move freely as the sunk cost is high in the real world. Our empirical results for Starbucks, Caribou Coffee, and Dunn Brothers Coffee have shown that they do co-locate. Thus, the location competition result of three players is unstable in theory but is stable in practice. This finding can help expand the application of game theory as some non-Nash equilibria have the same explaining power as Nash equilibria.

6 RELATED WORK

The concept of colocation was introduced by Shekhar et al. in [12]. Huang et al. [8] provided extensive experiments and rigorous discussions regarding the topic and the participation index as a prevalence measure between constituent features. Barua et al. [1] introduced statistical significance testing in global colocation and segregation pattern detection to avoid enumeration of chance patterns in the

dataset for both aggregation and segregation patterns but did not mention patterns which are regional (or local). Regional colocation was studied by Li et al. [10]. Our work extends their paper by the incorporation of statistical significance and enumeration of more explainable regions by space partitioning.

Prior works on Colocation pattern detection have primarily focused on generalized pattern detection where the search space would encompass the complete dataset [8], [12]. Other works on local pattern detection have focused on shapes [13], zonal patterns [3], and regional colocation patterns for sets of continuous variables [6]. This paper uses a pruning-based approach to reduce the search space of candidate colocations and a statistical test to verify the confidence level of our detection result. Thus, our results are statistically more robust than the previous literature.

A major focus of colocation pattern mining has been using spatial data to find aggregation patterns. The authors in [15] discuss the dynamics of events which mutually evolve with time. In [11] and [4], the focus is drawn more towards the temporal change in data points generated from complete spatial random data. Our work is primarily targeting emerging and vanishing colocation patterns where the spatial pattern might or might not exist at various instances of time. This type of analysis is made possible by newer datasets such as Safegraph and ours is the first work to explore such an approach with geographically meaningful regions as the output.

We also aim to make our detection algorithm robust by adding statistical significance to our method. The existing colocation detection literature [8], [6], [14] uses a threshold on prevalence measures such as the participation index (pi) to categorize a pattern as an aggregation or segregation. Recently an adaptive density threshold has been proposed [5] to find clusters of colocation instances. We intend to compare the pi of a candidate pattern observed in the dataset against the pis of the patterns in complete spatially random datasets, which serve as the null hypothesis. This would give us the probability of whether the observed pattern is rare in a complete spatial random dataset.

7 CONCLUSION AND FUTURE WORK

In this paper we discussed the problem of statistically significant regional colocation pattern detection. We proposed a maximal subgraph based approach which can enumerate candidate regions that are geographically and economically explainable. Experimental results on a real world data show that the proposed approach enumerates regional colocation patterns which are more statistically significant than the current space partitioning based methods. We also provided a case study using the maximal sub-graph based approach and provided a domain interpretation via game theory and regional economics.

Future Work: In future, we plan to explore the task while addressing the multiple comparisons problem. In this paper we perform multiple hypotheses tests simultaneously without addressing the effect of increased false positives with more inferences. We also plan to expand towards country-level retail establishment location data so as to apply our approach to enumerate states and other geographic regions within which a colocation pattern might be significant. Regional colocation pattern detection is a very computationally intensive problem. Adding statistical significance to the mining algorithm greatly increases the computational cost. Thus, we want to improve upon our current work to make it more computationally efficient. Finally, we plan to include the temporal component of emerging and vanishing regional colocation patterns with statistical significance.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants No. 2118285, 2040459, 1737633, 1901099, and 1916518, the USDOD under Grants No. HM0476-20-1-0009, the USDOE Advanced Research Projects Agency-Energy (ARPA-E) under Award No. DE-AR0000795, the NIH under Grant No. UL1 TR002494, KL2TR002492, and TL1 TR002493, the USDA under Grant No. 2021-51181-35861, and Minnesota Supercomputing Institute (MSI). The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. We also thank Kim Koffolt and the Spatial Computing Research Group for valuable comments and refinements.

REFERENCES

[1] Sajib Barua and Jörg Sander. 2013. Mining statistically significant co-location and segregation patterns. *IEEE Transactions on Knowledge and Data Engineering* 26, 5 (2013), 1185–1199.

[2] Julian Besag and Peter J Diggle. 1977. Simple Monte Carlo tests for spatial pattern. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 26, 3 (1977), 327–333.

[3] Mete Celik, James M. Kang, and Shashi Shekhar. 2007. Zonal Co-location Pattern Discovery with Dynamic Parameters. *Seventh IEEE International Conference on Data Mining (ICDM 2007)* (2007), 433–438.

[4] Mete Celik, Shashi Shekhar, James P Rogers, and James A Shine. 2008. Mixed-drove spatiotemporal co-occurrence pattern mining. *IEEE Transactions on Knowledge and Data Engineering* 20, 10 (2008), 1322–1335.

[5] Min Deng, Jiannan Cai, Qiliang Liu, Zhanjun He, and Jianbo Tang. 2017. Multi-level method for discovery of regional co-location patterns. *International Journal of Geographical Information Science* 31 (2017), 1846 – 1870.

[6] Christoph F. Eick, Rachana Parmar, Wei Ding, Tomasz F. Stepinski, and Jean-Philippe Nicot. 2008. Finding regional co-location patterns for sets of continuous variables in spatial datasets. In *GIS '08*.

[7] Harold Hotelling. 1929. Stability in Competition. *The Economic Journal* 39, 153 (1929), 41–57.

[8] Yan Huang, Shashi Shekhar, and Hui Xiong. 2004. Discovering colocation patterns from spatial data sets: a general approach. *IEEE Transactions on Knowledge and data engineering* 16, 12 (2004), 1472–1485.

[9] Janine Illian, Antti Penttinen, Helga Stoyan, and Dietrich Stoyan. 2008. *Statistical analysis and modelling of spatial point patterns*. Vol. 70. John Wiley & Sons.

[10] Yan Li and Shashi Shekhar. 2018. Local co-location pattern detection: a summary of results. In *10th International Conference on Geographic Information Science (GIScience 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[11] Brian D Ripley. 1976. The second-order analysis of stationary point processes. *Journal of applied probability* 13, 2 (1976), 255–266.

[12] Shashi Shekhar and Yan Huang. 2001. Discovering spatial co-location patterns: A summary of results. In *International symposium on spatial and temporal databases*. Springer, 236–256.

[13] Song Wang, Yan Huang, and Xiaoyang Sean Wang. 2013. Regional Co-locations of Arbitrary Shapes. In *SSTD*.

[14] Jin Soung Yoo and Shashi Shekhar. 2006. A joinless approach for mining spatial colocation patterns. *IEEE Transactions on Knowledge and Data Engineering* 18, 10 (2006), 1323–1337.

[15] Jin Soung Yoo, Shashi Shekhar, Sangho Kim, and Mete Celik. 2006. Discovery of co-evolving spatial event sets. In *Proceedings of the 2006 SIAM International Conference on Data Mining*. SIAM, 306–315.