Towards Statistically Significant Taxonomy Aware Co-Location Pattern Detection

Subhankar Ghosh
Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA

Arun Sharma
Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA

Jayant Gupta
Oracle Inc., Nashua, NH, USA

Shashi Shekhar
Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA

Abstract

Given a collection of Boolean spatial feature types, their instances, a neighborhood relation (e.g., proximity), and a hierarchical taxonomy of the feature types, the goal is to find the subsets of feature types or their parents whose spatial interaction is statistically significant. This problem is for taxonomy-reliant applications such as ecology (e.g., finding new symbiotic relationships across the food chain), spatial pathology (e.g., immunotherapy for cancer), retail, etc. The problem is computationally challenging due to the exponential number of candidate co-location patterns generated by the taxonomy. Most approaches for co-location pattern detection overlook the hierarchical relationships among spatial features, and the statistical significance of the detected patterns is not always considered, leading to potential false discoveries. This paper introduces two methods for incorporating taxonomies and assessing the statistical significance of co-location patterns. The baseline approach iteratively checks the significance of co-locations between leaf nodes or their ancestors in the taxonomy. Using the Benjamini-Hochberg procedure, an advanced approach is proposed to control the false discovery rate. This approach effectively reduces the risk of false discoveries while maintaining the power to detect true co-location patterns. Experimental evaluation and case study results show the effectiveness of the approach.

2012 ACM Subject Classification Information systems → Data mining; Computing methodologies → Spatial and physical reasoning

Keywords and phrases Co-location patterns, spatial data mining, taxonomy, hierarchy, statistical significance, false discovery rate, family-wise error rate

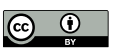
Digital Object Identifier 10.4230/LIPIcs.COSIT.2024.25

Category Short Paper

1 Introduction

Given a collection of Boolean spatial feature types (which satisfy an $is-a$ property), their instances, a neighborhood relation (e.g., proximity), and a hierarchical taxonomy of the feature types, the goal is to find the subsets of feature types or their parents whose spatial interaction is statistically significant. This problem is important due to its use in taxonomy-reliant applications such as ecology (e.g., finding new symbiotic relationships across the food chain), spatial pathology (e.g., immunotherapy for cancer), retail, etc. Figure 1 gives an example of a taxonomy used in retail.

Existing approaches [12, 13, 21, 15, 17] to co-location pattern detection primarily focus on identifying co-located features based on their spatial proximity and co-occurrence frequency. However, these approaches focus on a single spatial scale. In many real-world scenarios, spatial features are organized in taxonomies, where features at different levels of granularity



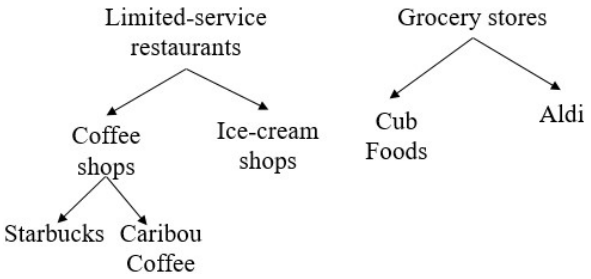


Figure 1 An example taxonomy from retail.

exhibit parent-child relationships. For example, in ecological studies, species are classified into taxonomic ranks such as genus, family, and order [22]. Similarly, in urban planning, points of interest (POIs) are categorized hierarchically (e.g., food, restaurants, and fast food [26]). Ignoring the hierarchical relationships among spatial features can lead to an incomplete or biased analysis of co-location patterns. If only the leaf-level features are considered, co-locations at higher taxonomy levels may be missed. Gupta and Sharma [14] developed the first approach for mining co-location patterns with hierarchical features. However, due to the spatial nature of the data and the multiple comparisons involved in the mining process, the method is likely to discover false positive patterns that arise by chance [3]. Traditional approaches [15, 1, 2] on predefined thresholds such as identifying abnormal gaps [19, 20] or user-specified parameters to determine the significance of co-location patterns [23]. Their lack of a rigorous statistical foundation however, may lead to an inflated rate of false discoveries. Moreover, the significance of co-location patterns may vary across taxonomy levels, requiring a comprehensive analysis that considers the hierarchical structure. For example, in Figure 1, if Starbucks co-locates with Aldi in Minneapolis, that does not imply that all coffee shops (generally) co-located with Aldi in Minneapolis. We propose a novel taxonomic-aware framework for detecting statistically significant co-location patterns. Our framework incorporates the hierarchical relationships among spatial features and employs statistical techniques to control the rate of false discoveries. We introduce two approaches within this framework: (1) a baseline approach that iteratively tests the significance of co-location patterns at different levels of the taxonomy, and (2) an approach based on the **Benjamini-Hochberg** procedure [5] for controlling the false discovery rate (FDR).

Contributions.

- We formalize the problem of statistically significant taxonomy-aware co-location pattern detection and highlight its importance in various application domains.
- We propose a Statistically Significant Taxonomy-aware co-location Miner (SSTCM) approach that incorporates taxonomies and assesses the statistical significance of co-location patterns at different levels of the hierarchy.
- We propose a refined FDR-based Taxonomy-aware co-location Miner (FDR-SSTCM) for controlling false discoveries using the Benjamini-Hochberg procedure.
- We evaluate the proposed approaches using synthetic and real-world datasets, demonstrating their effectiveness of the proposed approach.

2 Basic Concepts and Problem Definition

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 **neighbor relation** $\mathcal{R}$. For example, geographic proximity is represented as $\mathcal{R}_{f_i,f_j} \leq \theta$, where θ is the neighbor relation threshold. In a **neighbor graph**, we represent features that satisfy such relations as a *node* and their relationship as an *edge*.

A **co-location candidate** C is a set of features defined in a given study area (SA) or a sub-region (r_g) where $r_g \in SA$. An instance of a **co-location** satisfies the neighborhood relation $\mathcal{R}$ and forms a **clique**. **Co-location patterns** [21] are the set of prevalent co-location candidates (based on a prevalence measure, e.g., pi), i.e., candidates comprised of features having a high positive spatial interaction.

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 (SA). A **participation index (*pi*)** is the minimal participation ratio of all feature types in a co-location candidate. A **taxonomy (*T*)** of feature-types represents an "is_a" relation between two boolean feature types (e.g., Starbucks is_a Coffee Shop).

A **statistically significant co-location** determines whether an observed positive spatial interaction between features is genuine or could have occurred under complete spatial randomness (CSR). A **statistical significance test** for a co-location pattern assesses the probability of observed results if the features were spatially independent (null hypothesis).

The **multiple comparisons problem** [18] arises when multiple inferences are made simultaneously, increasing the likelihood of incorrectly rejecting the null hypothesis. Addressing this issue often requires stricter significance thresholds for individual comparisons. The **Benjamini-Hochberg procedure** [5] is used for controlling our false discovery rate (FDR), which is the expected proportion of false positives among all significant hypotheses.
(See the appendix for more definitions.)

Formal Problem Definition. Given a set F of spatial features, a significance level α, FDR control level q, and a neighbor relationship $\mathcal{R}$ in a taxonomy tree T, we aim to find statistically significant taxonomy-aware co-location patterns, C. Our objective is to reduce false discoveries, focusing on patterns of smaller size and higher statistical confidence.

Reasoning behind problem output. Testing for statistical significance in taxonomy-aware co-location analysis reduces the identification of spurious patterns, which may arise from class imbalance or spatial auto-correlation. High-cardinality features (e.g., f_B in 2) pose multiple comparisons problems, increasing the false discovery rate. Setting a specific α level is not enough; controlling the overall false discovery rate is necessary.

3 Methodology

3.1 Statistically Significant Taxonomy-aware Co-location Miner

In our study, we utilize a tree-like structure (taxonomy) to represent the hierarchical relationships among spatial features, where each node corresponds to a feature type and edges denote parent-child relationships. The most specific feature types are represented by leaf nodes, while internal nodes signify broader categories. An example of this taxonomy in a retail dataset is shown in Figure 1. To detect co-location patterns, our baseline approach initially identifies patterns at the leaf level, calculating co-location strength using metrics such as the participation index [15]. A pattern is considered significant if its strength exceeds a predetermined threshold, which may be set using domain knowledge or statistical methods like randomization tests [6]. For patterns involving leaf-level features, such as f_C and f_H, we use Algorithm 2 (in Appendix) to compare observed and null hypothesis datasets to compute a p-value, determining statistical significance against a threshold α. Additionally, for

patterns involving non-leaf features with children, such as f_B and f_H, a post-order traversal is performed to check the significance of interactions between the children nodes (f_D, f_E, and f_F) and the leaf node f_H. If they are significant, this implies the co-location between the parent node f_B and the leaf node f_H is also significant. Our approach systematically explores co-location patterns at varying levels of granularity, integrating the taxonomy's hierarchical structure into the analysis process.

Limitations and Challenges. While our baseline method effectively incorporates taxonomies into co-location pattern detection, it does present certain limitations and challenges. One significant issue is the multiple comparisons problem, where the iterative testing across various taxonomy levels increases the risk of Type I errors (false discoveries) due to the heightened number of tests, which boost the likelihood of significant results occurring by chance. Furthermore, the baseline algorithm requires each child node feature of an intermediate taxonomy tree to demonstrate a statistically significant co-location with other features in the candidate pattern, a stringent criterion that tends to elevate the rate of Type II errors (false negatives). To overcome these challenges, we propose an advanced approach that includes specific techniques aimed at controlling the false discovery rates inherent in the taxonomy-aware co-location pattern detection process. More details shown in Algorithm 3 (in Appendix).

3.2 FDR-based Taxonomy-aware co-location Miner

While the baseline $SSTCM$ algorithm can reduce spurious pattern detection, it may be overly conservative in some situations, reducing its statistical power. To strike a balance between controlling false discoveries and maintaining the ability to detect true co-location patterns, we propose a second approach that incorporates the Benjamini-Hochberg procedure [5] for controlling the false discovery rate (FDR). The Appendix A provides more details on the procedure.

False Discovery Rate and Benjamini-Hochberg Procedure. The false discovery rate (FDR) is the expected proportion of false positives among all significant hypotheses. Controlling the FDR is a less stringent criterion than the baseline, as it allows for a certain proportion of false positives while focusing on the overall rate of false discoveries.

The steps for the Benjamini-Hochberg procedure are as follows:

1. Order the p-values of the m hypothesis tests from smallest to largest: p(1), p(2), ..., p(m).
2. Set the desired FDR level q (e.g., 0.05).
3. Find the largest integer k such that $p(k) \leq (k/m) * q$.
4. Reject the null hypotheses for all tests with p-values smaller than or equal to p(k).
   By controlling the FDR, the Benjamini-Hochberg procedure ensures that, on average, no more than a specified proportion q of the rejected hypotheses are false positives.

Incorporating the Benjamini-Hochberg Procedure in Taxonomy-Aware Co-location Pattern Detection. In $FDR\text{–}SSTCM$, the Benjamini-Hochberg procedure is used at each level of the taxonomy when some of the constituent features of the candidate pattern are intermediate nodes in the taxonomy tree. As shown in Figure 2, when checking for the significance of the pattern (f_B, f_H), the $FDR\text{–}SSTCM$ algorithm checks for the significance of the children of f_B with f_H. Applying the Benjamini-Hochberg procedure in this step ensures that at least a few children of f_B have a significant co-location with f_H before we conclude that (f_B, f_H) is a statistically significant co-location pattern.

Algorithm 1 FDR-based Taxonomy-aware co-location Miner (FDR–$SSTCM$) snippet.

Input:
- A spatial dataset S consisting of features $\{f_A, f_B, ...\}$
- Distance threshold d (in meters)
- Taxonomy tree T
- FDR control level q
- Statistical significance level α
- Candidate pattern set C

Output:
A statistically significant subset of co-location patterns $C_s \subset C$ and their p-values ($p\text{-}value_{C_s}$)

1: **procedure** FDR–SSTCM
2: ⋮
22: **procedure** TRAVERSAL(f_p, f_l)
23: ⋮
29: **procedure** POSTORDER(f_p, f_l)
30: **for each:** $f_c \in children(f_p)$ **do**
31: $result, p-value = Traversal(f_c, f_l)$
32: $p-value_{list} \leftarrow$ append $p-value$
33: Sort $p-value_{list}$ in ascending order $p_{(1)}, p_{(2)}, ..., p_{(m)}$
34: $L = max\{j : p_{(j)} < qj/m\}$ ▷ Benjamini-Hochberg procedure
35: **if** $L \leq 2$ **then**
36: return $False$
37: $f_p^{'} \leftarrow$ Update f_p with it's children instances
38: return $True$

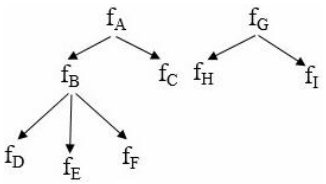


Figure 2 An example taxonomy.

The proposed approach using the Benjamini-Hochberg procedure has several advantages:

1. **FDR Control:** The approach controls the FDR and provides a means to balance between minimizing false discoveries and maintaining statistical power, allowing a manageable proportion of false positives to focus on the overall false discovery rate.
2. **Increased Power:** Compared to the baseline or family-wise error rate (FWER) control methods like the Holm-Bonferroni method, the Benjamini-Hochberg procedure generally has higher statistical power, meaning it is more likely to detect true co-location patterns.
3. **Adaptability:** The approach can be applied to a variety of co-location strength measures and significance tests, making it suitable for different datasets and problem settings.

However, there are also some limitations to consider:

1. **Assumption of Independence:** The Benjamini-Hochberg procedure can inflate FDR levels if its assumption of test independence or positive dependence is violated.
2. **Choice of FDR Level:** Specifying the FDR level q is crucial as it influences results, requiring domain knowledge or alignment with application needs. While the Benjamini-Hochberg procedure effectively manages FDR, balancing false and true positives, the results generated demand careful interpretation and possible further validation. Despite the potential for false positives, it remains a valuable alternative to baseline methods in detecting significant taxonomy-aware co-location patterns.

4 Experimental Evaluation

Our experimental goal was to compare solution quality between $SSTCM$ and $FDR\text{-}SSTCM$ and Figure 3 shows the experiment design.

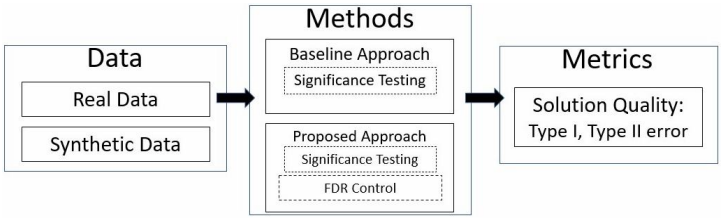


Figure 3 Experiment design.

We performed controlled experiments on synthetic data to compare the solution quality of $SSTCM$ with $FDR\text{-}SSTCM$. The metrics for comparison were $Type\text{-}I$ and $Type\text{-}II$ error rates. $P(\text{Type-I } error) = P(Reject\ H_0 | H_0\ is\ true)$ represents the probability of incorrectly rejecting the null hypothesis H_0, while $P(\text{Type-II } error) = P(Fail\ to\ reject\ H_0 | H_0\ is\ false)$ indicates the likelihood of failing to reject a false H_0. Table 1 displays the experiment results: FDR-SSTCM has a similar Type-I error rate but a significantly lower Type-II error rate compared to SSTCM. The baseline SSTCM algorithm mandates each child intermediate nodes within a taxonomy tree to co-locate with other features to consider the parent node's significance resulting in increased Type-II errors. In addition, unlike FDR-SSTCM, SSTCM does not address the multiple comparisons issue and may exacerbate Type-I errors in the presence of class imbalances.

Table 1 $FDR\text{-}SSTCM$ has lower $Type\text{-}I$ error rate and much lower $Type\text{-}II$ error rate as compared baseline $SSTCM$.

Pattern	***SSTCM Type-I***	***FDR-SSTCM Type-I***	***SSTCM Type-II***	***FDR-SSTCM Type-II***
f_A, f_G	0.05	0.03	0.35	0.21
f_E, f_H	0.02	0.02	0.2	0.13
f_B, f_H	0.04	0.03	0.27	0.16

Case Study. We extended our previous case study [12] to demonstrate the effectiveness of the proposed approach. The dataset, provided by SafeGraph – a vendor of mobility data – offers anonymized and aggregated location data to researchers studying the impact of COVID-19 on movement patterns towards various Points of Interest (POIs) across 1,473 retail brands in Minnesota. The data organization follows the hierarchical North American Industry Classification System (NAICS), which classifies businesses from level 1 to level 5 (e.g., if NAICS level 1 is "Retail Trade", then more detailed information is fleshed out at subsequent levels, from level 2 to level 5). In our case study, we found that the co-location of limited-service restaurants, specifically Subway and McDonald's, and coffee shops like Starbucks, was statistically significant within approximately 1400 meters in Hennepin County. Further global searches confirmed similar significant co-location patterns for Fast Food Restaurants and Coffee Shops, and other category pairings like Starbucks with Olive Garden showed significance, though this pattern did not hold for general categories of coffee shops and full-service restaurants.

5 Related Work

Co-location pattern detection is used to discover subsets of spatial features that frequently co-occur in close proximity [21] and has been extensively studied in the spatial data mining literature. Huang et al. [15] introduced the concept of spatial co-location patterns and proposed a framework for mining such patterns using a spatial join-based approach. Yoo et al. [25] developed a joinless approach for mining co-location patterns, which improves computational efficiency by avoiding the need for spatial joins. These early works focused on discovering co-location patterns at a single spatial scale without considering the hierarchical relationships among spatial features. Gupta and Sharma [14] explored the detection of co-location patterns within hierarchical relationships. However, their approach did not consider statistical significance, leading to spurious patterns, especially in the presence of class imbalance among child nodes in a taxonomy. Barua et al. [4] introduced the concept of statistical significance in global co-location pattern detection, and Ghosh et al. [12] introduced statistical significance in regional co-location pattern detection, but their work did not consider taxonomies or the multiple comparisons problem. Ghosh et al. [13] addressed the multiple testing problem in the context of co-location pattern detection. Using the Bonferroni Correction, they proposed a method for controlling the family-wise error rate (FWER). However, they did not consider the hierarchical relationships among spatial features. Recently, deep learning-based spatial association mining [7, 9, 8] has been proposed, but none of them considered the hierarchical relationship between features. Spatial association has also been studied in various other domains [11, 10, 16, 24] where considering the hierarchical relationship might be valuable.

6 Conclusion and Future Work

In this paper, we propose a statistically significant taxonomy-aware co-location miner (SSTCM) and refine the problem of statistical significance with the proposed FDR-based taxonomy-aware co-location miner, utilizing the Benjamini-Hochberg procedure. We evaluate our approaches using synthetic and real-world datasets, demonstrating their effectiveness. Finally, we provide a comparative analysis of the two approaches, discussing their strengths, limitations, and provide a case study on retail establishments in Minnesota using the proposed approach.

Future Work. We plan to explore other methods to reduce Type-I errors (false positives) further while also addressing Type-II errors (false negatives) and further providing computational efficiency to handle patterns of a large number of features. Finally, we plan to add a temporal dimension to these patterns.

References

1 Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993.

2 Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499. Citeseer, 1994.

3 Jared Aldstadt. Spatial clustering. In *Handbook of applied spatial analysis: Software tools, methods and applications*, pages 279–300. Springer, 2009.

4 Sajib Barua and Jörg Sander. Mining statistically significant co-location and segregation patterns. *IEEE TKDE*, 26(5):1185–1199, 2013.

5 Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

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

7 Majid Farhadloo, Carl Molnar, Gaoxiang Luo, Yan Li, Shashi Shekhar, Rachel L Maus, Svetomir Markovic, Alexey Leontovich, and Raymond Moore. Samcnet: towards a spatially explainable ai approach for classifying mxif oncology data. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2860–2870, 2022.

8 Majid Farhadloo, Arun Sharma, Jayant Gupta, Alexey Leontovich, Svetomir N Markovic, and Shashi Shekhar. Towards spatially-lucid ai classification in non-euclidean space: An application for mxif oncology data. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pages 616–624. SIAM, 2024.

9 Majid Farhadloo, Arun Sharma, Shashi Shekhar, and Svetomir N Markovic. Spatial computing opportunities in biomedical decision support: The atlas-ehr vision. *arXiv preprint arXiv:2305.09675*, 2023.

10 Subhankar Ghosh. Video popularity distribution and propagation in social networks. *Int. J. Emerg. Trends Technol. Comput. Sci.(IJETTCS)*, 2017.

11 Subhankar Ghosh, Shuai An, Arun Sharma, Jayant Gupta, Shashi Shekhar, and Aneesh Subramanian. Reducing uncertainty in sea-level rise prediction: A spatial-variability-aware approach. *arXiv preprint arXiv:2310.15179*, 2023.

12 Subhankar Ghosh et al. Towards geographically robust statistically significant regional colocation pattern detection. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on GeoSpatial Simulation*, pages 11–20, 2022.

13 Subhankar Ghosh et al. Reducing false discoveries in statistically-significant regional-colocation mining: A summary of results. In *12th International Conference on Geographic Information Science (GIScience 2023)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2023.

14 Jayant Gupta and Arun Sharma. Mining taxonomy-aware colocations: a summary of results. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–11, 2022.

15 Yan Huang et al. Discovering colocation patterns from spatial data sets: a general approach. *IEEE TKDE*, 16(12):1472–1485, 2004.

16 Yan Li, Majid Farhadloo, Santhoshi Krishnan, Yiqun Xie, Timothy L Frankel, Shashi Shekhar, and Arvind Rao. Cscd: towards spatially resolving the heterogeneous landscape of mxif oncology data. In *Proceedings of the 10th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, pages 36–46, 2022.

17 Yan Li and Shashi Shekhar. Local co-location pattern detection: a summary of results. In *GIScience*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018.

18 G Rupert Jr et al. *Simultaneous statistical inference*. Springer Science & Business Media, 2012.

19 Arun Sharma, Subhankar Ghosh, and Shashi Shekhar. Physics-based abnormal trajectory gap detection. *ACM Transactions on Intelligent Systems and Technology*, 2024.

20 Arun Sharma, Jayant Gupta, and Subhankar Ghosh. Towards a tighter bound on possible-rendezvous areas: preliminary results. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–11, 2022.

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

22 R. Whittaker. Evolution and measurement of species diversity. *Taxon*, 21(2-3):213–251, 1972.

23 Xiangye Xiao et al. Density based co-location pattern discovery. In *Proceedings of the 16th International conference on Advances in geographic information systems*, pages 1–10, 2008.

24 Mingzhou Yang, Bharat Jayaprakash, Matthew Eagon, Hyeonjung Jung, William F Northrop, and Shashi Shekhar. Data mining challenges and opportunities to achieve net zero carbon

emissions: Focus on electrified vehicles. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 953–956. SIAM, 2023.

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

26 Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 186–194, 2012.

A Appendix

A **null hypothesis** (H_0) posits no spatial interaction between dataset features, suggesting their independence. Conversely, an **alternative hypothesis** (H_a) asserts positive spatial interaction among features in a specified region, challenging H_0.

A **Type-I error** refers to the erroneous rejection of an actually true null hypothesis (or a false positive) while a **Type-II error** refers to the failure to reject a null hypothesis (H_0) that is actually false (or a false negative).

Null hypothesis generation.

- For an identical distribution, we generate the same number of feature instances across the study area using summary statistics, ensuring that the null hypotheses datasets closely model the observed dataset in each partition.
- We sample instances using a Poisson point process to ensure independence, analyzing them with a pair correlation function (PCF) up to a data-driven distance d; $g(d) > 1$ indicates clustering, while $g(d) = 1$ signifies complete spatial randomness (CSR).

Statistical significance test. The participation index (pi) quantifies spatial interaction strength; we compute the probability that the observed data's pi for a pattern C, represented as $pi_{obs}(C)$, differs from its null hypothesis index, $pi_{\emptyset}(C)$.

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

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 in the observed data ($pi_{obs}(C)$) and R refers to the total number of Monte Carlo simulations. If $p \leq \alpha$, we consider $pi_{obs}(C)$ as statistically significant at level α.

The **Benjamini-Hochberg** procedure [5] is a widely used method for controlling the false discovery rate (FDR) in multiple-hypothesis testing. The FDR is the expected proportion of false positives among all significant hypotheses. Unlike family-wise error rate (FWER) control methods, such as the Bonferroni correction, which controls the probability of making at least one false discovery, the Benjamini-Hochberg procedure focuses on the proportion of false discoveries among all rejected hypotheses. It provides a more powerful and less conservative approach to multiple testing by allowing a certain level of false positives while ensuring that the FDR is controlled at a desired level. The procedure works by ranking the p-values from smallest to largest and comparing each p-value to a threshold that depends on its rank and the desired FDR level. The Benjamini-Hochberg procedure has been widely applied in various fields, including genomics, neuroscience, and spatial data mining, where multiple comparisons are common, and controlling the FDR is crucial for drawing meaningful conclusions from the data.

Algorithm 2 Significance testing.

Input:
- A spatial dataset S consisting of features $\{f_A, f_B, ...\}$ in a study area
- Statistical significance level α and candidate co-location pattern C
- A set of R Null hypotheses ($NH_\emptyset$) data each modelled as co-location C

Output:
1. C is significant or not
2. $p\text{-}value_C$

procedure SIGNIFICANCE TESTING
  Statistically significant result $SSR_C \leftarrow$ False
  Counter $R^{\geq pi_{obs}} \leftarrow 0$
  Calculate pi_{obs} for C at d
  for $i \in [1, R]$ **do**
    Calculate the $pi_{\emptyset,i}$ of C at d in the i^{th} $NH_\emptyset$
    if $pi_{\emptyset,i} \geq pi_{obs}$ **then**
      $R^{\geq pi_{obs}} \leftarrow R^{\geq pi_{obs}} + 1$
  $p\text{-}value_C = \frac{R^{\geq pi_{obs}}+1}{R+1}$
  if $p\text{-}value_C \leq \alpha$ **then**
    $SSR_C \leftarrow$ True ▷ (i.e., C is statistically significant)
  else
    $SSR_C \leftarrow$ False ▷ (i.e., C is not statistically significant)
  return SSR_C, $p\text{-}value_C$

Algorithm 3 Statistically Significant Taxonomy-aware co-location Miner ($SSTCM$).

Input:
- A Spatial dataset S consisting of features $\{f_A, f_B, ...\}$
- Taxonomy tree T
- Statistical significance level α
- Candidate pattern set C

Output:
Subset of co-location patterns $C_s \subset C$ which are statistically significant and their p-values ($p\text{-}value_{C_s}$)

procedure Statistically Significant Taxonomy-aware co-location Miner
  for each: f_k in $\{f_A, f_B, ...\}$ **do**
    Generate R null hypotheses ($NH_\emptyset$) using summary statistics
  for each: candidate pattern $C_m \in \{C_1, C_2, ..., C_M\}$ **do**
    $f_1, f_2 \in C_m$
    if f_1, f_2 are leaf nodes in T: **then**
      return Significance Testing(S, α, C_m, $NH_\emptyset$)
    else if f_1 is leaf node and f_2 is a parent node in T: **then**
      return Traversal(f_2, f_1)
    else
      Replace f_2 with it's children instances
      $SSR^1_{C_m}$, $p - value_1 \leftarrow$ Traversal(f_1, f_2)
      Replace f_1 with it's children instances
      $SSR^2_{C_m}$, $p - value_2 \leftarrow$ Traversal(f_2, f_1)
      if $SSR^1_{C_m}$ and $SSR^2_{C_m}$ are both $True$: **then**
        return Significance Testing(S, α, C_m, $NH_\emptyset$)
procedure Traversal(f_p, f_l)
  if f_p is a leaf node in T: **then**
    return Significance Testing(S, α, (f_p, f_l), $NH_\emptyset$)
  post-order-result = PostOrder(f_p, f_l)
  if *post-order-result* is $True$ **then**
    return Significance Testing(S, α, ($f_p^{'}, f_l$), $NH_\emptyset$)
  else
    return $False, 1$
procedure PostOrder(f_p, f_l)
  $result \leftarrow True$
  for each: $f_c \in children(f_p)$ **do**
    $result, p - value = Traversal(f_c, f_l)$
    if $result$ is $False$ **then**
      return $result, 1$
  $f_p^{'} \leftarrow$ Update f_p with it's children instances