Towards Better Detection and Analysis of Massive Spatiotemporal Correlation Patterns

Yingcai Wu, Di Weng, Zikun Deng, Jie Bao, Zhangye Wang, Yu Zheng, and Wei Chen

***Abstract*—With the rapid development of sensing technologies, massive spatiotemporal data have been acquired from the urban space with respect to different domains, such as transportation and environment. Numerous correlation patterns (e.g., traffic speed < 10km/h, weather $=$ *foggy*, and air quality $=$ *unhealthy*) between the transportation data and other types of data can be obtained with given spatiotemporal constraints (e.g., within 3 kilometers and lasting for 2 hours) from these heterogeneous data sources. Such patterns present valuable implications for many urban applications, such as traffic management, pollution diagnosis, and transportation planning. However, extracting and understanding these patterns is beyond manual capability because of the scale, diversity, and heterogeneity of the data. To address this issue, a novel visual analytics system called *CorVizor* is proposed to identify and interpret these correlation patterns. CorVizor comprises two major components. The first component is a correlation mining framework involving three steps, namely, spatiotemporal indexing, co-occurring instance generation, and pattern mining. The second component is a visualization technique called *CorView* that implements a level-of-detail mechanism by integrating tailored visualizations to depict the extracted spatiotemporal correlations. Case studies and expert interviews are conducted to demonstrate the effectiveness of CorVizor.**

***Index Terms*—Heteorgeneous urban data, spatiotemporal data visualization, correlation pattern analysis.**

I. Introduction

THE rapid development of sensing technologies has resulted in a large amount of heterogeneous urban data acquired from different data sources, such as traffic and air quality data. These data inherently comprise numerous interesting *correlation patterns*, i.e., the combinations of the property value ranges that frequently co-occur with each other. These patterns appear frequently within a spatial range and a temporal window and may comprise properties from various data sources. For example, given three data sources, namely, transportation, weather, and air quality, and spatiotemporal constraints (within 3-kilometer range and 2-hour window), a fine-grained correlation pattern like $\{20 < \texttt{TrafficVolume} < 30,\ 100\text{m/s} < \texttt{WindSpeed} < 150\text{m/s},\ \texttt{AirQuality} = \text{healthy}\}$ may be identified. These fine-grained patterns reveal important spatiotemporal insights and anomalies (i.e., counterintuitive correlation patterns) across multiple data sources that support numerous urban decision-making applications, including traffic management and transportation planning.

However, neither have such correlation patterns been systematically studied and detected, nor effectively interpreted and understood. Two challenges arise from the identification and interpretation of these patterns: a) **efficient extraction** and b) **interactive visualization**.

Efficient extraction of the correlation patterns is considerably difficult, particularly from the heterogeneous urban data sources that comprise various properties, such as PM2.5 and PM10 in air quality data and temperature and humidity in meteorological data. Without an efficient approach, exhaustively testing all possible combinations of these properties to find the potential patterns will result in poor computational performance as the number of properties increases. Furthermore, the aforementioned patterns are fine-grained on continuous value domains and may involve both categorical and numerical properties. However, most of the traditional correlation mining techniques, such as Apriori [1], [2], are specifically designed to detect the coarse association rules (i.e., correlation patterns), such as $\{\texttt{butter}, \texttt{bread}\} \Rightarrow \{\texttt{milk}\}$, among categorical attributes only. Some recent techniques [23], [49] that extract patterns from numerical data generally require the continuous domains of property values to be initially discretized, thereby resulting in the severe loss of latent patterns.

The extracted correlation patterns also require a well-designed visualization technique, with which domain experts can examine these correlations, identify spatiotemporal trends, and study correlation anomalies. However, it is difficult to develop an appropriate technique that visualizes these patterns because of three identified design challenges: a) **diversity:** the data properties involved in the visualized correlation patterns may have different types, scales, and semantics; b) **volume:** numerous patterns with overlapping value ranges can be extracted from heterogeneous urban data; c) **organization:** a combination of properties can be shared by many patterns, thereby forming a two-level hierarchical structure where the designed visualizations should enable experts to explore the hierarchy flexibly and efficiently. To the best of our knowledge, none of the existing visualization approaches, including parallel coordinates and parallel sets [24], can be applied directly to address these challenges, which demand a set of considerate visualizations specifically tailored on the basis of the unique characteristics of the extracted patterns.

In this study, we develop a novel data mining model that extracts correlation patterns efficiently based on three main modules: a) *spatiotemporal indexing* that builds a unified index structure to accelerate the following mining process, b) *co-occurring instance generation* that identifies co-occurring instances and builds a pruning graph to reduce the search space of patterns, and c) *pattern mining* that aggregates the

Yingcai Wu, Di Weng, Zikun Deng, Zhangye Wang, and Wei Chen are with the State Key Lab of CAD&CG, Zhejiang University. E-mail: {ycwu, dweng, zikun_rain}@zju.edu.cn, {zywang, chenwei}@cad.zju.edu.cn.

Jie Bao and Yu Zheng are with Urban Computing Business Unit, JD Finance. E-mail: baojie1985@gmail.com, msyuzheng@outlook.com.

Manuscript received April 19, 2005; revised August 26, 2015.

data by using *value matrices* and extracts distinct patterns via an efficient *sweep-line* algorithm. We also propose a new matrix-based visualization technique named *CorView*, which effectively depicts the extracted patterns that comprise properties of different data types, scales, and semantics in an aligned fashion. Particularly, we address the scalability issue by adopting a level-of-detail mechanism that integrates brick-like glyphs, scatterplots, parallel coordinates, and a stacked line chart. Furthermore, we design *CorVizor*, a novel visual analytics system that helps users reliably detect and analyze the correlation patterns in cross-domain urban data. The major contributions of this study are as follows.

- ◇ We characterize the problem of identifying and interpreting fine-grained spatiotemporal correlations among cross-domain urban data sources;
- ◇ We formulate an efficient framework that extracts the correlation patterns from heterogeneous spatiotemporal data based on three main modules;
- ◇ We design *CorView*, a multi-scale visual representation for depicting the massive patterns that comprise properties of different types, scales, and semantics;
- ◇ We develop *CorVizor*, a visual analytics system that assists experts in exploring, interpreting, and analyzing the extracted correlation patterns effectively.

II. Related Work

This section discusses related studies in the following three parts, namely, correlation pattern mining, spatiotemporal visualization, and correlation visualization.

A. Correlation Pattern Mining

Correlation pattern mining (i.e. association analysis) techniques aims to identify interesting correlations among categorical or numerical data properties.

Traditional techniques like Apriori [1], [2] and PrefixSpan [3] attempt to extract the patterns from transactional datasets and thus were limited to handling categorical data. Similar methods have been adapted to transactional urban datasets, where the extraction of correlation patterns from the moving object [10], boolean [34], [55], and event [7], [29], [36] datasets in spatiotemporal contexts were extensively studied. However, these techniques cannot be directly applied to solve our problem since most of the properties like traffic speed and volume in heterogeneous urban data have continuous domains.

Techniques [23], [38], [49] were also proposed to handle numerical data by dividing the continuous domains of properties into a number of bins. However, such discretization may lead to the severe loss of latent patterns. Other studies have attempted to avoid the discretization with topological methods in spatiotemporal contexts. Chirigati et al. [14] developed a topology-based method named data polygamy. This method efficiently extracts the relationships between extrema in urban datasets. Nonetheless, it can only identify the correlations that comprise the peaks or valleys of data properties.

We define correlation patterns as the flexible value range combinations of both numerical and categorical properties in heterogeneous urban datasets. Without the discretization of domains, our model has finer granularity and extracts the patterns more reliably than those models in the prior studies.

B. Spatiotemporal Visualization

The rapid development of smart cities enables authorities to collect citywide spatiotemporal data via sensors more efficiently than ever, making data-driven solutions possible for urbanization problems like air pollution and traffic congestions. To integrate human in the analysis loop, spatiotemporal data visualization has been investigated and applied in many settings, such as billboard location selection [27], public utility analysis [54], and hotspot prediction [30]. Andrienko et al. [4] provided an excellent taxonomy of existing spatiotemporal visualization methods for movement data, which are classified into three categories, namely, direct depiction (e.g. points [19], polylines [5], stacked bands [44], and space-time cubes [6]), summarization (e.g. density maps [40], [50], graphs [45], and flow maps [20]), and pattern extraction [12], [22], [52]. Sun et al. [42] also explored the better integration of temporal information in spatial contexts by transforming maps. To handle large-scale spatiotemporal data, many novel methods have been incorporated into visualizations, such as tailored query model [19], topological methods [16], [33], uncertainty analysis [12], and anomaly detection [9]. However, most of the prior studies focus on single-source data only, including trajectory [39], cellphone [52], [56], and weather data [37]. In contrast, our study targets at visualizing multi-source heterogeneous data, which poses difficult design challenges arisen from the unique characteristics of correlation patterns.

This work establishes a pattern extraction method that aims to detect and visualize an extensive number of fine-grained correlation patterns among heterogeneous datasets. In particular, various types of urban data from multiple domains were analyzed and explored through a mining model and a set of tailored visualization techniques.

C. Correlation Visualization

Visually understanding and analyzing the massive extracted correlation patterns remain a difficult and challenging task. Many correlation visualization methods targeting categorical data have been proposed based on scalable techniques like 2D plots [26], [28], graphs [17], parallel coordinates [53], and matrices [21], [51]. For numerical datasets, Bothorel et al. [8] proposed a visual mining pipeline based on the Apriori algorithm, yet the value ranges must be discretized.

Recently, the visual analysis of spatiotemporal correlations has attracted wide research interests. Qu et al. [37] studied the visualization of the correlations between various weather attributes. TelCoVis [52] was designed to illustrate the human co-occurrence patterns with mobile phone data. Furthermore, a few studies have considered the complex correlation patterns among multiple data sources. Urbane [18] combines datasets from diverse domains for target building selection. VAUD [13] allows users to explore cross-domain correlations based on visual query and reasoning. COPE [25] detects various co-occurrence patterns of spatiotemporal events via a well-designed visual interface. However, most of these visualization techniques neither integrate with an automated mining model nor involve the correlation patterns characterized by combinations of continuous value ranges and categorical sets. Thus, finding and interpreting interesting patterns will become increasingly difficult with the growing size of datasets.

TABLE I: Properties of three data sources.

Data Source	Property	Value Range
Air Quality	PM 2.5, PM 10 (ug/m^3)[1]	[0, 500]
	O$_3$ (ug/m^3)	[0, 2300]
	NO$_2$ (ug/m^3)	[0, 300]
	CO (ug/m^3)	[0, 70]
	SO$_2$ (ug/m^3)	[0, 150]
	AQI Level	6 levels
Weather	Temperature °C	[-20, 40]
	Humidity	[0, 100]
	Wind Speed (m/s)	[0, 300]
	Wind Direction	[1, 24]
	Cloud Conditions	14 conditions
Traffic	Total Cars	[9, 200]
	Low Speed $(0-20$ km/h) %	[6, 80]
	Medium Speed ($20-50$ km/h) %	[17, 74]
	High Speed (above 50 km/h) %	[0, 53]

[1] PM 2.5 is particulate matter 2.5 micrometers or less in diameter; PM 10 is particulate matter 10 micrometers or less in diameter.

In this paper, we design a novel analytics system that combines several interactive visualizations specifically tailored for the massive fine-grained correlation patterns extracted by the proposed model in spatiotemporal contexts.

III. Background and System Overview

This section presents the background, problem, and overview of the proposed system.

A. Background

Our study mainly focuses on extracting, visualizing, and evaluating frequent spatiotemporal patterns obtained from heterogeneous urban data. We introduce the following terminologies in the extraction of spatiotemporal correlation patterns. For each annotation, the superscript is used to distinguish different objects, and the subscript is to indicate the association of the current object.

- ◇ **Data Source**: A data source $s \in S = \{s^1, s^2, ..., s^n\}$ comprises a set of spatial locations $\{l^1, l^2, ...\} \in L$, each of which is associated with a set of time-varying properties $\{p^1, p^2, ...\}$ observed at the location. The data sources used in this study are shown in Table I with their detailed properties.
- ◇ **Instance**: An instance φ, associated with a data source s, comprises a spatial location l, a time point t, and a value v_p of property p observed at time t and location l.
- ◇ **Property Value Range**: We denote a specific value range of property p of data source s with $s_p|\mathcal{C}$. The range $\mathcal{C}$ can be either numeric (e.g. $[5, 8]$) or ordinal (e.g. $\{\text{cloudy}\}$), depending on the type of property p. An instance φ satisfies a property value range $s_p|\mathcal{C}$ iff (1) φ is collected from the property p and (2) the observed value v_p of property p is within the range of $\mathcal{C}$, namely, $v_p \in \mathcal{C}$.
- ◇ **Co-occurrence of Instances**: Instances are co-occurring with each other iff they co-occur within the user-specified spatial and temporal thresholds. Fig. 1(a) shows an example of five instances $\varphi^2_{s^\tau}$, $\varphi^1_{s^\alpha}$, $\varphi^2_{s^\alpha}$, $\varphi^1_{s^\beta}$, and $\varphi^2_{s^\beta}$ of data sources s^τ, s^α, and s^β that co-occur with instance $\varphi^1_{s^\tau}$ under the spatial and temporal threshold d and t.

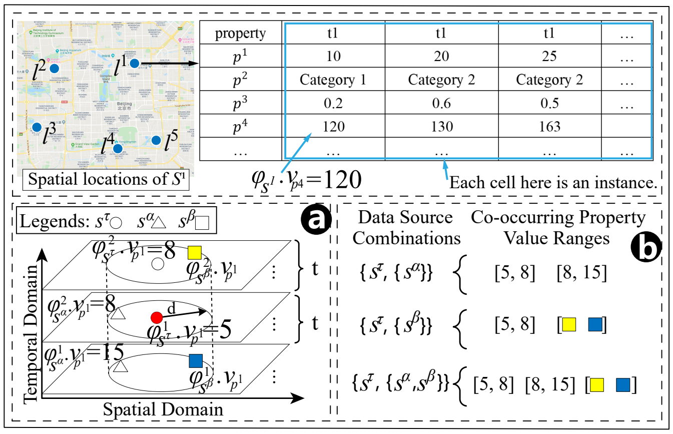

property
t1
t1
t1
…
p^1
10
20
25
…
p^2
Category 1
Category 2
Category 2
…
p^3
0.2
0.6
0.5
…
p^4
120
130
163
…
…
…
…
…
l^1
l^2
l^3
l^4
l^5
Spatial locations of S^1
$\varphi_{S^1} \cdot v_{p4}$=120
Each cell here is an instance.
Legends: s^τ○ s^α△ s^β□
a
$\varphi^2_{s^\tau} \cdot v_{p^1}$=8
$\varphi^2_{s^\beta} \cdot v_{p^1}$
$\varphi^2_{s^\alpha} \cdot v_{p^1}$=8
d
$\varphi^1_{s^\tau} \cdot v_{p^1}$=5
$\varphi^1_{s^\alpha} \cdot v_{p^1}$=15
$\varphi^1_{s^\beta} \cdot v_{p^1}$
t
t
Temporal Domain
Spatial Domain
b
Data Source Combinations
Co-occurring Property Value Ranges
$\{s^\tau, \{s^\alpha\}\}$ [5, 8] [8, 15]
$\{s^\tau, \{s^\beta\}\}$ [5, 8] [🟨 🟦]
$\{s^\tau, \{s^\alpha, s^\beta\}\}$ [5, 8] [8, 15] [🟨 🟦]

Fig. 1: (a) Co-occurring instances and (b) examples of co-occurring property value ranges.

- ◇ **Co-occurrence of Property Value Range**: By extracting co-occuring instances, the co-occurrences of the property value ranges constituted by the values of these instances can be obtained. Fig. 1(b) shows an example of the co-occurring value ranges of the properties in data sources s^α and s^β w.r.t. data source s^τ.
- ◇ **Correlation of Property Value Ranges**: Value ranges of different properties are correlated iff there exists a sufficient number of co-occurring instances, each of which satisfies its corresponding property value range. The sufficiency is determined by the *correlation coverage* introduced below.
- ◇ **Correlation Coverage**: Regarding s^τ as the *target data source* of a given correlation of property value ranges which involves one and only one property in s^τ, we compute the correlation coverage λ as the number of correlated instance combinations involving s^τ divided by the number of instances of s^τ (denoted by $|s^\tau|$).
- ◇ **Frequent Spatiotemporal Pattern**: With respect to target data source s^τ, a frequent spatiotemporal pattern is a set of correlated property value ranges, and the correlation coverage λ based on s^τ is larger than the threshold specified by the user. A pattern is denoted with $\{s^\tau_{p^u}|\mathcal{C}, \{s^\alpha_{p^x}|\mathcal{C}, s^\beta_{p^y}|\mathcal{C}, ...\}\}$, where s^τ, s^α, and s^β are correlated data sources, p^u, p^x, and p^y are data properties, and $s^\tau_{p^u}|\mathcal{C}$, $s^\alpha_{p^x}|\mathcal{C}$, and $s^\beta_{p^y}|\mathcal{C}$ are correlated property value ranges. We developed optimization techniques to remove redundant patterns by maximizing the correlation coverage.

B. Problem Definition

Given target data source s^τ, a group of cross-domain data sources $S = \{s^1, s^2, ..., s^n\}$, and a set of mining parameters, including spatial distance d, temporal window t, and a user specified threshold λ_e (i.e., a desired correlation coverage value), a set of frequent spatiotemporal correlation patterns are identified with the likelihood of occurrence (i.e., correlation coverage λ) being larger than λ_e from S, that is, $\lambda > \lambda_e$. We denote the identified patterns as $\{s^\tau_{p^u}|\mathcal{C}, \{s^\alpha_{p^x}|\mathcal{C}, s^\beta_{p^y}|\mathcal{C}, ...\}\}$. The objective is to identify all distinctive spatiotemporal

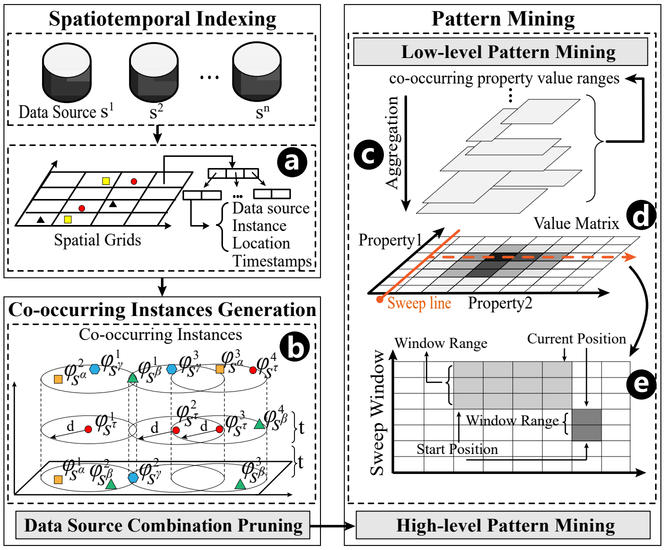


Fig. 2: Overview of the mining framework.

patterns by efficient correlation mining and comprehend the patterns through interactive visualization.

C. System Overview

We develop CorVizor, a web-based visual analytics system that can assist urban experts in interpreting and analyzing patterns extracted from heterogeneous urban data. CorVizor comprises two components: data mining and interactive visualization. The mining component, which is implemented in C#, indexes heterogeneous spatiotemporal data, extracts co-related instances, and performs pattern mining, thereby transforming raw data into interpretable patterns. The visualization component, which is implemented in TypeScript, visualizes patterns with four tailored views, enabling users to interactively filter, compare, and evaluate patterns across multiple data sources.

IV. Mining Framework

Detecting fine-grained correlation patterns from heterogeneous urban data sources is difficult because spatiotemporal urban data generally do not have transactions. Moreover, diverse data sources possess dissimilar value scales, including the scales of numeric and categorical values. Furthermore, the correlation patterns characterized by combinations of continuous value ranges and categorical sets can easily lead to the exponential size of possible correlation patterns. Given these issues, traditional mining techniques, such as Apriori [2], cannot be applied directly. Thus, we propose a mining framework with these three modules (Fig. 2) to tackle the challenges: (1) **indexing spatiotemporal data** with a nested data structure to ensure the efficiency of the subsequent mining process, (2) **generating co-occurring instances** by identifying these instances with unified indexes and pruning the impossibly co-occurring instances, and (3) **mining frequent patterns** with a novel sweep-line algorithm based on the *value matrices* constructed from instances.

A. Indexing Spatiotemporal Data

At this stage, we build a unified index [11], [15], [47] for large-scale heterogeneous spatiotemporal data (Fig. 2(a)) to enable faster data retrieval with different mining parameters specified and accelerate the subsequent mining stages. To construct the index, we divide the map into $n \times m$ spatial grids, each of which has an area of 1×1 km^2. Each grid maintains the covered instances with a temporal index, where the instances are organized by their timestamps with a B^+ tree. Each leaf node of the B^+ tree records the ID, data source ID, GPS location, and timestamp of an instance.

B. Generating Co-occurring Instances

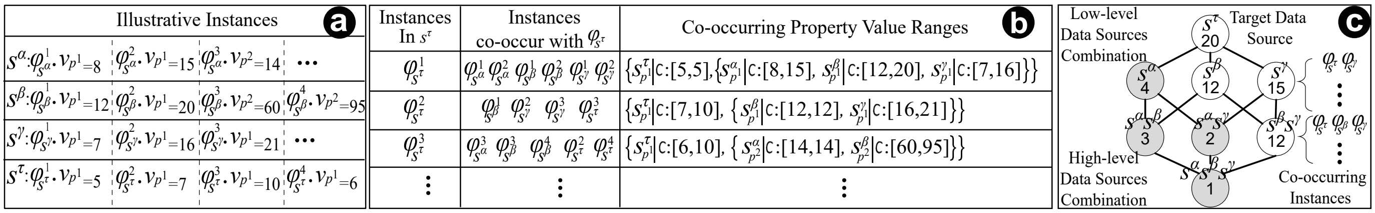
Using the target data source s^τ, spatial threshold d, temporal threshold t, and expected correlation coverage λ_e, the proposed mining framework attempts to extract co-occurring instances with *co-occurrence tables* and *pruning graphs* (Fig. 3) from the spatiotemporal index built at the previous stage.

First, range queries based on the spatial and temporal thresholds are issued for each instance of target data source s^τ to identify co-occurring instances within the same property of the target data source or in other data sources. Based on the given spatial distance and temporal range, the co-occurring instances (Fig. 3(a)) in a spatiotemporal cylinder of each instance in s^t (Fig. 2(b)) are organized into a co-occurrence table (Fig. 3(b)) by the instances in s^t. Additionally, the values of the co-occurring instances associated with the same property are aggregated and represented with a value range.

Next, a pruning graph (Fig. 3(c)) is created to characterize all combinations of the data sources associated with the detected co-occurring instances, such that impossible combinations are pruned at the data source level. The basic idea is that no frequent spatiotemporal patterns of two data sources is present if no sufficient co-occurring instances are found from the two data sources. Each node represents a potential combination of data sources with (1) the IDs of the involved data sources, (2) a list of the co-occurring instances of the involved data sources, and (3) a counter storing the number of the co-occurring instances. For example, the node labeled $\{S^\alpha S^\beta, 3\}$ in Fig. 3(c) indicates that three co-occurring instances can be extracted from the data sources S^α, S^β, and S^τ. Links between the nodes depict downward closure relations [1], i.e., an upper-level node contains all the instances of its linked lower-level nodes. Therefore, the insignificant combinations of data sources can be quickly detected and invalidated from the top to the bottom at this stage, by removing the nodes whose number of associated co-occurring instances is lower than the specified threshold ($\lambda_e \cdot |T|$, as per the definition of correlation coverage), as illustrated with gray nodes in Fig. 3(c). Corresponding rows in the co-occurrence table are removed thereafter. Hence, the property combinations belonging to invalid data source combinations are eliminated, thereby accelerating the subsequent pattern mining stages.

C. Mining Frequent Patterns

We propose a two-fold approach to extract frequent patterns from the co-occurrence table and pruning graph. This approach comprises two steps: (1) **Low-Level Pattern Mining** entails



Illustrative Instances (a)			
S^{α}:$\varphi^{1}_{S^{\alpha}}\cdot v_{p^1}$=8	$\varphi^{2}_{S^{\alpha}}\cdot v_{p^1}$=15	$\varphi^{3}_{S^{\alpha}}\cdot v_{p^2}$=14	···
S^{β}:$\varphi^{1}_{S^{\beta}}\cdot v_{p^1}$=12	$\varphi^{2}_{S^{\beta}}\cdot v_{p^1}$=20	$\varphi^{3}_{S^{\beta}}\cdot v_{p^2}$=60	$\varphi^{4}_{S^{\beta}}\cdot v_{p^2}$=95
S^{γ}:$\varphi^{1}_{S^{\gamma}}\cdot v_{p^1}$=7	$\varphi^{2}_{S^{\gamma}}\cdot v_{p^1}$=16	$\varphi^{3}_{S^{\gamma}}\cdot v_{p^1}$=21	···
S^{τ}:$\varphi^{1}_{S^{\tau}}\cdot v_{p^1}$=5	$\varphi^{2}_{S^{\tau}}\cdot v_{p^1}$=7	$\varphi^{3}_{S^{\tau}}\cdot v_{p^1}$=10	$\varphi^{4}_{S^{\tau}}\cdot v_{p^1}$=6

Instances In s^{τ}	Instances co-occur with $\varphi_{S^{\tau}}$	Co-occurring Property Value Ranges (b)
$\varphi^{1}_{S^{\tau}}$	$\varphi^{1}_{S^{\alpha}}$ $\varphi^{2}_{S^{\alpha}}$ $\varphi^{1}_{S^{\beta}}$ $\varphi^{2}_{S^{\beta}}$ $\varphi^{1}_{S^{\gamma}}$ $\varphi^{2}_{S^{\gamma}}$	$\{S^{\tau}_{p^1}\vert$c:[5,5],$\{S^{\alpha}_{p^1}\vert$c:[8,15], $S^{\beta}_{p^1}\vert$c:[12,20], $S^{\gamma}_{p^1}\vert$c:[7,16]$\}\}$
$\varphi^{2}_{S^{\tau}}$	$\varphi^{1}_{S^{\beta}}$ $\varphi^{2}_{S^{\gamma}}$ $\varphi^{3}_{S^{\gamma}}$ $\varphi^{3}_{S^{\tau}}$	$\{S^{\tau}_{p^1}\vert$c:[7,10], $\{S^{\beta}_{p^1}\vert$c:[12,12], $S^{\gamma}_{p^1}\vert$c:[16,21]$\}\}$
$\varphi^{3}_{S^{\tau}}$	$\varphi^{3}_{S^{\alpha}}$ $\varphi^{3}_{S^{\beta}}$ $\varphi^{4}_{S^{\beta}}$ $\varphi^{2}_{S^{\tau}}$ $\varphi^{4}_{S^{\tau}}$	$\{S^{\tau}_{p^1}\vert$c:[6,10], $\{S^{\alpha}_{p^2}\vert$c:[14,14], $S^{\beta}_{p^2}\vert$c:[60,95]$\}\}$
⋮	⋮	⋮

c
Low-level Data Sources Combination
Target Data Source
S^{τ} 20
S^{α} 4
S^{β} 12
S^{γ} 15
$S^{\alpha}S^{\beta}$ 3
$S^{\alpha}S^{\gamma}$ 2
$S^{\beta}S^{\gamma}$ 12
$S^{\alpha}S^{\beta}S^{\gamma}$ 1
$\varphi_{S^{\tau}}$ $\varphi_{S^{\gamma}}$
$\varphi_{S^{\tau}}$ $\varphi_{S^{\beta}}$ $\varphi_{S^{\gamma}}$
High-level Data Sources Combination
Co-occurring Instances

Fig. 3: (a) The instances extracted from data sources (one row for each data source), (b) a co-occurrence table (20 co-occurring instances, each described by a row), and (c) a pruning graph built with 20 co-occurring instances.

identifying low-level patterns that comprise correlated value ranges between the properties of target data source s^{τ} and another data source and (2) **High-Level Pattern Mining** involves identifying high-level patterns that comprise correlated property value ranges across multiple data sources.

1) Low-level Pattern Mining: Low-level pattern mining aims to find significant combinations of correlated value ranges between a property in the target data source s^{τ} and one in another data source by a) aggregating the co-occurring property value ranges discovered at the previous stage with *value matrices* and b) performing a novel sweep-line based algorithm on the value matrices to identify salient correlated property value ranges that satisfy the given threshold λ_a.

Range Aggregation. We first enumerate every possible pair of properties. Then, we aggregate all combinations of co-occurring property value ranges (which we have discovered at the previous step) between two properties in each pair. The aggregation is achieved with a value matrix (Fig. 2(c)). Axes of the value matrix represent the categorical or discretized numerical domains of properties p^x and p^y, while each cell (i, j) in the matrix denotes a property value combination of $v_{p^x} = i$ and $v_{p^y} = j$. We overlay the combinations of co-occurring property value ranges extracted from the co-occurrence table (rectangles in Fig. 2(d)) on the value matrix and maintain the combination and instance counts collected from the covered range combinations for each cell.

Pattern Identification. Given a value matrix, salient patterns appear as the rectangular areas on the matrix in which the instance count of every cell covered by the area is larger than $\lambda_e \cdot |s^{\tau}|$, where$\lambda_e$ means a user-desired correlation coverage and $|s^{\tau}|$ represents the number of instances of s^{τ}. To detect these areas, we develop a fast algorithm based on the sweep line (Alg. 1). Specifically, we sweep the domain of a property column by column and construct rectangular areas along the way. For each column, the algorithm detects vertically continuous *sweep windows* (cf. L3) in which every cell satisfies the constraint and maintains two states, namely, *ASW* for the active sweep windows detected in the previous column and *NSW* for the new ones detected in the current column. To replace ASW with NSW, the algorithm considers three cases:

- ◇ ***Case 1: Continued.*** A sweep window in ASW entirely continues in NSW, thereby remaining active (cf. L5-6).
- ◇ ***Case 2: Discontinued.*** A sweep window in ASW completely disappears in NSW, thereby being removed from ASW. A new rectangular area will be constructed from the swept area and inserted into the result set RS (cf. L7-9).
- ◇ ***Case 3: Partially continued.*** A sweep window in ASW only partially continues in NSW. This sweep window will be invalidated and regarded as a discontinued window (Case 2), and a new shrunk sweep window will be created with the rows that continue from ASW to NSW (cf. L11-14).

Algorithm 1: Sweep-line Pattern Mining Algorithm

```
Data: Value Matrix $VM$, desired coverage coverage $\lambda_e$.
Result: The result set $RS$ with maximal rectangles in the
        matrix.
1  $ASW \leftarrow \emptyset$, $NSW \leftarrow \emptyset$ ;
2  for Each column col in $VM$ do
3      $NSW \leftarrow$ continuous qualified (satisfying $\lambda > \lambda_e$) cells
         in col ;
4      for sweep window $sw \in ASW$ do
5          if sw continues in col then
6              keep $sw$ in $ASW$ ;            /* cont'd */
7          else if sweep window sw not continue in col then
8              $RS \leftarrow$ result($sw$ and $col$) ;   /* discont'd
                 */
9              remove $sw$ from $ASW$ ;
10         else
11             $RS \leftarrow$ result($sw$ and $col$) ;
                /* part. cont'd */
12             remove $sw$ from $ASW$;
13             shrink $sw$ to the partially overlapped range $sw'$;
14             $ASW \leftarrow sw'$ ;
15     for sweep window $sw \in NSW$ do
16         if sw does not have the same window in ASW then
17             $ASW \leftarrow sw$
```

In addition, new sweep windows in NSW which are not covered by the above cases will be added to ASW (cf. L15-17). Hence, the low-level patterns between two properties (i.e., the combinations of property value ranges satisfying the given threshold λ_e) are obtained from the result set RS.

2) High-level Pattern Mining: The low-level patterns enable the framework to generate and validate high-level patterns that involve three or more data sources (target included).

Candidate Generation. High-level pattern candidates can be generated by intersecting low-level patterns. For example, pattern candidate $\{s^{\tau}_{p^u}|(\mathcal{C}' \cap \mathcal{C}''), \{s^{\alpha}_{p^x}|\mathcal{C}, s^{\beta}_{p^y}|\mathcal{C}\}\}$ can be generated from the intersection of low-level patterns $\{s^{\tau}_{p^u}|\mathcal{C}', \{s^{\alpha}_{p^x}|\mathcal{C}\}\}$ and $\{s^{\tau}_{p^u}|\mathcal{C}'', \{s^{\beta}_{p^y}|\mathcal{C}\}\}$. We only keep candidates whose property value ranges are not empty.

Pattern Validation. A pattern candidate is considered valid only if the number of the instance combinations it covers is larger than $\lambda_e \cdot |s^{\tau}|$. Valid high-level patterns are inserted into the result set for further interactive analysis.

V. Visual Design

This section discusses analysis tasks, design rationales, and the visualization techniques specifically designed for interpreting the extracted patterns.

A. Design Rationales

Although the model can efficiently extract correlation patterns, interpreting the massive correlations, detecting anomalies, and obtaining high-level insights remain challenging. Visualization techniques are highly necessary to help explore the extracted correlation patterns.

In this study, we have conducted a user-centered design process with three interdisciplinary urban planning experts over the past year. These experts have more than 10 years of experience in developing data-driven solutions for various urban problems, such as location selection, energy planning, and pollution analysis. They approached us to seek an interactive visualization system for interpreting and analyzing the correlation patterns among different heterogeneous data sources, including city-wide meteorological and traffic data, collected in urban environments. Through frequent discussions with the experts, two important analysis tasks, *macro-* and *micro-level analyses*, were identified.

Macro-level analysis. Users select proper mining parameters and wish to see the statistical distribution of all value ranges regarding individual properties. Users also select properties and value ranges that they are interested in for further analysis. A visual summary of correlation patterns should be provided to help users determine a specific property combination and proceed to the micro-level analysis to inspect the correlation patterns in this combination.

Micro-level analysis. A clear overview of the correlation patterns of a given property combination should be provided. Subsequently, users may group interesting patterns for observation and comparison. The spatiotemporal distribution of the instances of the patterns should also be provided for further validation and analysis of the patterns.

Based on these two analysis tasks, the design rationales behind our system are derived and summarized below.

R1 **Generating a visual summary of massive patterns**
A large number of patterns is difficult to analyze individually, but these patterns are included in various property combinations. Thus, users highly desire a visual summary.

R2 **Allowing statistical analysis of properties**
Obtaining an intuitive understanding of the overall data range distribution is difficult. Users need a visualization that presents the statistical information of the range distribution of each data property.

R3 **Enabling interactive visual exploration of patterns**
The system should allow domain experts to interact with the patterns directly by supporting various interactions like filtering, ranking, and grouping to unfold the patterns.

R4 **Visualizing the spatiotemporal information of patterns**
A pattern can be associated with many spatiotemporal instances. The system should show the spatiotemporal trend of the pattern.

R5 **Applying different model parameters**
The mining model may not always produce the desired results. Thus, user interaction with the model should be supported to select different results of the model.

In the design process, we identified three challenges, namely, diversity, volume, and organization (detailed in Section I). We tackle these challenges by designing CorVizor with four linked views, including CorView, STView, StatView, and PatTable (Fig. 4), based on the aforementioned rationales. CorView is the core component and provides a matrix-style visual summary of the patterns of all property combinations (**R1**). Multi-level interactive exploration is naturally supported (**R3**). StatView displays the distributions of value ranges of different data source properties (**R2**). STView shows the spatiotemporal information of the target instances associated with the patterns (**R4**). PatTable presents the details of the selected patterns in a table (**R3**). Choosing different model parameters are also supported (**R5**) in the Info Panel (Fig. 4(a) and Fig. 10(a)).

B. CorView

This section presents the design of CorView, which visually summarizes the patterns of property combinations (Section V-B1) and interactively unfolds those of a selected property combination (Section V-B2).

1) Visualization of Property Combinations: We adopt a scalable matrix-based approach (the **volume** challenge) to visualize the property combinations shared by massive correlation patterns (**R1**) and provide an unified overview for the diverse data properties among the patterns (the **diversity** challenge). The matrix-based approach is easy to understand and allows users to make efficient visual comparisons of the property combinations in an aligned manner.

Each column in the CorView represents a property, and each row (Fig. 4(i)) represents a group of correlation patterns with the same set of properties (i.e., the same property combination). The properties in the target data source and other data sources are marked in green and orange, respectively. The properties in each row are encoded by a set of linked brick-like density glyphs. Each glyph contains a density map (Fig. 4(c)), which reveals the distribution of the value ranges in the patterns with the same combination of properties for the corresponding property. The links between glyphs indicates the correlation of properties. The numbers of patterns and pattern instances for each property combination are visualized with two bars in each row (Fig. 4(d)). The blue area in each bar (Fig. 9(j)) indicates the number of the selected patterns or instances in other views. By clicking on the column headers (Fig. 4(b, k), users can filter out the property combinations that do not contain the selected properties or sort the combinations based on the numbers of patterns or instances.

2) Visualization of Correlation Patterns: Users can unfold a property combination in the CorView and analyze the patterns with the selected combination using a similarity-based scatterplot (**R1**), a tailored parallel coordinates plot, and a stacked line chart (Fig. 4(h)) in the expanded view (Fig. 4(e)). The scatterplot provides an overall picture of the similarity among correlation patterns, thereby enabling users to group the patterns and detect anomalies. The parallel coordinates plot depicts the correlation among multiple properties, where the value range distributions of the patterns w.r.t. each property are

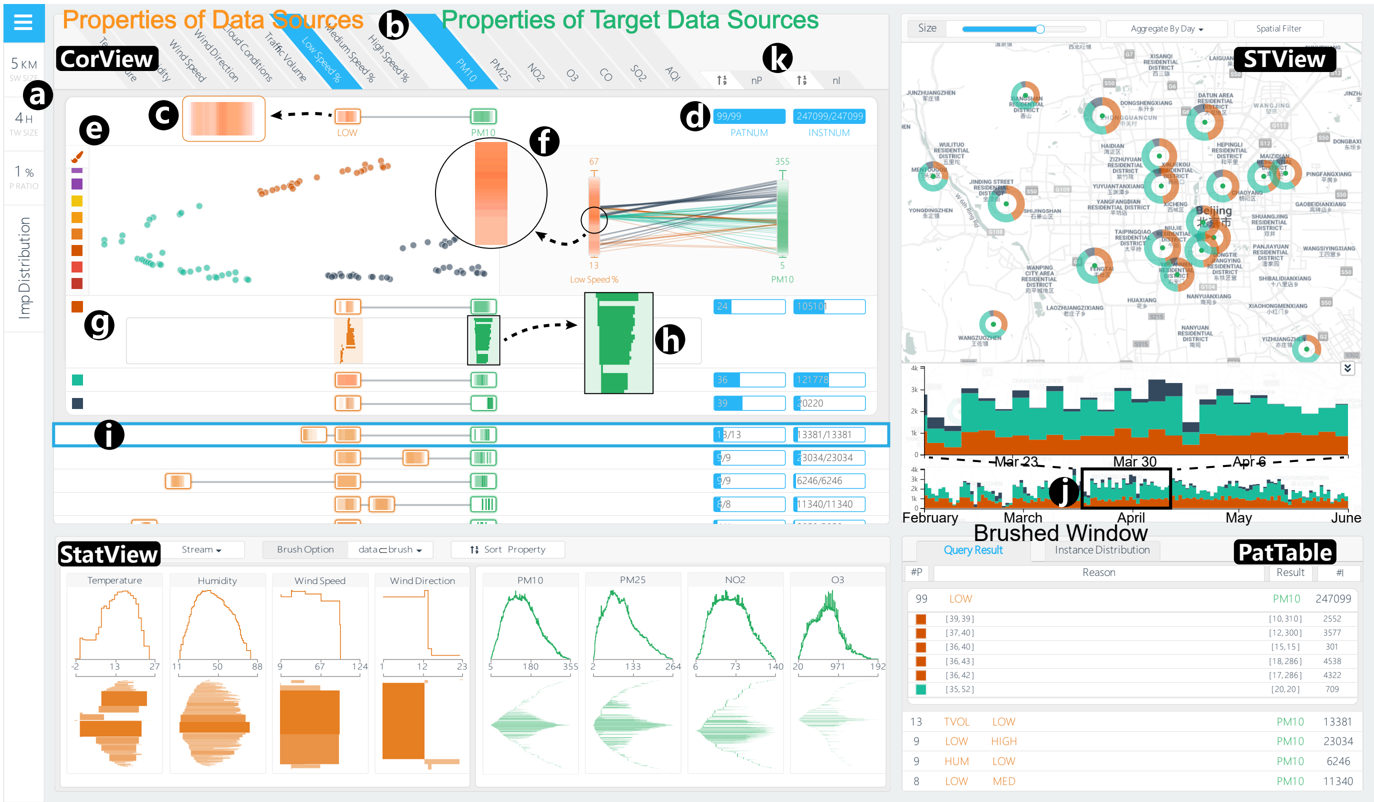


Fig. 4: CorVizor consists of four main views (CorView, STView, StatView, and PatTable) for detecting and understanding correlation patterns.

encoded with a density map on the axis. The stacked line chart further provides a compact visualization of the value ranges of the selected patterns. These scalable visualizations are combined to organize patterns without severe visual clutters and facilitate the effective exploration of both the overview and details (the **volume** and **organization** challenges).

Scatterplot. Analyzing the relationship between patterns is important for pattern grouping, comparison, and anomaly detection. A scatterplot is used to display the correlation patterns, such that similar patterns are naturally grouped together. This scatterplot provides a concise overview of pattern relationships with less clutter than parallel coordinates. The multidimensional scaling (MDS) is used to create the scatterplot. The distance of patterns i and j is computed with $\sqrt{\sum_k^n d(i_k, j_k)^2}$, where n is the number of properties in the pattern, k denotes property k, and $d(i_k, j_k)$ is the distance between the two ranges with regard to property k of pattern i and j. The Jaccard index and KL divergence were tested to calculate the distance. However, distance is regarded as a constant value by both measures when two ranges are disjoint regardless of how far the ranges semantically appear. Thus, a new measure is used. In this new measure, four features are extracted from each range: lower bound (lb), upper bound (ub), median (mid), and length (len). All features are normalized. The range distance of property k is measured with the Euclidean distance of the pair of ranges, namely, $d(i_k, j_k) = \sqrt{\Delta lb^2 + \Delta ub^2 + \Delta mid^2 + \Delta len^2}$, where Δ represents the difference between two feature values. For those categorical properties that can be ordered, we assign numeric values for each category starting from 1 by the categorical order and compute the range distance based on these values. For those categorical properties that cannot be ordered, we map the text descriptions of those categories to high-dimensional space with word2vec [31], [32]. The word2vec model can generate a high-dimensional vector for each word considering their semantics in a series of sentences. The Euclidean distance between two vectors indicates the semantic similarity between two corresponding words. As such, the distance between the two categories can be measured. Users can group patterns and highlight anomalies by brushing the corresponding points with various colors.

Parallel coordinates. Correlation patterns can have a high-level form (Section IV-C2) with more than three properties involved. Thus, parallel coordinates are used as a uniform view to display the multidimensional correlation patterns. Each axis represents a property. The medians of the ranges are used as the end points of the line segments to connect the value ranges in various property axes. Considering that overlaps exist among ranges of the same property, we do not adopt parallel sets as it is more suitable for categorical and disjoint data.

The range distribution is displayed with a density map on its corresponding axis (Fig. 4(f)). A density map is used instead of other methods, such as histogram, because it consumes less space and compactly shows the density information of the property value. In each density map, value ranges are drawn along each coordinate with equal opacity. The ranges are overlaid and their opacity values are combined to encode the density (i.e., dark areas indicate that the corresponding values

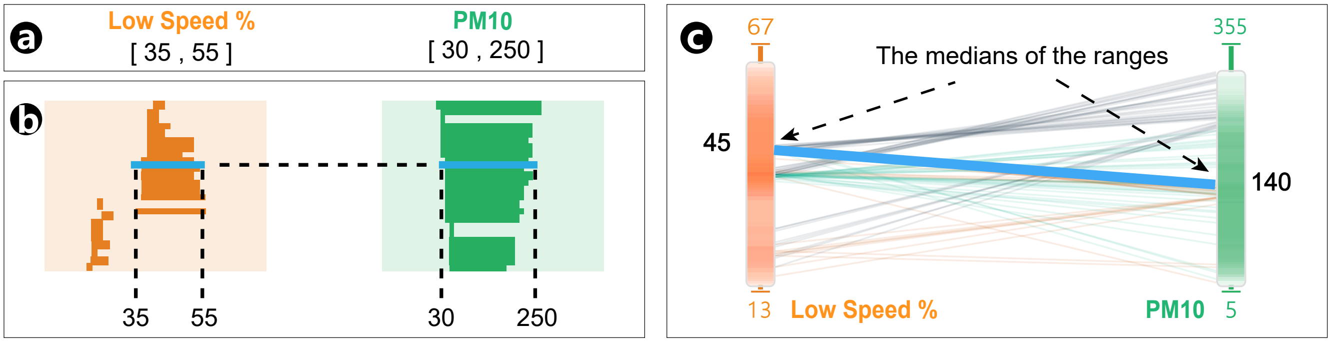


Fig. 5: Example of the visual encodings of the stacked line chart (b) and parallel coordinates (c) for a pattern in (a).

or categories are covered by many ranges). Fig. 5 (a) shows an example of a pattern and Fig. 5 (c) shows the corresponding visual representation with tailored parallel coordinates.

Stacked line chart. When a set of patterns are grouped in the scatterplot, a new row summarizing the pattern group is automatically generated and attached under the scatterplot and parallel coordinates. Fig. 4(g) shows one of the three rows of the grouped patterns. When a user unfolds a row, the row is expanded to show a stacked line chart (Fig. 4(h)). In the chart, the value range for each property is represented by a fine line segment. The segments are stacked to compose a distribution map. Fig. 5(b) shows an example of the chart, where the left and right endpoints of the line segments denote the two endpoints of the range.

3) Design Alternatives: In the aforementioned user-centric design process, we attempted to refine the visual design of CorVizor iteratively by proposing and evaluating alternatives. In this section, two design alternatives are discussed to reveal the rationales behind our design choices in terms of the macro- and micro-level analyses of correlation patterns.

Visualization of patterns in many property combinations. Instead of organizing property combinations with a matrix-based CorView, we attempted to maintain the structure of these combinations with a node-link diagram (Fig. 6(a)). Each node in the diagram represents a property combination. The directed edges in the diagram indicate the composition of subsequent combinations. In each node lies a glyph, which encodes the distribution of property value ranges, and the size of the glyph shows the number of pattern instances. Moreover, we allow users to apply filters to keep the desired combinations by selecting properties on the top. Although such an alternative clearly reveals the inherent structure of property combinations, three major weaknesses prohibit it from being applied in our system, that is, a) the proposed node-link diagram costs excessive screen space; b) the crossing edges introduce serious visual clutters and are thus not scalable; and c) the distributions of property value ranges in different nodes are difficult to compare because they are not aligned. Hence, we decided to adopt a compact matrix-based view and facilitate the comparison between properties with alignment.

Visualization of patterns of a property combination. To help analysts grasp the similarity among massive correlation patterns, we initially projected these patterns into a 2D view via dimensional reduction techniques. Inspired by Liu et al. [27], we attempted to depict these patterns with glyphs embedded directly into the view (Fig. 6(b)). On the edge of the glyph lies a circular histogram that encodes the temporal pattern distribution of a single pattern, and the properties involved in the pattern are represented by homocentric donut charts. However, such an approach is not scalable with the number of patterns. The value ranges encoded with radians can also be misleading. Hence, we iterated our design by dissecting the high-dimensional information in these patterns with multiple coordinated views as described previously.

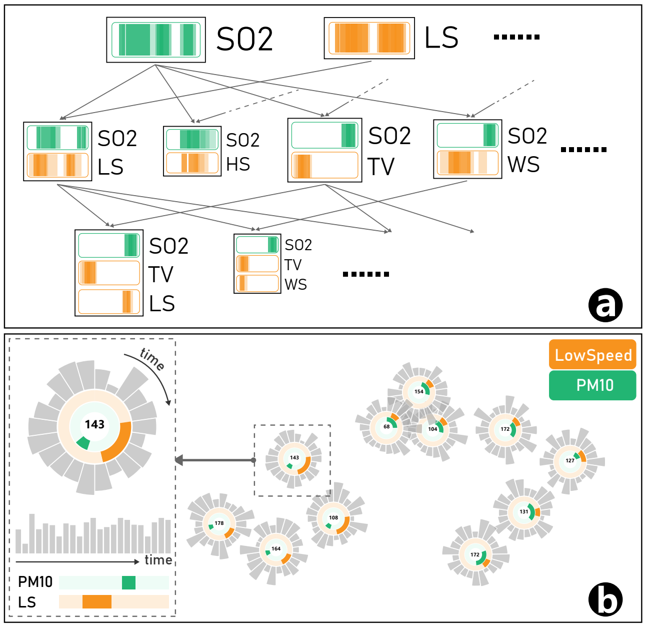


Fig. 6: Design Alternatives for CorView. (a) Node-link visualization for many property combinations; (b) Glyph-based visualization for correlation patterns.

C. StatView

Although CorView shows the property combinations and their correlation patterns, the value range distribution aggregated by properties remains unavailable. This information is essential for high-level exploration. Thus, StatView is used to display the pattern distribution of each property (Fig. 4) (**R2**). This pattern distribution comprises small multiples that display the distributions of the value ranges for the properties. An individual plot of a property displays a value distribution in a line chart (upper part) and presents a range distribution in a stacked rectangle chart (lower part). Both parts have their counterparts in CorView. The upper and lower parts correspond to the density map and stacked line chart, respectively.

StatView uses a line chart to encode the density distribution of the values or categories of each property. Position information is considered perceptually effective in encoding magnitude [35]. Moreover, StatView has more space to display the distribution information. Thus, we utilize position rather than luminance to encode the distribution information. The same value or category ranges are aggregated, and a rectangle is used to represent each unique range in the stacked rectangle chart. The height of a rectangle represents the number of occurrences of the associated range.

Users can brush a span of the property value on the horizontal axis of any line chart to perform filtering. Other views can be subsequently updated. StatView supports three types of filtering interactions. The selected patterns should satisfy the following constraints (1) $s \cap d \neq \emptyset$, (2) $s \subset d$, and (3) $s \supset d$, where s indicates the brushed span and d denotes all data correlation patterns.

D. STView

STView allows users to gain insights into the spatiotemporal trends of the patterns (**R4**). The bottom of the view shows a histogram to visualize the temporal distributions of pattern instances with multiple scales. Users can easily select patterns by brushing a temporal window (Fig 4(j)). When the patterns are grouped in CorView, the histogram shows the temporal distributions of pattern instances of different groups by using stacked bars. The top of the view shows the spatial distribution of the pattern instances on a map. In this study, we use air quality data as the target data source. Each air quality station is represented by a donut chart whose radius encodes the total number of pattern instances in this location. The sectors in different colors indicate the ratios of the pattern instances of each group selected from CorView in each location. Donut charts are used instead of pie charts because the former has a blank center. Users can see through it to observe the details.

When a user hovers his or her mouse on a glyph, a circle around the glyph is displayed to show the coverage of the associated station, namely, the size of the spatial window used in the mining model. The circle covers the spatial area whose correlation instances can be viewed as being co-located with the air quality station. Users can select several stations to see the related correlation patterns in other views.

E. PatTable

PatTable is a table-like component that allows users to inspect raw patterns directly on demand. Each row represents a property combination. The combinations can be unfolded to show the corresponding patterns. Detailed information for each individual pattern, such as the number of instances and the correlated property value ranges, is depicted in the unfolded view. Moreover, PatTable is coordinated with CorView, where user interactions in one view are reflected in the other view.

F. User Interactions

CorVizor supports various basic and advanced interactions.

- ◇ **Showing overview first and details on demand.** CorView shows a succinct overview of all property combinations. Users can click on a row to explore the corresponding property combination in detail.
- ◇ **Brushing and filtering.** Users are allowed to group patterns or spot anomalies by brushing the patterns with colors in a scatterplot in CorView. Users can filter by spatial area, time range, and property value range in STView and StatView.
- ◇ **Changing the model parameters.** Users can change the parameters including the spatiotemporal threshold and minimum correlation coverage (Fig. 10(a)) and see new results (**R5**). The histogram (Fig. 10(b)) shows the distribution of the normalized range widths of all correlation patterns.

VI. Experiments

This section presents model evaluation, case studies, and expert interview to evaluate the effectiveness and usability of the proposed system. The experimental data contain the 16 data properties from the three data sources listed in Table I in Section III-A. The data were collected from a large city. Data collection was conducted from February 1 to May 31 in 2014. **Weather data** were collected hourly from 20 weather monitoring stations around the city. **Air quality data** were collected hourly from 36 air quality monitoring stations in the city, and **traffic data** were collected from $100,215$ segments of the city road network every half hour from a geospatial mapping platform. To sum up, there are 103 thousand records in the air quality data, 57 thousand records in the weather data, and 577,238 thousand records in the traffic data. All experiments were evaluated on a laptop running Windows 10 with Intel Core i7 3.4GHz CPU, 256GB SSD drive, and 16 GB RAM.

A. Model Evaluation

The proposed sweep-line algorithm is the core component of our pattern mining model. We compared it with a naive approach to demonstrate its efficiency and effectiveness.

Naive Approach. The naive method to identify distinctive rectangles from a value matrix follows the following steps. First, every cell in the matrix is scanned. Second, if a qualified cell (fulfilling the correlation coverage requirement) is identified, the naive approach considers the cell the left-up corner of certain distinctive rectangles and traverses toward right and down directions to find the rectangles as candidates. Third, each candidate rectangle is tested to see if it is completely covered by other rectangles identified previously. If overlapping cases exist, the candidate rectangle is discarded. Otherwise, the identified rectangle is inserted into the result set. The cost of the approach is prohibitively high. Assuming that an $M \times N$ matrix exists, the approach needs to traverse the entire matrix to identify qualified cells in the outer iteration. For each qualified cell, the approach needs to traverse the remainder of the matrix to identify the distinctive rectangles and test their qualification, which may result in $\mathcal{O}(M^2N^2)$ in the worst case. In contrast, our method has the time complexity of $O(MN)$ as we only need to scan the matrix once.

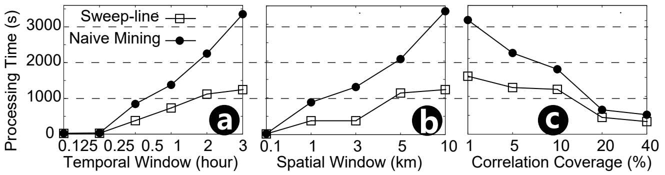


Fig. 7: Comparison of the time performances of the proposed sweep-line algorithm and naive mining method.

The comparison was performed with the varying correlation coverage, spatial window, and temporal window. Fig. 7 shows the results of the time performance comparison.

1) Temporal window. Fig. 7(a) shows that the time of both approaches increases with the increase in the temporal window. A large temporal window usually generates a large value range, which increases the probability of finding qualified cells in the value matrix. The sweep-line algorithm outperforms the naive approach, given that a significant number of redundant cell examinations are avoided during the process.

2) Spatial window. Fig. 7(b) provides two observations. First, with a large spatial window, the processing time of both approaches increase because many co-related instances are to be analyzed in a spatiotemporal partition. Second, the difference in the performance of the two approaches is significant, as the naive approach examines more qualified cells and incurs more computational time.

3) Correlation Coverage. Fig. 7(c) presents two observations. First, the processing time of both approaches decreases. Second, the sweep-line algorithm performs better than the naive one because a large correlation coverage value results in a small chance of finding a qualified cell (i.e., qualified patterns) in the value matrix.

B. Case Studies

The case studies were conducted with the domain experts to evaluate the effectiveness of our system.

1) Macro Analysis: Correlations Relevant to High SO_2*:* This case study demonstrated the effectiveness of CorVizor for the macro-level analysis (detailed in Section V-A).

Selecting proper mining parameters was the first step to explore the correlation patterns. The domain experts suggested 5 km and 4 hours for spatial and temporal window based on their experience. However, the correlation coverage threshold is difficult to choose. The experts attempted the thresholds 0.5%, 1%, and 3%. The distributions of the normalized range widths generated by these thresholds are presented in Fig. 8(a), and the associated statistical information on the extracted patterns is depicted in StatView (Fig. 8(b)). Based on their observations, the experts selected 1% as the threshold because: a) although the histograms generated by the thresholds 0.5% and 1% seemed similar, the patterns represented as stacked rectangles in StatView with the threshold 1% were more organized and meaningful than those with 0.5%; and b) the patterns extracted with the threshold 3% was too coarse to reveal any useful insights. Thus, the threshold 1% (Fig. 10(a)) was selected by the experts for further explorations.

Urban air pollution, which is crucially related to the well-being of city residents, has attracted increasing concerns in recent years. Therefore, the experts attempted to identify the correlations between air quality and other urban data sources with CorVizor. In particular, they were interested in the correlation patterns relevant to high SO_2 because SO_2 was one of the major pollutants produced by human activities in cities. Hence, the experts selected the patterns that comprised high SO_2 in StatView. Fig. 9(b) showed that low traffic volume was strongly correlated with high SO_2 because the bars in the range distribution view of traffic volume indicated that the corresponding value ranges were relatively low and narrow.

These findings seemed contradicted with the experts' intuition, as they believed that only huge traffic volume would result in severe air pollutant emission. Hence, they selected traffic volume and SO_2 in CorView for further exploration (Fig. 9(c)). Only the property combinations that involved these two selected properties remained in the view. The glyphs in the first row (Fig. 9(d) and 9(e)) validated the aforementioned observation with StatView. By analyzing other rows, the experts discovered that the number of low-speed vehicles (Fig. 9(f)) was considerably larger than that of high-speed vehicles (Fig. 9(g)) while the traffic volume (Fig. 9(h)) was low and SO_2 (Fig. 9(i)) was high. Thus, the experts suggested that the large number of slow vehicles and small traffic volume could be a sign of potential traffic congestions, which resulted in the high SO_2 emission. The correlation patterns among traffic volume, low-speed vehicles, and AQI level were also explored with the identical approach. The result was similar: the air quality appeared to be bad with small traffic volume and the large number of slow vehicles. This insight confirmed that the small traffic volume correlated with severe traffic congestions, which were a significant contributing factor to the deteriorated urban air quality.

2) Micro Analysis: Correlations Involving Air Pollution: The second case study demonstrates the usefulness of the system in analyzing the correlation patterns associated with a specific property combination.

Road space rationing policies were widely adopted by

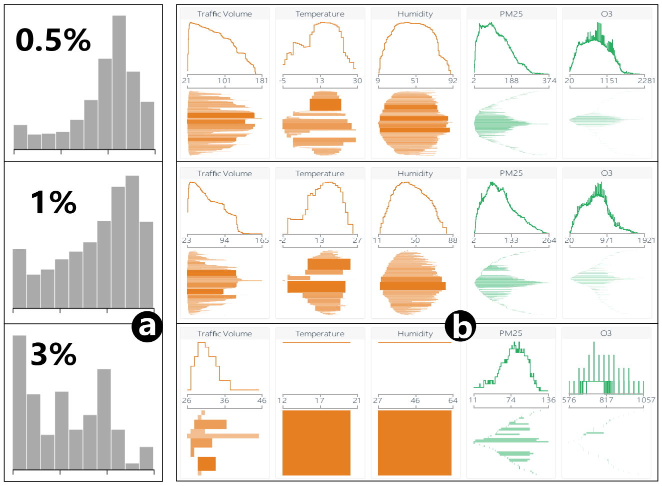


Fig. 8: Selecting proper parameters: (a) histograms showing the distribution of the normalized range widths and (b) statistical information for correlation coverage thresholds 0.5%, 1% and 3% in StatView.

governments to alleviate serious air pollution. However, the experts doubted the effectiveness of these policies. Hence, they would like to analyze the correlation between traffic and air pollution with our system. The combination of low speed (%) and PM10 in CorView was selected in this study. The corresponding row was expanded to show its details for an in-depth exploration. Patterns that were extremely general were filtered out by brushing the histogram of the normalized range width (Fig. 10(b)). The patterns in the scatterplot under the expanded row were grouped in different colors based on the closeness of the patterns in the plot (Fig. 10(c)). The parallel coordinates were colored accordingly (Fig. 10(d)). The red group is considerably different from the blue and green groups in parallel coordinates (Fig. 10(d)). The red group represents "less low speed (%) and high PM10," whereas the blue and green groups indicate "more low speed (%) and high PM10."

The experts were particularly interested in the red pattern group. The stacked line chart of the group in Figs. 10(f) and 10(g) also indicates "less low speed (%) (i.e., traffic congestion is unlikely to occur) and high PM10." STView was used to examine the spatiotemporal distribution of the patterns of the selected groups (Fig. 10(e)). To experts' suprise, the red group only occurred in the area between Rings 5 and 6 of the expressways, which is the suburban area of the city. One expert indicated that there were several garbage incineration plants in this area. Comparison between the distribution of the red group patterns (Fig. 10(h)) and that of the garbage incinerators (Fig. 10(i)) showed a clear match. They speculated that the garbage incinerators could be highly correlated with the high PM10 in the area, in which traffic congestion did not occur. Further investigation and analysis in the field were required to verify this conjecture and determine its plausible cause.

Furthermore, CorVizor was used to explore the correlation patterns involving air quality index (AQI), which would increase as air quality worsens. The experts were curious about the reasons behind the worst air quality represented by the highest AQI level with the value of 5. Thus, they drew a selection on the AQI property in StatView (Fig. 11(a)) to select the patterns that involved the highest AQI level. From STView, the experts observed that most of these patterns occurred around February (Fig. 11(b)). They suggested that air pollution might be caused by coal heating in the winter, which emitted massive pollutants and severely deteriorated the air quality. To confirm this hypothesis, the experts selected the temperature and AQI properties in CorView (Fig. 11(c)) and discovered that the highest AQI level correlated with low temperature. Moreover, the temporal distribution of these selected correlation patterns was identical to that of the patterns involving the highest AQI level (Fig. 11(d)). The temperature ranges in PatTable were around 0 °C (Fig. 11(e)), which also provided useful hints for this correlation. Furthermore, the experts attempted to verify the correlation by selecting the patterns with low temperature in StatView. They were satisfied to discover that these patterns were all correlated to medium and high AQI levels (Fig. 11(f)). These observations effectively supported the experts' hypothesis and helped them link the deteriorating air quality with coal heating.

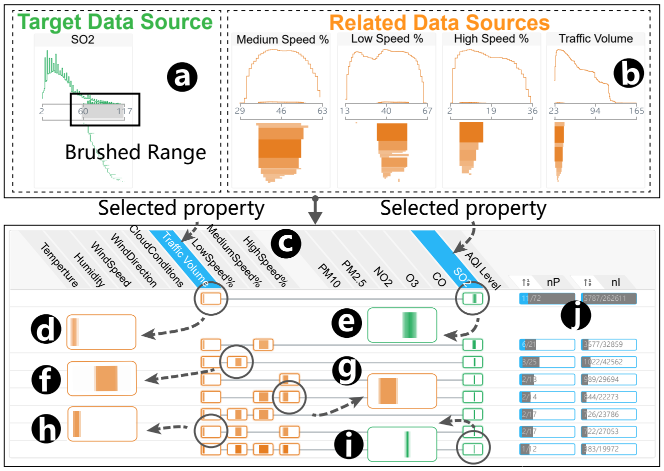


Fig. 9: Macro-level analysis of the correlation patterns that are relevant to high SO_2.

In this case study, the correlation patterns involving both numerical and categorical properties were explored and analyzed in detail. These detailed exploration and analysis demonstrate the effectiveness of CorVizor in handling the micro-level analysis tasks and providing interesting insights into the correlation patterns for further verification and analysis.

C. Interview with Domain Experts

After the case studies, we collected and summarized the feedback from the experts as follows.

Overall System Usability. CorVizor was well received by the experts. They were pleased to explore and analyze the massive heterogeneous patterns intuitively with the proposed interactive visualizations. *"The visualization system makes the correlation patterns produced by the data mining model much more meaningful,"* an expert said. Both experts acknowledged that the analytical workflow of our system could help them gain considerable insights into the spatiotemporal correlations. Moreover, they believed that our system could be extended to identify interesting correlations in various scenarios, such as business location selection and travel recommendation.

Visual Design and Interactions. Both the experts were impressed by the visual design and interactions. They praised CorView, which presents the correlations among various data properties explicitly. An expert commented *"the matrix-like layout is familiar to me and the hierarchical visualization method well organizes the exploration process."* He was also highly satisfied with the intuitive visual summary of the correlation patterns provided in CorVizor. Another expert was deeply impressed by the spatiotemporal view. *"Without this system, it would be impossible to discover interesting cases related to the spatiotemporal distribution of the correlation patterns,"* he said. Both experts appreciated the interactive features of the proposed system. They especially appreciated the usefulness of filtering and brushing. The experts said that these techniques help in anomaly detection and pattern grouping and comparison.

Suggestion. The usability of our system was confirmed by the experts, who immediately became familiar with the system after a brief training. However, they suggested that the design of our system could be simplified further, such as by replacing

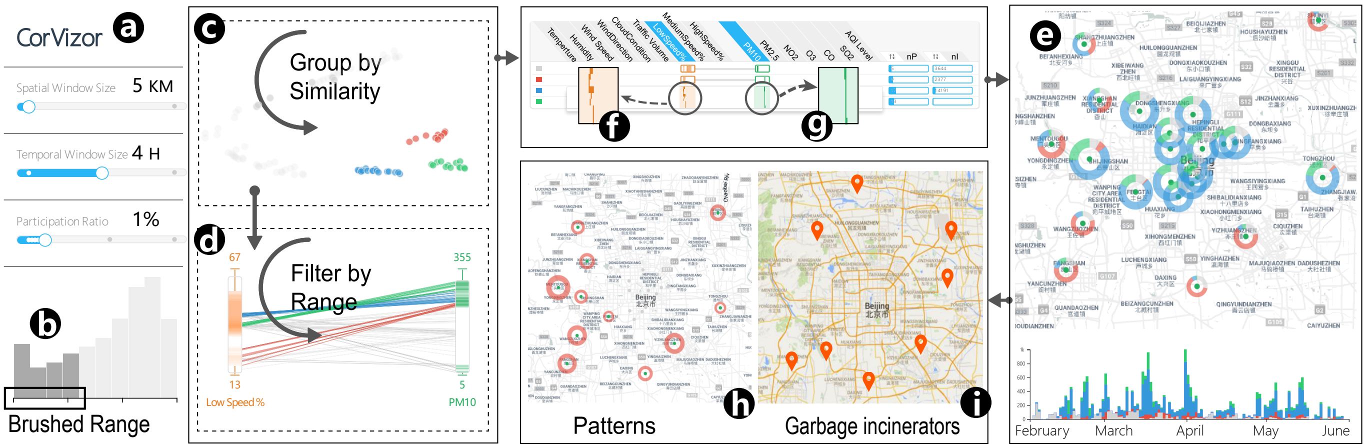


Fig. 10: Micro-level analysis of correlation patterns between traffic congestion and air pollution.

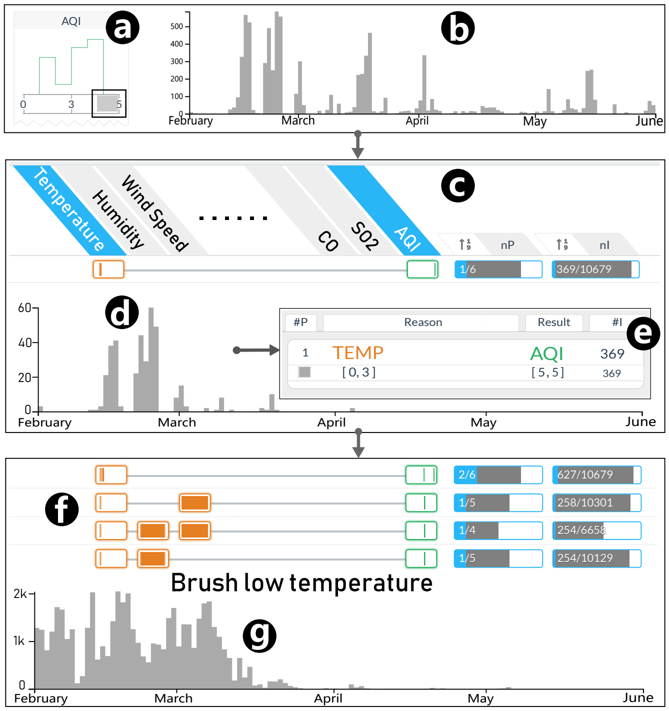


Fig. 11: Micro-level analysis of correlation patterns between temperature and air quality index (AQI).

the scatterplot and parallel coordinate plot with a plain list with numbers and figures. They also suggested that the design should be integrated with visual guides to allow average users, such as government officials to monitor the city dynamics and grasp interesting insights conveniently. We will leave this simplified version of our system as a part of our future work.

VII. Discussion

In this section, we discuss the implications, limitations, and generalization of the proposed system.

Implications. CorVizor can identify interesting correlation patterns that may facilitate numerous *transportation applications*, such as traffic management and transportation planning. Important insights revealed by these patterns, including how the traffic speed and volume in a local area affect the concentrations of air pollutants, provide strong decision-making contexts for urban planners to establish informed road policies and long-term planning strategies in advance. Nevertheless, correlations do not necessarily imply causation. Analysts may not be able to come up with a clear actionable plan with only correlations, and inferencing of causal relationships remains a challenge. However, the present work still has several important implications with regard to causal inference. First, pattern correlations can reduce the search space of causal inference. Second, the special characteristics of pattern correlations can have significant implications for research on causal inference. Moreover, with CorVizor, data mining researchers can easily obtain an intuitive overview of a large number of correlation patterns while checking the credibility of any specific correlation pattern or group of correlation patterns. As such, researchers can be informed of imperfections of the data mining model, consequently inspiring them to enhance the model's effectiveness.

Limitations. The time performance of the correlation mining framework is not highly optimized. Running the model for our experimental dataset usually requires nearly an hour. Data mining results for possible parameter combinations were computed in advance to support the interactive adjustment of the model results. We plan to optimize the model and adapt it to a high-performance distributed computing platform, such that the interactive adjustment of the model setting is made possible. As for the design part of our system, MDS adopted by the scatterplot is widely used in the visualization literature, but it may be misleading at times [41]. To enhance the scatterplot, the method for visualizing dimensionally-reduced data [41] can be further incorporated into our system.

Generalization. CorVizor can be directly applied to various urban analysis applications, such as urban planning, pollution diagnoses, and location selection, to detect and understand the correlation patterns in spatiotemporal datasets that support effective decision-making processes. The case studies we presented were conducted for pollution diagnosis. However, the target data source can be changed to identify other interesting correlations in other domains. For example, traffic congestion [13], [46], [48] can be analyzed efficiently by setting traffic data as the target. In addition, the evolution of business is closely related to many latent correlation patterns extracted from various urban datasets [43], which can also be captured by our framework.

VIII. Conclusion and Future Work

In this work, we studied the extraction and interpretation of fine-grained spatiotemporal patterns that comprise various properties of different types, scales, and semantics. Based on the proposed data mining framework and interactive multi-scale visualization technique, we developed CorVizor, a visual analytics system that assists users in exploring these patterns. This study contributes an important step towards the in-depth understanding of urban dynamics formed by the complex correlation patterns extracted from heterogeneous spatiotemporal data sources, including transportation data.

We will continue on improving our system in several ways as follows. First, we plan to migrate the correlation mining module to a high-performance distributed computing platform. Users can directly interact with the model and see the results instantly in CorVizor. Second, we will deploy CorVizor in the field, such that the streaming datasets collected from diverse sources can be fed into the system in real-time, thereby enabling a proactive analysis workflow of urban problems.

References

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22(2):207–216, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proc. of VLDB*, pages 487–499, 1994.

[3] R. Agrawal and R. Srikant. PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth. In *Proc. of ICDE*, pages 215–224, 2001.

[4] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel. *Visual Analytics of Movement*. Springer, 2013.

[5] G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel. Scalable analysis of movement data for extracting and exploring significant places. *IEEE TVCG*, 19(7):1078–1094, 2013.

[6] N. Andrienko, G. Andrienko, L. Barrett, M. Dostie, and S. P. Henzi. Space transformation for understanding group movement. *IEEE TVCG*, 19(12):2169–2178, 2013.

[7] B. Aydin, A. Kucuk, R. A. Angryk, and P. C. Martens. Measuring the significance of spatiotemporal co-occurrences. *ACM TSAS*, 3(3):9, 2017.

[8] G. Bothorel, M. Serrurier, and C. Hurter. From visualization to association rules: an automatic approach. In *Proc. of the Spring Conference on Computer Graphics*, pages 57–64, 2013.

[9] N. Cao, C. Lin, Q. Zhu, Y.-R. Lin, X. Teng, and X. Wen. Voila: Visual anomaly detection and monitoring with streaming spatiotemporal data. *IEEE TVCG*, 24(1):23–33, 2018.

[10] M. Celik, S. Shekhar, J. P. Rogers, and J. A. Shine. Mixed-drove spatiotemporal co-occurrence pattern mining. *network*, 11:15, 2008.

[11] V. P. Chakka, A. C. Everspaugh, and J. M. Patel. Indexing large trajectory data sets with SETI. In *Proc. of CIDR*, 2003.

[12] S. Chen, X. Yuan, Z. Wang, C. Guo, J. Liang, Z. Wang, X. Zhang, and J. Zhang. Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. *IEEE TVCG*, 22(1):270–279, 2016.

[13] W. Chen, Z. Huang, F. Wu, M. Zhu, H. Guan, and R. Maciejewski. Vaud: A visual analysis approach for exploring spatiotemporal urban data. *IEEE TVCG*, (1):1–1, 2017.

[14] F. Chirigati, H. Doraiswamy, T. Damoulas, and J. Freire. Data polygamy: The many-many relationships among urban spatio-temporal data sets. In *Proc. of ACM SIGMOD*, pages 1011–1025, 2016.

[15] P. Cudré-Mauroux, E. Wu, and S. Madden. TrajStore: An adaptive storage system for very large trajectory data sets. In *Proc. of ICDE*, pages 109–120, 2010.

[16] H. Doraiswamy, N. Ferreira, T. Damoulas, J. Freire, and C. T. Silva. Using topological analysis to support event-guided exploration in urban data. *IEEE TVCG*, 20(12):2634–2643, 2014.

[17] G. Ertek and A. Demiriz. A framework for visualizing association mining results. In *Proc. of ICCIS*, pages 593–602, 2006.

[18] N. Ferreira, M. Lage, H. Doraiswamy, H. Vo, L. Wilson, H. Werner, M. Park, and C. Silva. Urbane: A 3d framework to support data driven decision making in urban development. In *Proc. of IEEE VAST*, pages 97–104. IEEE, 2015.

[19] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE TVCG*, 19(12):2149–2158, 2013.

[20] D. Guo and X. Zhu. Origin-destination flow data smoothing and mapping. *IEEE TVCG*, 20(12):2043–2052, 2014.

[21] M. Hahsler and S. Chelluboina. Visualizing association rules in hierarchical groups. In *Proc. of the Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms*, 2011.

[22] X. Huang, Y. Zhao, C. Ma, J. Yang, X. Ye, and C. Zhang. TrajGraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data. *IEEE TVCG*, 22(1):160–169, 2016.

[23] Y. Ke, J. Cheng, and W. Ng. Mining quantitative correlated patterns using an information-theoretic approach. In *Proc. of ACM SIGKDD*, pages 227–236. ACM, 2006.

[24] R. Kosara, F. Bendix, and H. Hauser. Parallel Sets: Interactive exploration and visual analysis of categorical data. *IEEE TVCG*, 12(4):558–568, 2006.

[25] J. Li, S. Chen, K. Zhang, G. Andrienko, and N. Andrienko. Cope: Interactive exploration of co-occurrence patterns in spatial time series. *IEEE TVCG*, 2018.

[26] X. Lin, A. Mukherji, E. A. Rundensteiner, and M. O. Ward. SPIRE: Supporting parameter-driven interactive rule mining and exploration. *Proc. of the VLDB Endowment*, 7(13):1653–1656, 2014.

[27] D. Liu, D. Weng, Y. Li, J. Bao, Y. Zheng, H. Qu, and Y. Wu. Smartadp: Visual analytics of large-scale taxi trajectories for selecting billboard locations. *IEEE TVCG*, 23(1):1–10, 2017.

[28] G. Liu, A. Suchitra, H. Zhang, M. Feng, S.-K. Ng, and L. Wong. Assocexplorer: an association rule visualization system for exploratory data analysis. In *Proc. of ACM SIGKDD (Demo)*, pages 1536–1539. ACM, 2012.

[29] Z. Liu, Y. Huang, and J. R. Trampier. Spatiotemporal topic association detection on tweets. In *Proc. of ACM SIGSPATIAL*, page 28. ACM, 2016.

[30] R. Maciejewski, R. Hafen, S. Rudolph, S. G. Larew, M. A. Mitchell, W. S. Cleveland, and D. S. Ebert. Forecasting hotspots—a predictive analytics approach. *IEEE TVCG*, 17(4):440–453, 2011.

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[32] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proc. of NIPS*, pages 3111–3119, 2013.

[33] F. Miranda, H. Doraiswamy, M. Lage, K. Zhao, B. Gonçalves, L. Wilson, M. Hsieh, and C. T. Silva. Urban pulse: Capturing the rhythm of cities. *IEEE TVCG*, 23(1):791–800, 2017.

[34] P. Mohan, S. Shekhar, J. A. Shine, and J. P. Rogers. Cascading spatiotemporal pattern discovery. *IEEE TKDE*, 24(11):1977–1992, 2012.

[35] T. Munzner. Chapter 5: Marks and channels. In *Visualization Analysis and Design*, page 102. CRC Press, 2014.

[36] K. G. Pillai, R. A. Angryk, and B. Aydin. A filter-and-refine approach to mine spatiotemporal co-occurrences. In *Proc. of ACM SIGSPATIAL*, pages 104–113. ACM, 2013.

[37] H. Qu, W.-Y. Chan, A. Xu, K.-L. Chung, K.-H. Lau, and P. Guo. Visual analysis of the air pollution problem in hong kong. *IEEE TVCG*, 13(6):1408–1415, 2007.

[38] R. Rastogi and K. Shim. Mining optimized association rules with categorical and numeric attributes. *IEEE TKDE*, 14(1):29–50, 2002.

[39] R. Scheepens, C. Hurter, H. van de Wetering, and J. J. van Wijk. Visualization, selection, and analysis of traffic flows. *IEEE TVCG*, 22(1):379–388, 2016.

[40] R. Scheepens, N. Willems, H. van de Wetering, G. Andrienko, N. Andrienko, and J. J. van Wijk. Composite density maps for multivariate trajectories. *IEEE TVCG*, 17(12):2518–2527, 2011.

[41] J. Stahnke, M. Dörk, B. Müller, and A. Thom. Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. *IEEE TVCG*, 22(1):629–638, 2016.

[42] G. Sun, R. Liang, H. Qu, and Y. Wu. Embedding spatiotemporal information into maps by route-zooming. *IEEE TVCG*, (5):1506–1519, 2017.

[43] G. Sun, R. Liang, F. Wu, and H. Qu. A web-based visual analytics system for real estate data. *Science China Information Sciences*, 56(5):1–13, 2013.

[44] C. Tominski, H. Schumann, G. Andrienko, and N. Andrienko. Stacking-based visualization of trajectory attribute data. *IEEE TVCG*, 18(12):2565–2574, 2012.

[45] T. von Landesberger, F. Brodkorb, P. Roskosch, N. Andrienko, G. Andrienko, and A. Kerren. MobilityGraphs: Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering. *IEEE TVCG*, 22(1):11–20, 2016.

[46] F. Wang, W. Chen, F. Wu, Y. Zhao, H. Hong, T. Gu, L. Wang, R. Liang, and H. Bao. A visual reasoning approach for data-driven transport assessment on urban roads. In *Proc. of IEEE VAST*, pages 103–112. IEEE, 2014.

[47] L. Wang, Y. Zheng, X. Xie, and W.-Y. Ma. A flexible spatio-temporal indexing scheme for large-scale GPS track retrieval. In *Proc. of MDM*, pages 1–8, 2008.

[48] Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. Van De Wetering. Visual traffic jam analysis based on trajectory data. *IEEE TVCG*, 19(12):2159–2168, 2013.

[49] G. I. Webb. Discovering associations with numeric variables. In *Proc. of ACM SIGKDD*, pages 383–388. ACM, 2001.

[50] N. Willems, H. van de Wetering, and J. J. van Wijk. Visualization of vessel movements. *CGF*, 28(3):959–966, 2009.

[51] P. C. Wong, P. Whitney, and J. Thomas. Visualizing association rules for text mining. In *Proc. of IEEE Symposium on Information Visualization*, pages 120–128, 1999.

[52] W. Wu, J. Xu, H. Zeng, Y. Zheng, H. Qu, B. Ni, M. Yuan, and L. M. Ni. TelCoVis: Visual exploration of co-occurrence in urban human mobility based on telco data. *IEEE TVCG*, 22(1):935–944, 2016.

[53] L. Yang. Visualizing frequent itemsets, association rules, and sequential patterns in parallel coordinates. In *Proc. of ICCSA*, pages 21–30, 2003.

[54] J. Zhang, E. Yanli, J. Ma, Y. Zhao, B. Xu, L. Sun, J. Chen, and X. Yuan. Visual analysis of public utility service problems in a metropolis. *IEEE TVCG*, 20(12):1843–1852, 2014.

[55] Z. Zhang and W. Wu. Composite spatiotemporal co-occurrence pattern mining. In *Proc. of WASA*, pages 454–465. Springer, 2008.

[56] Y. Zheng, W. Wu, H. Zeng, N. Cao, H. Qu, M. Yuan, J. Zeng, and L. M. Ni. Telcoflow: Visual exploration of collective behaviors based on telco data. In *Proc. of ICBD*, pages 843–852. IEEE, 2016.