

Article

# Identify Road Clusters with High-Frequency Crashes Using Spatial Data Mining Approach

Zhonggui Zhang <sup>1,2</sup>, Yi Ming <sup>3</sup> and Gangbing Song <sup>2,\*</sup> 

<sup>1</sup> School of Architecture and Materials Engineering, Hubei University of Education, Wuhan 430205, China; zzhang66@uh.edu

<sup>2</sup> Department of Mechanical Engineering, University of Houston, Houston, TX 77204, USA

<sup>3</sup> Department of Information System, Arizona State University, Tempe, AZ 85281, USA; yming2@asu.edu

\* Correspondence: gsong@uh.edu

Received: 1 November 2019; Accepted: 27 November 2019; Published: 4 December 2019



**Abstract:** This paper develops a three-step spatial data mining approach to directly identify road clusters with high-frequency crashes (RCHC). The first step, preprocessing, is to store the roads and crashes in a spatial database. The second step is to describe the conceptualization of road–road and crash–road spatial relationships. The spatial weight matrix of roads (SWMR) is constructed to describe the conceptualization of road–road spatial relationships. The conceptualization of crash–road spatial relationships is established using crash spatial aggregation algorithm. The third step, spatial data mining, is to identify RCHC using the cluster and outlier analysis (local Moran’s I index). This approach was validated using spatial data set including roads and road-related crashes (2008–2018) from Polk County, IOWA, U.S.A. The findings of this research show that the proposed approach is successful in identifying RCHC and road outliers.

**Keywords:** data mining; road network; traffic crash; road clusters with high-frequency crashes (RCHC); spatial weight matrix (SWM); local Moran’s I index; cluster identifying

## 1. Introduction

According to the World Health Organization, ~1.25 million people die each year on the roads as a result of crashes (traffic accidents) [1]. The road traffic network is an integrated and complex system consisting of four elements: “people, vehicle, roads, and environment” [2,3]. As a carrier of traffic, roads with ancillary facilities have an important impact on the frequency of crashes. From the perspective of transportation authorities and safety specialists, strategies such as renovating road facilities, improving road traffic conditions, and using prompt signs of crash warning at road clusters with high-frequency crashes (RCHC) are effective in reducing crashes. Therefore, given the massive roads, how to identify RCHC is one of the most significant challenges faced by transportation authorities and safety specialists.

A review of previous studies shows that data mining [4–6] has been widely used to traffic crash analysis. Kumar et al. [7] used the latent class clustering and k-modes clustering technique on road accident data from Haridwar, India. Castro and Kim [8] explored the role of different factors on injury risk using a Bayesian network, decision trees, and artificial neural networks to detect factors of the greatest influence on car accidents. Taamneh et al. [9] established a set of rules that can be used by the United Arab Emirates Traffic Agencies to identify the main factors that contribute to accident severity. Li et al. [10] applied statistics analysis and data mining algorithms on the fatal accident dataset as an attempt to discover variables that are closely related to fatal accidents.

The above studies focus on using data mining approach to obtain the relationships between non-spatial factors and traffic crashes, neglecting mining geospatial features associated with traffic crashes. Unlike non-spatial data mining, few studies have been dedicated to the spatial autocorrelation

measure [11], an important method of spatial data mining [12–14], to identify crash hotspot. Ouni and Belloumi [15] examined the stability of the performance of two spatial autocorrelation measures based on a road safety risk index through the comparison of the results in Tunisia. Xie and Yan [16] integrated network kernel density estimation with local Moran's I for hot spot detection of traffic accidents. Blazquez and Celis [17] identified critical areas with high child pedestrian crash risk in the city of Santiago, Chile, using kernel density estimation and Moran's I index in a GIS environment.

In the above spatial autocorrelation measures, including Moran's I [18], Geary's ratio [19], and Getis-Ord  $G_i^*$  [20], the Moran's I was most favored by researchers, as its distributional characteristics are more desirable and the indicator has greater general stability and flexibility [18,21]. The Moran's I, first known as a global single indicator of assessing spatial autocorrelation, can qualitatively detect whether the spatial distribution is dispersed, random, or clustered in the entire space with respect to their attribute values. In this context, it is important to note that the global Moran's I cannot quantitatively describe traffic crashes that is mainly concentrated on those roads.

Therefore, it is necessary to calculate the local Moran's I index [22] of each road, and perform clustering and outlier analysis to reveal RCHC. The clustering and outlier analysis method examines the local Moran's I index of individual road based on a comparison with the neighboring roads, which is as an effective method to identify RCHC.

In addition, some studies have used network kernel density estimation [23–25] as a spatial data mining method to compute spatial concentrations of point-based crashes in a road network. The spatial weight between crashes is used as a distance or spatial closeness of crashes along the road network. These studies take point-based crashes as the research object and used point-based spatial clustering analysis method, which is effective to detect hazardous locations by clusters of crashes.

The above studies provide a foundation for the research content of this paper. However, previous studies neglected some issues by using spatial data mining methods. First, these studies mainly focus on the point-based spatial clustering to find hotspots of crashes, and thus cannot directly identify RCHC. Second, some studies neglected the road–road or crash–road spatial relationships that affect accuracy of the result of spatial data mining.

To solve the issues, first, this paper focuses on the line-based spatial clustering method and takes linear roads as the research object, which can directly identify RCHC. Second, in this study, road–road and crash–road spatial relationships are applied in spatial data mining methods. Crashes are spatially aggregated as the attribute of the count of crashes of road (ACCR) by considering the road-crash geometric and attribute relationships. Then, a spatial weight matrix of roads (SWMR) [26–28] is established based on the road–road topological and geometric relationships. The ACCRs and SWMR are used as the input parameters in the cluster and outlier analysis (local Moran's I) to improve accuracy of the result of spatial data mining.

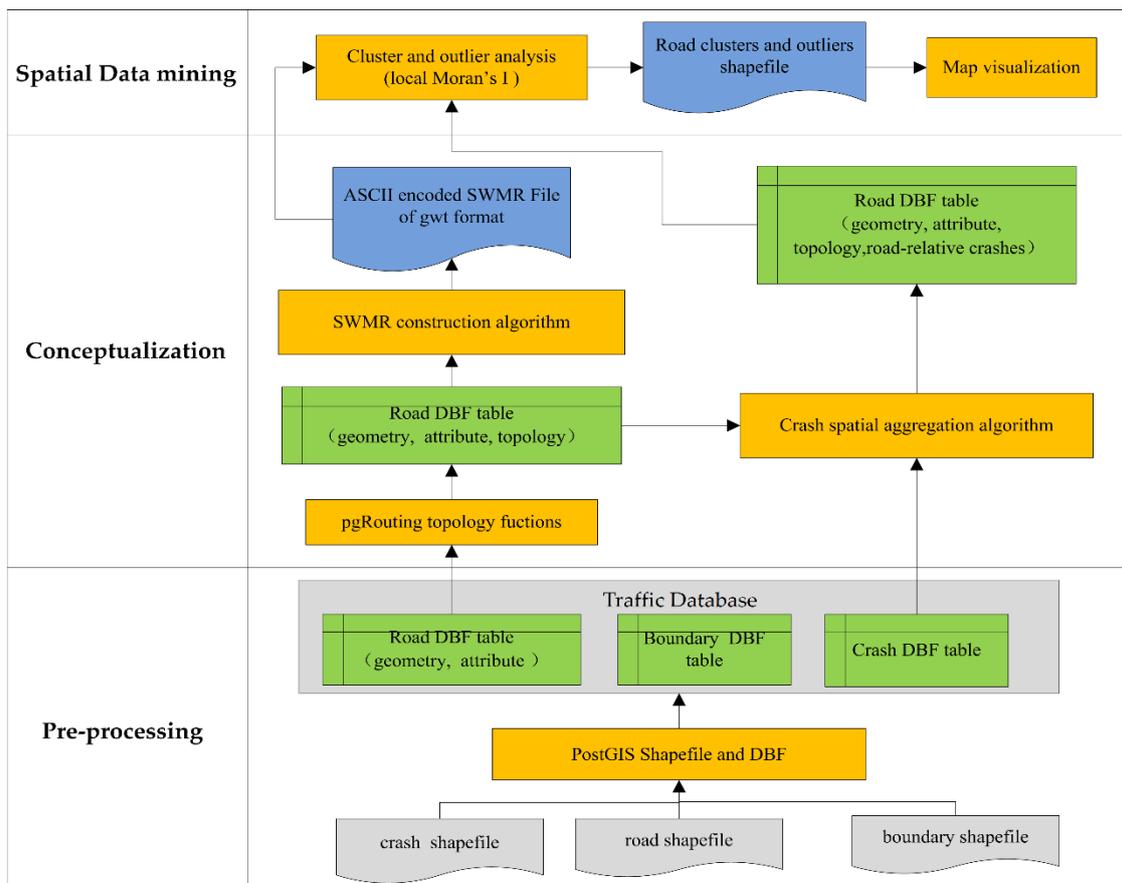
The aim of this study is to (a) create the accurate SWMR of complex road network respect to overpass crossing or underpass crossings to support further spatial statistics analysis (e.g., high–low clustering, hotspots analysis) (b) identify the RCHC to help transportation authorities and safety specialists to identify and prioritize roads that require more safety attention to reduce crashes.

The rest of the paper is organized as follows. Section 2 presents the methodology used in this study. Section 3 describes the spatial data, including the traffic crashes and roads of Polk County, Iowa used in this study. Section 4 illustrates and discusses the results by using the methodology within the study area. Section 5 recommends future work. Finally, Section 6 concludes the paper.

## 2. Methodology

### 2.1. Process Map

This paper presents a three-stepped spatial data mining approach to directly identifying RCHC by using: (1) preprocessing, (2) conceptualization, and (3) spatial data mining. The process map for the approach is shown in Figure 1.



**Figure 1.** Process map for the approach of directly identifying road clusters with high-frequency crashes (RCHC).

### (1) Preprocessing

The first step (preprocessing) is storage of the road network, crashes, and region boundary in spatial database. In this paper, we use PostgreSQL database [29] and PostGIS spatial data engine [30] to store and query massive spatial roads and crashes. The attribute and geometry info of road and crash table inherited from the input road and crash shapefile.

### (2) Conceptualization

The second step (conceptualization) is to build the conceptualization of spatial relationships. There are two types of conceptualization of spatial relationships: (1) the conceptualization of linear road relationships and (2) the conceptualization of road–crash relationships. Spatial weight matrix of roads (SWMR) is constructed to describe the conceptualization of road spatial relationships. The road network topology is created using pgRouting [31], which extends the PostGIS/PostgreSQL geospatial database to provide road network geospatial routing ability functions (e.g., `pgr_createTopology`, `pgr_createVertices`, `pgr_analyzeGraph`, `pgr_nodeNetwork`). The road network DBF table with count of crashes is calculated using crash spatial aggregation algorithm to describe the conceptualization of spatial relationships between roads and crashes.

### (3) Spatial data mining

The third step (data mining) is to identify RCHC using the cluster and outlier analysis (local Moran's I). According to studies of the point-based spatial clustering analysis method, this paper proposes a method of directly identifying RCHC by using line-based Moran's I index. We can

quantitatively identify RCHC (positive autocorrelation) and road outliers (negative autocorrelation) via map visualization of road clusters and outliers shapefile.

## 2.2. File Types

There are three types of file: (1) the input files, (2) the intermediate files, and (3) output results used in the approach, as shown in Table 1.

**Table 1.** The overview of file types.

Name	File Type	Format Type	Feature Type	Description
Road	Input	Shapefile	Line	The feature file of road centerlines
Crash	Input	Shapefile	Point	The feature file of crashes
Boundary	Input	Shapefile	Region	The Boundary of study Area
Traffic Database	Intermediate	PostgreSQL Database	Line/Point/Region	The spatial database converted from input shapefiles
SWMR file	Output	GWT File	/	ASCII encoded SWM File of road
Road clusters and outliers	Output	Shapefile	Line	The result of cluster and outlier analysis

### 2.2.1. The Input Files

Input files include point-based shapefile of crash, linear shapefile of road, and the regional shapefile of boundary. All the files contain the geometry and the important attributes (e.g., location, names, etc.).

### 2.2.2. The Intermediate Files

Intermediate files include PostgreSQL traffic database and SWMR file. We convert the input shapefiles to traffic database that contains the road, crash DBF table and boundary DBF table using the tool-“PostGIS 2.0 Shapefile and DBF”. The road DBF table contains the information: geometry, attribute, network topology, and road-relative crashes.

### 2.2.3. Output Results

Output results include the SWMR file and road clusters and outliers shapefile. The SWMR file is constructed based on geometric and topological adjacency of road network to describe the conceptualization of road spatial relationships as a foundation of spatial analysis. To improve the versatility of the approach, we use ASCII encoded gwt format file, which is compatible with spatial analysis software such as ArcGIS [32] and GeoDa [33], to store SWMR data. The first line of the file of the SWMR file in gwt format is the name of the unique identifier field (e.g., *ID*). After that, each row in the file is formatted into three columns: the *ID* of the road *i* ( $ID_i$ ), the *ID* of the road *j* ( $ID_j$ ), and the spatial weight ( $W_{ij}$ ). The road clusters and outliers shapefile contains the high–high and low–low road clusters, and the high–low and low–high outliers are the result of cluster and outlier analysis (local Moran’s *I* index). The high–high road clusters are RCHC we should identify in this research.

### 2.2.4. Remark on Availability of Input Files

The availability of input data determines the practicability of the approach. The input data can be easily accessed from Geofabrik’s free download server and state Department of Transportation noted in Section 3. However, the traffic crash shapefile, not being core input data, may be difficult to obtain in few cities or regions. As a result of taking roads (line type) as the core research objects instead of crashes (points) in this approach, csv format plain code file or excel xls file of crash have the basic information (e.g., crash id and location) also can be used to statistic the crash count of each road applying the attribute matching method in the proposed approach if we do not have the detailed crash shapefile in few cities or regions. Thus, the practicability of the approach is improved.

## 2.3. Methods

### 2.3.1. Crash Spatial Aggregation Algorithm

To find the RCHC, we should aggregate crashes in roads as the count of traffic crashes on the road. First, we add a fields (<crash\_count>, <crashlist >) of the road DBF table. Second, we calculate the attributes (<crash\_count>, <crashlist>) for each road. To determine whether the crash is on a road, the following two premises shall be considered.

- Premise 1 As the accuracy of crash global positioning system (GPS) coordinate has a positioning error of approximately 10 m [34]. A crash occurs on the road if its coordinates are within 10 m of the buffer of the linear road considering the positioning error of crash.
- Premise 2 As the Interstate Highway standards for the U.S. Interstate Highway System use a 12 foot (3.7 m) standard lane width [35], a crash occurring on the road can be determined if its coordinates are within 3.7 m of the buffer of the road.

Under above premises, we can determine the crash is on the road if the geometric and attribute relationships between crash and road meet both conditions 1 and 2 (means geometric and attribute integrated matching). If the crash cannot match any road by geometric and attribute integrated matching, the crash can be spatially aggregated to the road meet both conditions 3 and 4 (means spatial fuzzy matching).

1. Condition 1: The shortest distance between traffic crash and road less than 47 m (consider 10 lane roads and GPS positioning accuracy) order by the distance.
2. Condition 2: Attribute matching between name of road and location of crash.
3. Condition 3: The shortest distance between traffic crash and road is less than 10 m (consider GPS positioning accuracy).
4. Condition 4: The shortest distance between traffic crash and road is minimum in all datasets.

We use qt platform to realize the crash spatial aggregation algorithm in PostGIS/PostgreSQL database following the conditions. Figure 2 shows the definition of main data types. Figure 3 shows the main structure of the algorithm.

```

enum CRASH_TYPE{
Intersection_Related_Crash = 0,
Road_Related_Crash = 1,
};
typedef struct tagCrash Info{
int id; // unique identifier
QString geomWktInfo; // spatial info
QString literal; // location
QString systemstr;
enum CRASH_TYPE crash_type;
bool matchFlag12, matchFlag34;
}CrashInfo;

typedef struct tagRoadInfo{
int id; // unique identifier
QString name;
QString roadRef;
QString geomWktInfo; // spatial info
int stNode, endNode; // topological info
QString fclass;
int crash_count; // the count of crashes occur on the road
QList<QString> crashlist; // crashes occur on the road
short linkRoadcount;
bool bBridge, bTunnel, bTopoLink, bCrossLink;
}RoadInfo;

```

**Figure 2.** Definition of main data types.

```

QMap<int, RoadInfo> mRoadCache; QList<CrashInfo> mCrashCache;
1: Connect to PostGIS/PostgreSQL databases.
PGconn mConn = PQconnectdb("dbname=' dbname' host=ip port=5432 user='postgres' password='psw'
sslmode=disable connect_timeout=30");
2: Cache road DBF table to memory
SQLResultes =mConn.Execute("select id,st_astext(geom) as geom, name , ref, fclass ,bridge
,tunnel, source, target from road order by id");
for( each rowi to SQLResultes) mRoadCache.Set(rowi.id, rowi );
3: Cache crash DBF table to memory
SQLResultes =mConn.Execute("select id, st_astext(geom) as geom, literal, systemstr from crash order by
id");
for( each rowi to SQLResultes) mCrashCache.Set(rowi.id, rowi );
4: determine the crash is on the unique road if the spatial relationship between crashes and roads meet
both condition 1 and 2.
for (int i = 0; i < mCrashCache.count(); i++){
    SQLResultes =mConn.Execute(QStringLiteral ("select id from road where st_crosses(geom,
st_geomfromtext(st_astext(st_buffer(geography(st_geomfromtext('%1'),47)),4326)
or st_within(geom,st_geomfromtext(st_astext(st_buffer(geography(st_geomfromtext('%1'),47)),4326))
order by road.geom <-> ST_PointFromText('%1',4326)").arg(mCrashCache [i].geomWktInfo));
    // geometric matching
    for( each rowi to SQLResultes) {
        if( mRoadCache.find(rowi.id).name like mCrashCache[i]. literal ) {
            mCrashCache[i]. matchFlag12=true; mRoadCache.find(rowi.id).crash_count++;
            mRoadCache.find(rowi.id).crashIDs<< QStringLiteral("%1").arg(mCrashCache[i].id); break;}
    }
    // attribute matching
5: If the crash cannot matched to any road by geometric/attribute integrated matching, the crash can be
aggregated to the road meet both condition 3 and 4.
    for (int i = 0; i < mCrashCache.count()&&mCrashCache[i].matchFlag1=false; i++){
        SQLResultes =mConn.Execute(QStringLiteral ("select id from road where st_crosses(geom,
st_geomfromtext(st_astext(st_buffer(geography(st_geomfromtext('%1'),10)),4326)
or st_within(geom,st_geomfromtext(st_astext(st_buffer(geography(st_geomfromtext('%1'),10)),4326))
order by road.geom <-> ST_PointFromText('%1',4326) limit 1").arg(mCrashCache[i].geomWktInfo));
    }
    if( SQLResultes is not null) {
        mCrashCache[i]. matchFlag34=true; mRoadCache.find(row0.id).crash_count++;
        mRoadCache.find(row0.id).crashIDs<< QStringLiteral("%1").arg(mCrashCache[i].id);
    }
6: Save matching result to Road DBF table
    for (const_iterator it = mRoadCache.constBegin(); it != mRoadCache.constEnd();it++){
        mConn.Execute(QStringLiteral("update road set crash_count = %1 ,crashlist = '%2' where id = %3 ").
        arg(it.value().crash_count).arg(it.value().crashIDs.join(','))).arg(it.key());
    }
    mConn.close();

```

Figure 3. Structure of crash spatial aggregation algorithm.

### 2.3.2. SWMR Construction Algorithm

A spatial weight matrix (SWM) is a representation of the spatial structure of the data, and it is designed to generate, store, reuse, and share the conceptualization of the relationships among a set of features [36]. A SWM is the key input parameters in cluster and outlier analysis. The input SWM directly determines the correctness of the calculation results of cluster and outlier analysis. Consequently, when using an inappropriate SWM, cluster and outlier analysis cannot be trusted in general.

As taking linear roads as the research object in this paper, we need to build the SWMR that matches with spatial distribution characteristics of roads for identification of RCHC by using cluster

and outlier analysis. Conceptually, the SWMR is an  $N \times N$  matrix (as shown in Equation (1) [26,27]. There is one row for every road and one column for every road.

$$\text{SWMR} = \begin{matrix} & \begin{matrix} 1 & 2 & j & n \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ i \\ n \end{matrix} & \begin{matrix} W_{11} & W_{12} & W_{1j} & W_{1n} \\ W_{21} & W_{22} & W_{2j} & W_{2n} \\ W_{i1} & W_{i1} & W_{ij} & W_{in} \\ W_{n1} & W_{n2} & W_{nj} & W_{nn} \end{matrix} \end{matrix} \tag{1}$$

where  $N$  is the number of roads;  $i, j$  is the unique identifier of roads; and  $W_{ij}$  is the weight of matrix (means the cell value for any given row  $i$  and column  $j$  combination) that quantifies the spatial relationship between roads.

Typically,  $W_{ij}$  in SWMR are defined using Euclidean distance measurements and contiguity, fixed, or inverse distance weighting schemes [37,38]. However, road traffic system with crashes is based on road network which has complex topology and geometric relationships [39,40]. For identification of RCHC, defining spatial relationships in terms of road network is more appropriate.

SWMR models road spatial relationships and should follow the topology between roads that are restricted to the adjacency of road network. At the most basic level, there is a binary strategy for creating  $W_{ij}$  to quantify the spatial relationships among roads. If the road  $i$  is a 1st-order neighbor to road  $j$  in the road network, then  $W_{ij} = 1$ , else then  $W_{ij} = 0$ .

In this paper, we consider topological or geometric adjacency roads, which share the same intersection or the same node, equal to 1st-order neighbors. That means there is a topological or geometric adjacency between road  $i$  and road  $j$ , then  $W_{ij} = 1$ ; otherwise,  $W_{ij} = 0$  using binary strategy. However, to find the geometric or topological adjacency roads is a difficult problem since the different road types (e.g., highway, local, bridge, and tunnel) in real road network have complex topological and geometric relationships. Note that roads with an overpass crossing or underpass crossing lack a true intersection. For instance, Hul Ave and I235 highway have two intersections in the 2D map; however, they do have two overpass crossings (nonintersecting) in the photo, as shown in Figure 4.

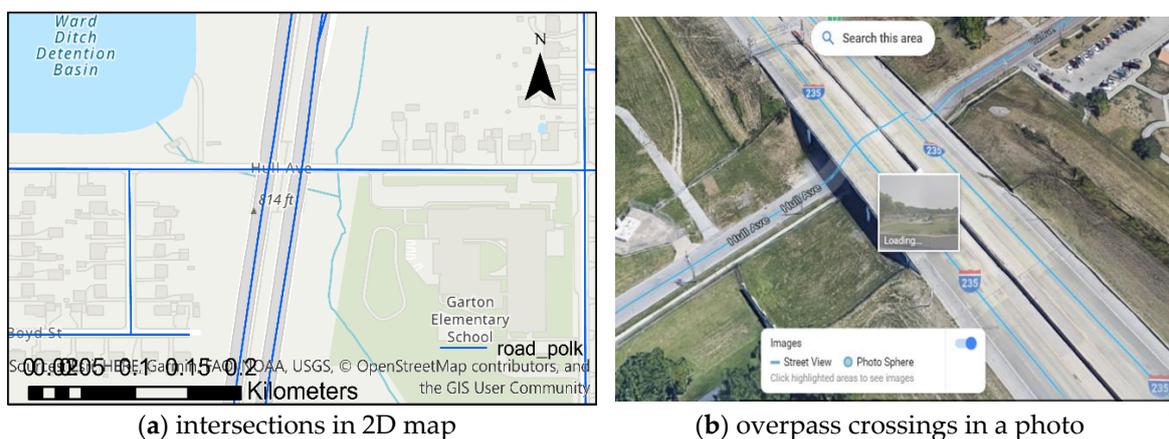


Figure 4. Overpass crossing (nonintersecting) of the highway and local road.

In GIS, there are different ways to model the topological and geometric relationship of road network in real world. A total of 11 kinds of graphs were summarized in this paper, as shown in Figure 5, to demonstrate the typical topological and geometric relationships of road network. Considering the typical topological and geometric relationships of road network, we can find the geometric or topological adjacency roads to calculate spatial weights of SWMR:

- The spatial weights of SWMR are 1 if the roads are topological adjacent for the following cases. (a) T-intersection with node, (c) T-intersection of highway bridge with node, (f) cross-intersection

with node, (j) topological adjacency, (k) topological adjacency with bridge, and (l) topological adjacency with tunnel.

- The spatial weights of SWMR are 1 if the roads are geometric adjacent for the following cases. (b) T-intersection without node, (d) T-intersection of highway bridge without node, and (e) cross-intersection without node.
- The spatial weights of SWMR are 0 if the roads are neither geometric adjacent nor non topological adjacent for the following cases; (g) overpass crossing and (h) underpass crossing.

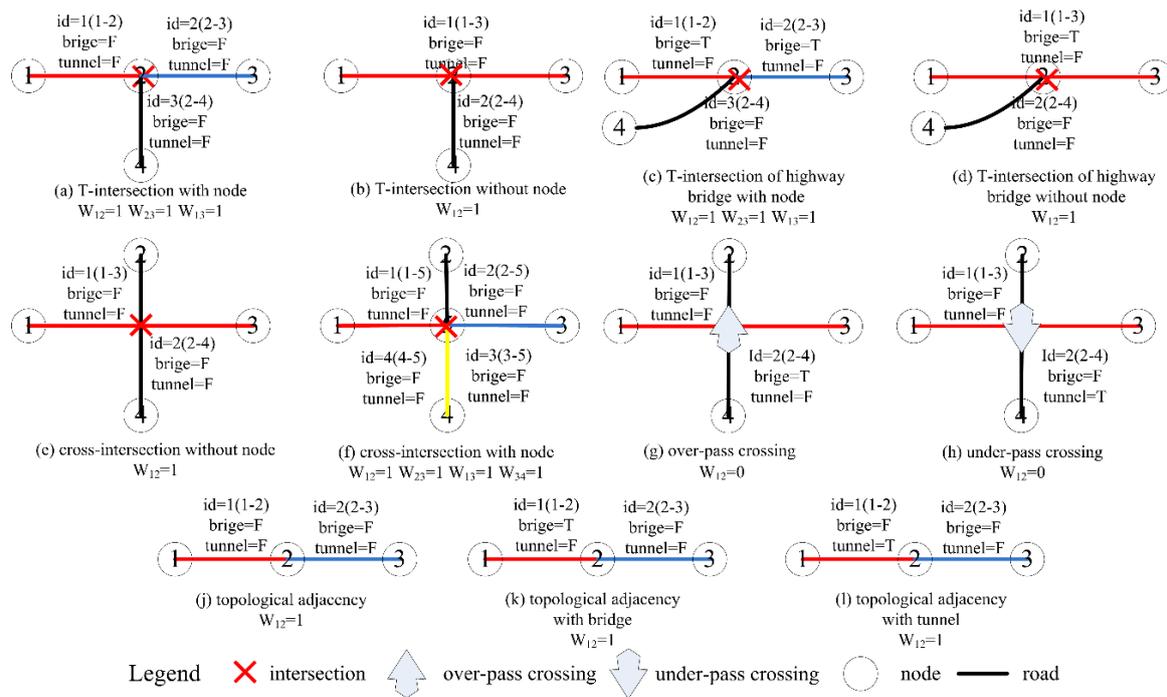


Figure 5. Typical topological and geometric relationships of road network.

Figure 6 shows the main structure of the SWMR construction algorithm.

### 2.3.3. Cluster and Outlier Analysis (Local Moran’s I)

The First Law of Geography [41], according to Professor Waldo Tobler, is “everything is related to everything else, but near things are more related than distant things.” Based on the first law of geography, geographical phenomena or attributes are related to each other in spatial distribution, have spatially related characteristics; that is, the closer the distance is, the more similar the things are [41]. Therefore, there are three types of distributions of crashes: dispersed, random, or clustered.

The global Moran’s I developed by Professor Moran in 1948, is one of the most preferred measure of spatial autocorrelation [42]. The global Moran’s I use a single index to detect the degree of autocorrelation of the same variable in the spatial region, and can verify the spatial distribution pattern in entire spatial extent.

In this study, the local Moran’s I (suggested in Professor Anselin based on global Moran’s I) [19] is used as a local indicator of spatial autocorrelation to find RCHC. The local Moran’s I, one of the most widely used local indicators of spatial association statistics [16], is calculated for each road to reveal the degree of spatial autocorrelation and is used to analyze whether the same variable (<crash\_count> in this research) has autocorrelation at a specific local location. The local Moran’s I index is expressed as [19,22,42]

$$I_i = \frac{X_i - \bar{X}}{S_i^2} \sum_{j=1, j \neq i}^n W_{ij}(X_j - \bar{X}) \tag{2}$$

where  $I_i$  is the calculated local Moran’s I index of Road  $i$ .  $I_i$  is a relative measure and can only be interpreted within the context of its computed  $z$ -score or  $p$ -value. The  $p$ -value is a probability, a type of statistics to express confidence level. The  $z$ -score is a standardized local Moran’s I index.  $W_{ij}$  is the weight in SWMR (as discussed in Section 2.3.2) that quantifies the spatial relationship between road  $i$  and road  $j$ .  $x_i, x_j$  is the crash count (as discussed in Section 2.3.1) of road  $i$  and road  $j$ .  $\bar{X}$  is the average crash count of all roads.  $n$  is the total number of roads,  $i = 1, 2, n$  and  $j = 1, 2, n$ .  $S_i^2$  is the measure of sample variance, defined as

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^n X_i^2 (X_j - \bar{X})^2}{n - 1} \tag{3}$$

```

QMap<int, RoadInfo> mRoadCache; QMap<int, QList<int>> mSWMRCacheGeom, mSWMRCacheTopo;
1: Connect to PostGIS/PostgreSQL database.
PGconn mConn = PQconnectdb("dbname='dbname' host=ip port=5432 user='postgres' password='psw'
sslmode=disable connect_timeout=30");
2: Cache road DBF table to memory
SQLResultes =mConn.Execute("select id, st_astext(geom) as geom, name , ref, fclass ,bridge
,tunnel, source, target from road order by id");
for( each rowi to SQLResultes) mRoadCache.Set(rowi.id, rowi );
3: Cache SWMR to memory
for (const_iterator it = mRoadCache.constBegin(); it != mRoadCache.constEnd(); it++){
    QList<int> roadIDsTopo = TopologicalAdjacency(it.value());
    mSWMRCacheTopo. insert(it.value().id, roadIDsTopo);
}
for (const_iterator it = mRoadCache.constBegin(); it != mRoadCache.constEnd(); it++){
    QList<int> roadIDsGeom ;
    RoadInfo Roadinfoi= it.value();
    SQLResultes =mConn.Execute(QStringLiteral("select id from road where ST_Intersects(
geom,ST_GeographyFromText('%1') order by id").arg(tomultilinestring(Roadinfoi.geomWktInfo)));
    for( each rowi to SQLResultes){ // 2D geometric adjacency
        if( collision of Roadinfoi.id and rowi.id not over-pass crossing or under-pass crossing )
            roadIDsGeom << rowi.id ;
    }
    mSWMRCacheGeom. insert(Roadinfoi.id, roadIDsGeoms);
}
QMap<int, QList<int>> mSWMRCache= mSWMRCacheTopo U mSWMRCacheGeom;
4: Save mSWMRCache to ASCII encoded gwt format file.
QFile file("file path and name"); QTextStream stream(&file); stream<<"id"<<endl;
for (const_iterator it = mSWMRCache[i].constBegin(); it != mSWMRCache[i].constEnd(); it++) {
    int roadID= it.key(); QList<int> roadIDs= it.value();
    for (int j = 0; j < roadIDs.count();j++)
        stream << QStringLiteral("%1 %2 1").arg(roadID).arg(roadIDs[j]) << endl;
}
file.close();
    
```

Figure 6. Structure of the spatial weight matrix of roads (SWMR) construction algorithm.

In the cluster and outlier analysis, it is necessary to make assumptions about the spatial distribution of crashes, which is randomization null hypothesis of spatial distribution. The test of randomization null hypothesis of spatial distribution can be performed based on the  $z$ -score and  $p$ -value along with the local Moran’s I index. The equation to calculate the  $z$ -score ( $Z_{I_i}$ ) for  $I_i$  is shown as

$$Z_{I_i} = \frac{I_i - E[I_i]}{\sqrt{V[I_i]}} \tag{4}$$

where  $E[I_i] = -\frac{\sum_{j=1, j \neq i}^n W_{ij}}{n-1}$ , and  $V[I_i] = E[I_i^2] - E[I_i]^2$ .

The  $Z_i$ -score is a standardized local Moran's I value of road  $i$ . The  $z$ -scores and  $p$ -values for roads are measures of statistical significance which tell us whether or not to reject the randomization null hypothesis, road by road [1]. In this study,  $p$ -value  $\leq 0.05$  (95% confidence level) is used to indicate significant clusters, which is applied to each road. For either road, if its  $p$ -value is smaller than 0.05 and  $z$ -score is greater than 1.96 or less than  $-1.96$ , that road will be considered as one of the cluster or outlier. The road who  $z$ -score is positive and greater than 1.96 and  $p$ -value is smaller than 0.05, with the neighboring roads have similar  $z$ -score and  $p$ -value, form the high-high road clusters. The high-high road clusters are the RCHC, which can be used to directly identify the dangerous roads.

### 3. Data Description

This study focuses on the Polk County, Iowa, United States. Based on the 2010 census, its population was 430,640, representing 14% of the state's residents, making it the Iowa's most populous county. The study considers all types of roads (e.g., local, highway, bridge, and tunnel) and crashes occur on road (not in intersection) within the Polk County boundary. The data can be downloaded from the website: <https://geodata.iowa.gov/dataset/county-boundaries-iowa>.

#### 3.1. The Spatial Data of Roads

The spatial data of IOWA statewide roads that this study employ can be download form OpenStreetMap [43] website (<http://download.geofabrik.de/north-america.html>). We extract spatial data of roads from the Iowa statewide road using (a) select layer by location tool (roads intersect with the Polk County boundary) and (b) select layer by attribute (roads suitable for cars) from ArcGIS geoprocessing tool box. A total of 27,606 roads are successfully recorded in ArcGIS software, as shown in Figure 7.

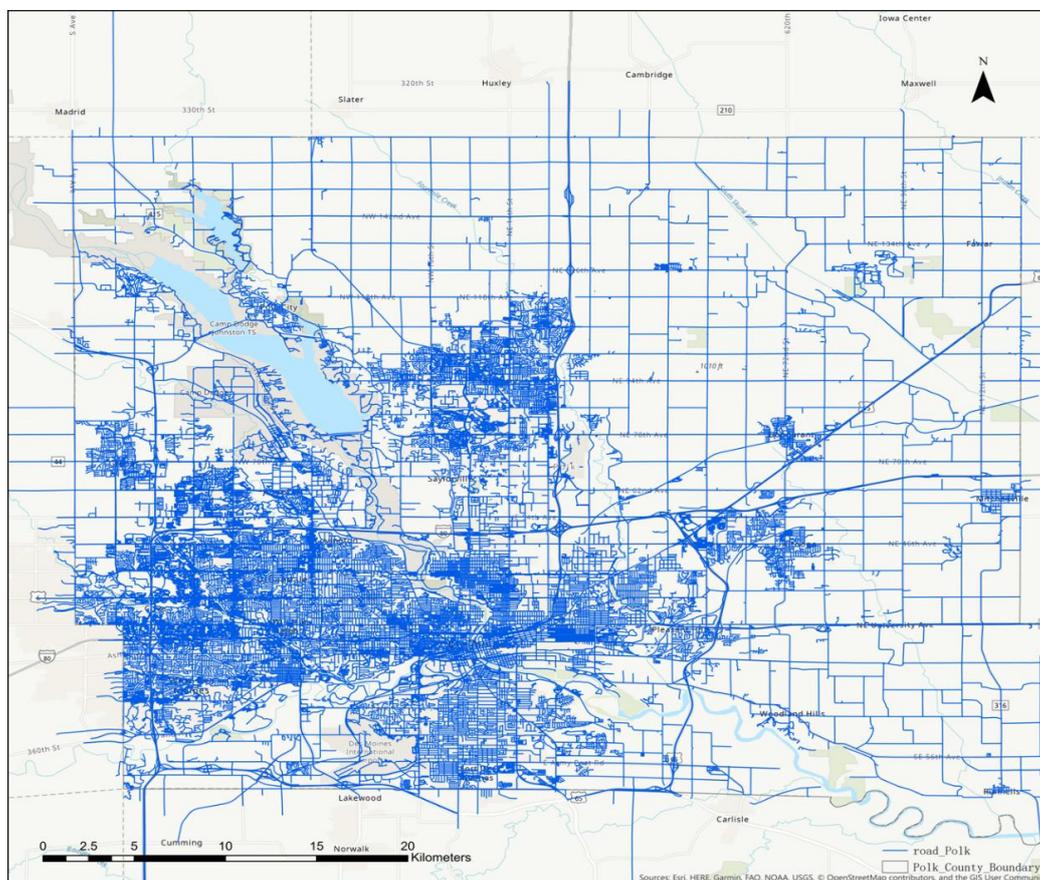


Figure 7. Distribution of roads in Polk County.

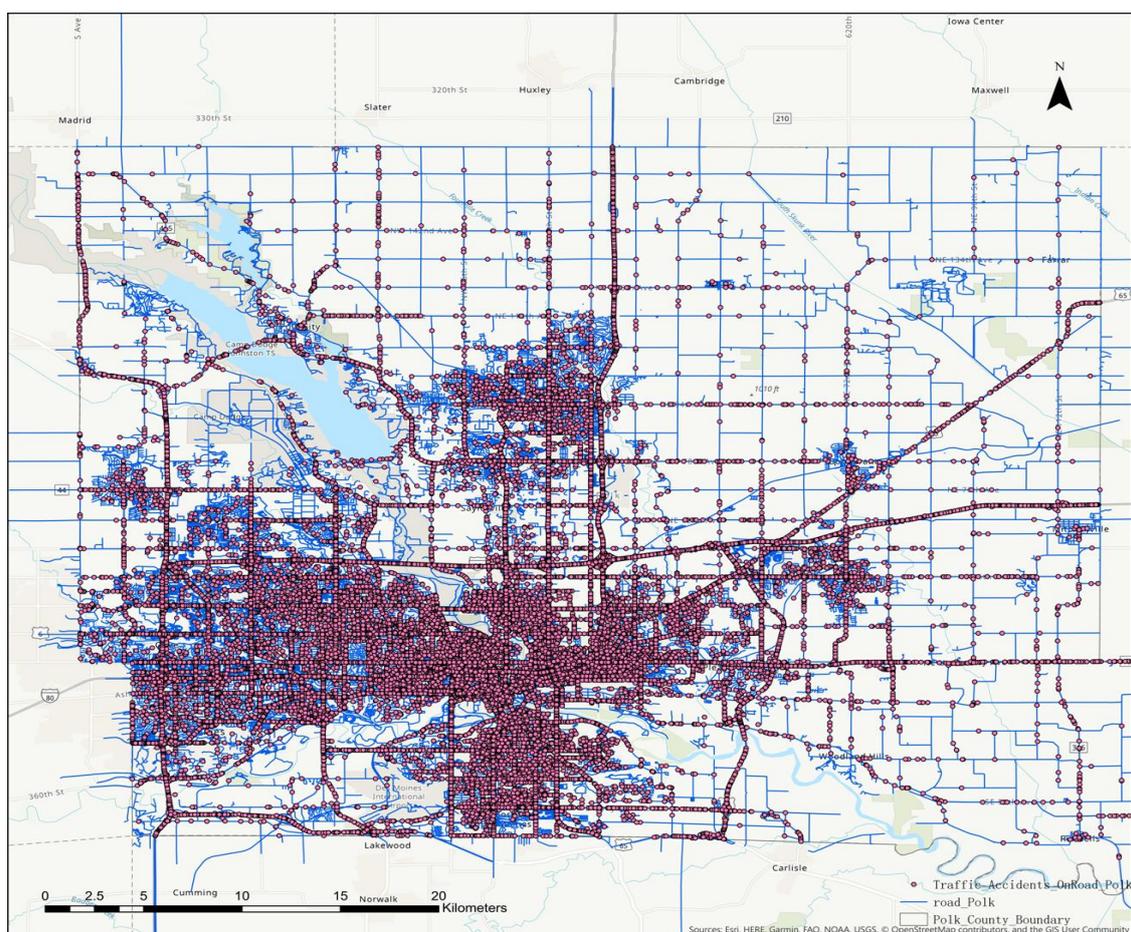
Note that the road shapefile of other cities or regions can be easily downloaded from Geofabrik's free download server, which has the latest global spatial road data normally update every day from the OpenStreetMap project.

### 3.2. The Spatial Data of Road-Related Crashes

This study employs the spatial data of crashes provided by the Iowa Department of Transportation's public platform (<https://data.iowadot.gov/>), which has the statewide data of general traffic crashes from the prior 10 years.

The database of crashes contains 49 types of information (e.g., crash\_key, casenumber, crash\_day, crash\_date, district, county\_num, literal, locfsthrm, locfsthrm, light, weather, rdtype, xcoord, and ycoord). For this study, a dataset of road-related crashes that occurred in Polk County was selected and analyzed.

We extract spatial data of crashes of the Polk County from the Iowa statewide traffic crash shapefile using (a) select layer by location tool (crashes intersect with the Polk County boundary) and (b) select layer by attribute tool (crashes are not intersection-related) from ArcGIS geo-processing tool box. A total of 41,734 road-related crashes that happened in Polk County from 1 January 2008–6 August 2018 are successfully recorded in ArcGIS, as shown in Figure 8.



**Figure 8.** Distribution of crashes in Polk County.

Note that we can also access other cities or regions crash shapefile published by state Department of Transportation in USA using the same extracting method.

### 4. Results and Discussion

#### 4.1. The SWMR of Polk County

As described earlier, SWMR is the critical input parameters for the local Moran’s I analysis. Therefore, it is necessary to establish the SWMR of the Polk County to accurately express the spatial relationship between roads.

As discussed in Section 2.2.3, we use ASCII encoded gwt format file to store SWMR data. The first line of the file of the SWMR in gwt format is the name of the unique identifier field (we use the field ‘ID’ as the unique identifier field in this paper). The spatial weight ( $W_{ij}$ ) is calculated by considering topological and geometric relationships between roads. Due to space limitations, we take the typical highway–local link road graph shown in Figure 9 as an example to demonstrate the SWMR of Polk County.

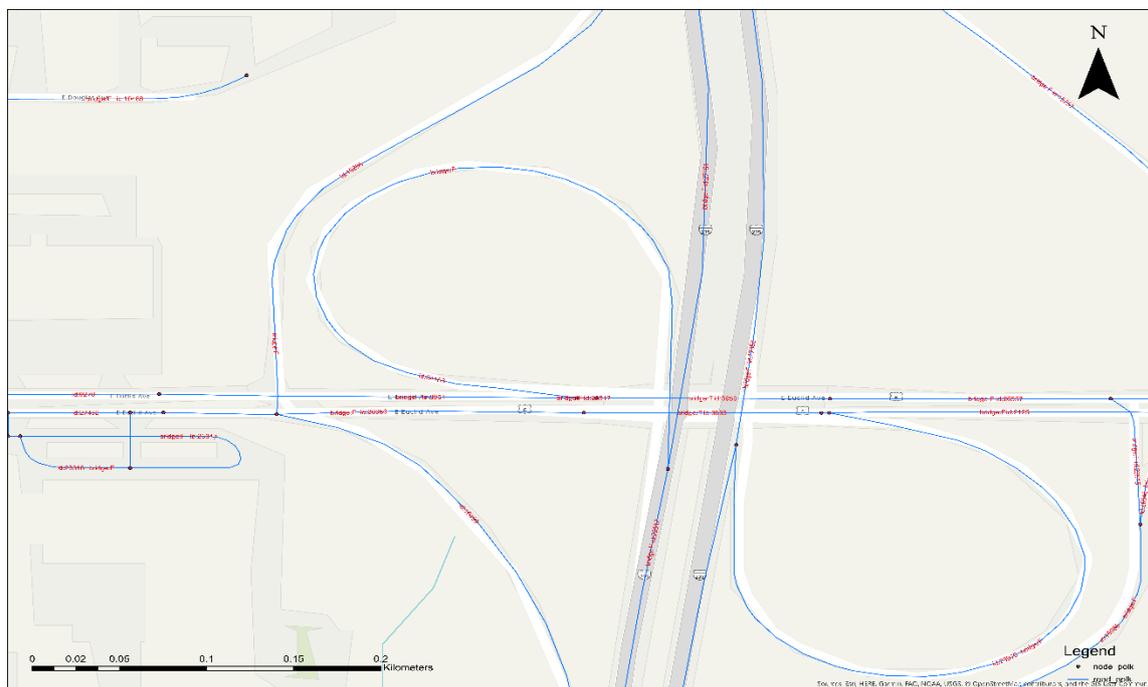


Figure 9. Typical highway–local link road graph.

Table 2 lists the content of the ASCII encoded gwt format file of roads, including motorway links (ID. 23970, 24103) and bridges (ID. 3833, 3950) created by the algorithm of constructing of SWMR (discussed in Section 2.3.2), to save the conceptualization data of road spatial relationships. It should be noted that the SWMR is a sparse matrix, and there is a large amount of zero  $W_{ij}$  data. Therefore, in this paper, the rows with a spatial weight of 0 are omitted since the default setting for spatial weights is 0 in the spatial data mining approach, which can effectively reduce the file storage space.

Table 2. The content of the ASCII encoded gwt format file.

First Row: ID (Unique Identifier Field)							
Row No.	IDS <sub>i</sub>	IDS <sub>j</sub>	$W_{ij}$	Row No.	IDS <sub>i</sub>	IDS <sub>j</sub>	$W_{ij}$
1	23970	2125	1	2	23970	17352	1
3	24103	3951	1	4	24103	23512	1
5	24103	23817	1	6	24103	27191	1
7	3833	2125	1	8	3833	26956	1
9	3950	23817	1	10	3950	26957	1

#### 4.2. The Results of Road Cluster and Outlier Analysis of Polk County

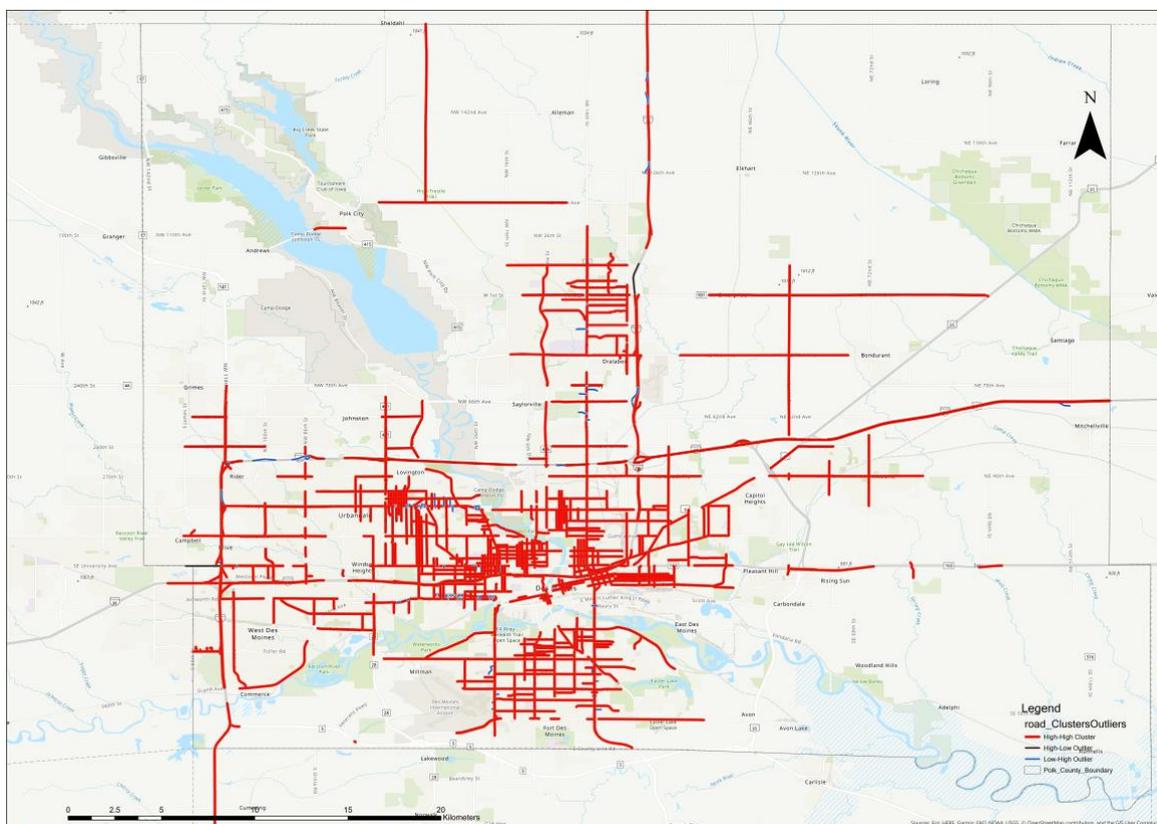
The study uses cluster and outlier analysis (the local Moran’s I, as discussed in Section 2.3.3) from geo-processing tools in ArcGIS [44,45], by taking the following input parameters, as shown in Table 3, to calculate local Moran’s I index, z-score, and p-value for each linear road to obtain road clusters and outliers across Polk County.

**Table 3.** Input Parameters of cluster and outlier analysis (Anselin local Moran’s I).

Input Parameter	Input Value
Input Feature Class	road_polk
Input Field	crash_count
Output Feature Class	c:\data\myproject.gdb\road_clustersoutliers
Conceptualization of Spatial Relationships	get_spatial_weights_from_file
Standardization	none
Distance Band or Threshold Distance	none
Weights Matrix File	c:\data\swmr.gwt
Apply False Discovery Rate Correction	no_fdr

All roads should have at least one neighbor [32] according to the best practice guidelines of cluster and outlier analysis. In this research, a total of 27,491 roads were selected in the input feature class since we find 115 roads in OpenStreetMap roads of Polk County have no neighbor when we construct the SWMR of Polk County.

In general, there are four types of road cluster and outlier in the road cluster and outlier shapefile, as discussed in Section 2.2.3: high–high cluster (cotype is HH), low–low cluster (cotype is LL), high–low outlier (cotype is HL), and low–high outlier (cotype is HL). The high–high cluster, high–low outlier, and low–high outlier are colored in red, black, and blue, respectively. The results, as shown in Figure 10 using map visualization, clearly demonstrate the road clusters and outliers.



**Figure 10.** Distribution of road clusters and outliers of Polk County.

High-high road cluster indicates that there is a positive autocorrelation. The roads in this cluster all have high crashes and the neighbor roads also have high-frequency crashes. That means, High-high road cluster is the RCHC that we should identify from the 27,606 roads of Polk County.

The RCHC of Polk County, as shown in Figure 10, centered along with I 35, I235, US69, I80, and US 6, can be discovered is relevant to hazardous roads detection. There are 738 roads in the RCHC of Polk County, which account for 2.67% of all roads and 24, 652 crashes occurred in RCHC, accounting for 59.07% of all crashes. That means 59.07% crashes occurred in 2.67% roads in Polk County. In addition, we can discovery that 85.60% crashes in RCHC occurred on major roads, including motorway, trunk, primary, secondary, and tertiary, as shown in Figure 11. We can quantitatively identify that 59.07% crashes of Polk County have strong positive spatial autocorrelation with topological or geometric adjacency of roads.

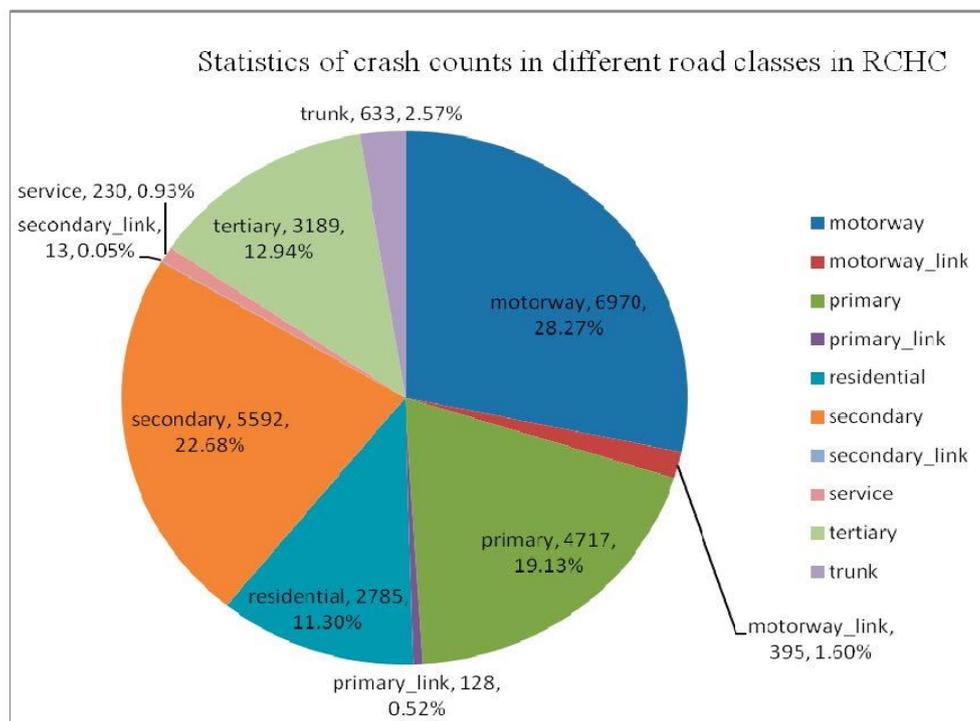


Figure 11. Statistics of crash counts in different road classes in RCHC.

Portion of calculated local Moran’s I, z-score, and p-value for each road in RCHC are shown in Table 4, based on which, transportation authorities can develop targeted mitigation strategies for the roads in RCHC to effectively reduce the number of crashes.

The high-low and low-high outlier shows that there is a negative autocorrelation. The roads in high-low outlier have high-frequency crashes; however, their neighboring roads have low-frequency crashes. Contrarily, the roads in low-high outlier have low-frequency crashes; however, their neighboring roads have high-frequency crashes. A portion of the calculated local Moran’s I, z-score, and p-value of each road in low-high and high-low outlier are shown in Table 5. Transportation authorities should also pay attention to roads in high-low and low-high outliers to find why there is a negative autocorrelation.

**Table 4.** Local Moran’s I, z-score, and p-value of roads in high–high cluster (RCHC).

<b>Id</b>	<b>Fclass</b>	<b>Name</b>	<b>Ref</b>	<b>Crash_Count</b>	<b>Local Moran I</b>	<b>z-Score</b>	<b>p-Value</b>	<b>Cotype</b>
24305	motorway		I 35	492	209.13	94.51	0.00	HH
17387	motorway		I 235	343	141.99	54.23	0.00	HH
24483	secondary	Fleur Drive		295	422.79	70.28	0.00	HH
26939	motorway		I 235	290	8.57	3.06	0.00	HH
26629	motorway		I 80; I 35	267	10.70	4.41	0.00	HH
27435	motorway		I 80; US 6	262	141.85	58.53	0.00	HH
26612	primary	South Ankeny Boulevard	US 69	252	242.99	59.57	0.00	HH
26533	motorway		I 80; I 35	250	141.82	82.74	0.00	HH
26381	primary	Douglas Avenue	US 6	246	751.06	90.18	0.00	HH
24779	secondary	University Avenue		241	1570.78	229.30	0.00	HH
26953	motorway		I 35	230	130.27	49.76	0.00	HH
26621	motorway		I 235	209	107.74	62.86	0.00	HH
26230	primary	Southeast 14th Street	US 69	200	455.77	88.67	0.00	HH
26884	motorway		I 80	199	91.34	46.15	0.00	HH
27436	tertiary	Ingersoll Avenue		199	39.29	4.36	0.00	HH
26954	motorway		I 35	198	39.23	14.99	0.00	HH
24841	motorway		I 35; I 80	193	7.77	3.93	0.00	HH
24971	motorway		I 80; I 35	188	160.12	72.36	0.00	HH
24800	primary	Southeast 14th Street	US 69	181	438.95	87.03	0.00	HH
24845	motorway		I 80	180	14.64	10.46	0.00	HH
26781	secondary	University Avenue		179	133.61	27.01	0.00	HH
24918	primary	Southeast 14th Street	US 69	177	132.56	50.64	0.00	HH
26558	motorway		I 80; I 35	172	7.82	4.56	0.00	HH
24782	primary	Northeast 14th Street	US 69	167	78.74	18.26	0.00	HH
25995	secondary	Martin Luther King Jr. Parkway		161	378.83	102.33	0.00	HH
24849	motorway		I 80; I 35	155	72.62	51.89	0.00	HH

**Table 5.** Local Moran’s I, z-score, and p-value of roads in high–low and high–low outliers.

<b>Id</b>	<b>Fclass</b>	<b>Name</b>	<b>Crash_Count</b>	<b>Local Moran i</b>	<b>z-Score</b>	<b>p-Value</b>	<b>Cotype</b>
1970	residential	Willowmere Drive	0	−4.91	−2.48	0.01	LH
23513	tertiary	Watrous Avenue	0	−5.11	−2.58	0.01	LH
24306	residential	Wakonda Drive	0	−4.89	−2.02	0.04	LH
2749	secondary	University Avenue	0	−4.60	−3.28	0.00	LH
15523	residential	Southwest 16th Street	0	−3.86	−2.25	0.02	LH
1670	residential	Southlawn Drive	0	−4.95	−2.89	0.00	LH
911	tertiary	Porter Avenue	0	−4.99	−2.52	0.01	LH
13993	residential	Northeast 69th Place	0	−2.58	−2.61	0.01	LH
14706	tertiary	Maurry Street	0	−4.46	−2.02	0.04	LH
23514	secondary	Indianola Avenue	0	−8.43	−3.48	0.00	LH
12238	residential	Hart Avenue	0	−5.87	−3.42	0.00	LH
10097	residential	Hackley Avenue	0	−4.86	−2.01	0.04	LH
23676	tertiary	East Watrous Avenue	0	−6.39	−3.23	0.00	LH
4286	tertiary	Cowles Drive	0	−4.49	−2.03	0.04	LH
4770	secondary	Bell Avenue	0	−4.63	−2.09	0.04	LH
8815	tertiary	Aurora Avenue	0	−5.08	−2.10	0.04	LH
14734	residential	41st Street	0	−3.98	−2.32	0.02	LH
1970	residential	Willowmere Drive	0	−4.91	−2.48	0.01	LH
23513	tertiary	Watrous Avenue	0	−5.11	−2.58	0.01	LH
24789	motorway	I 35	100	−3.59	−2.09	0.04	HL
27604	secondary	University Avenue	119	−19.65	−4.68	0.00	HL

### 5. Recommendation of Future Work

In this paper, we have developed a spatial data mining approach to directly identify road clusters with high-frequency crashes (RCHC) by using spatial weight matrix of roads (SWMR) and the local Moran’s I for cluster and outlier analysis. We believe that the proposed approach can be extended to the following fields, which can be considered as future work.

### 5.1. Spatiotemporal Data Mining Approach.

In this study, the ten-year crashes were equally applied in spatial data mining approach. However, different temporal crash factors, such as light, season, weather, may be varied. It is necessary to differentiate the treatment of crashes according to different temporal factors. As such, we will develop the spatiotemporal data mining approach considering the several spatiotemporal correlations between the crashes and main factors to discover the spatial and temporal patterns of traffic crashes according to different factors.

### 5.2. Identify Traffic Bottleneck.

In recent years, identify traffic patterns, including traffic bottlenecks, has received much attention [46–48]. The proposed approach in this paper has the potential to identify the traffic bottleneck. The spatial relationship between the bottleneck and broad network can be established by a fuzzy spatial aggregation algorithm, and the spatial weight matrix of road network are then used to study the degree of autocorrelation of the traffic congestion and discover the spatial distribution pattern of the traffic bottleneck under the constraints of the road network.

### 5.3. Identify Certain Roadway Damages.

Due to various adverse factors, such as pounding and impact [49], chloride diffusion [50] and corrosion [51,52], and freeze and thaw [53,54], roadways are subject damages, such as potholes and cracks, which will negatively impact traffic flow [55], generating abnormal traffic patterns, such as sudden slow down and lane changes, which will worsen over time. Therefore, a spatiotemporal data mining approach by using the historical data will be developed. Cluster and outlier analysis can be used to discover roadway damages by checking the degree of autocorrelation of traffic flow and the outliers represent with abnormal traffic flow.

### 5.4. Cloud-Based RCHC Identification

The proposed approach was developed and validated using a personal computer that has the capacity to identify RCHC in large cities or regions with populations of ~430,000 people, such as Polk County; however, it is difficult to process the big data of crashes and roads in megacities. As a future work, we will deploy, test, and amend the proposed approach in the cloud computing environment [56] to provide high performance computing solutions for identify RCHC in megacities, to further improve the data processing capability of this approach.

## 6. Conclusions

As important carriers of traffic, roads and their ancillary facilities have important impacts on the frequency of crashes [57,58]. This paper successfully demonstrates a spatial data mining approach to directly identify road clusters with high-frequency crashes (RCHC). The application of methodology was illustrated by using the spatial data set (stored in SHP file format) including traffic crashes (2008–2018) and roads of Polk County, Iowa, U.S.A. The proposed crash spatial aggregation algorithm uses geometric and attribute integrated matching and spatial fuzzy matching to build the crash–road spatial relationships considering GPS location accuracy. The developed spatial weight matrix of roads (SWMR) algorithm has the ability to detect and accommodate overpass crossing and underpass crossing with the consideration of the 11 typical topological and geometric relationships of roads. The algorithm, creates accurate SWMR of complex road network, have the added value that can support further spatial statistics (e.g., high–low clustering and Getis-Ord  $G_i^*$  analysis) of road network crashes. As a major contribution, the research adopts a new idea and focuses on line-based local Moran's I analysis by taking line-based roads as the core research objects instead of point-based crashes. As a result, the proposed method can directly identify RCHC.

**Author Contributions:** Z.Z. and G.S. developed the original idea. Z.Z. and Y.M. proposed the method. Z.Z. and Y.M. analyzed the data. Z.Z., G.S., and Y.M. wrote the paper.

**Funding:** This work was supported by the Construction Science and Technology Program of Hubei Province in 2016 (Transportation, municipal No. 02) and the Research Start-Up Funding of Hubei University of Education Grant NO. 18RC09.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Harirforoush, H.; Bellalite, L.; Bénié, G.B. Spatial and Temporal Analysis of Seasonal Traffic Accidents. *Am. J. Traffic Transp. Eng.* **2019**, *4*, 7–16. [[CrossRef](#)]
2. Pelaez, C.G.A.; Garcia, F.; de la Escalera, A.; Armingol, J.M. Driver Monitoring Based on Low-Cost 3-D Sensors. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 1855–1860. [[CrossRef](#)]
3. Carmona, J.; García, F.; Martín, D.; Escalera, A.; Armingol, J. Data Fusion for Driver Behaviour Analysis. *Sensors* **2015**, *15*, 25968–25991. [[CrossRef](#)] [[PubMed](#)]
4. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*; Elsevier/Morgan Kaufmann: Amsterdam, The Netherlands; Boston, MA, USA, 2012; ISBN 978-0-12-381480-7.
5. Kumar, S.; Toshniwal, D. A data mining framework to analyze road accident data. *J. Big Data* **2015**, *2*, 26. [[CrossRef](#)]
6. Chen, F.; Deng, P.; Wan, J.; Zhang, D.; Vasilakos, A.V.; Rong, X. Data Mining for the Internet of Things: Literature Review and Challenges. *Int. J. Distrib. Sens. Netw.* **2015**, *11*, 431047. [[CrossRef](#)]
7. Kumar, S.; Toshniwal, D.; Parida, M. A comparative analysis of heterogeneity in road accident data using data mining techniques. *Evol. Syst.* **2017**, *8*, 147–155. [[CrossRef](#)]
8. Castro, Y.; Kim, Y.J. Data mining on road safety: Factor assessment on vehicle accidents using classification models. *Int. J. Crashworthiness* **2016**, *21*, 104–111. [[CrossRef](#)]
9. Taamneh, M.; Alkheder, S.; Taamneh, S. Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. *J. Transp. Saf. Secur.* **2017**, *9*, 146–166. [[CrossRef](#)]
10. Li, L.; Shrestha, S.; Hu, G. Analysis of road traffic fatal accidents using data mining techniques. In Proceedings of the 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), London, UK, 7–9 June 2017; IEEE: London, UK, 2017; pp. 363–370. [[CrossRef](#)]
11. Soltani, A.; Askari, S. Exploring spatial autocorrelation of traffic crashes based on severity. *Injury* **2017**, *48*, 637–647. [[CrossRef](#)]
12. Shekhar, S.; Jiang, Z.; Ali, R.; Eftelioglu, E.; Tang, X.; Gunturi, V.; Zhou, X. Spatiotemporal Data Mining: A Computational Perspective. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 2306–2338. [[CrossRef](#)]
13. Zheng, Y. Trajectory Data Mining: An Overview. *ACM Trans. Intell. Syst. Technol.* **2015**, *6*, 29. [[CrossRef](#)]
14. Chen, W.; Panahi, M.; Pourghasemi, H.R. Performance evaluation of GIS-based new ensemble data mining techniques of adaptive neuro-fuzzy inference system (ANFIS) with genetic algorithm (GA), differential evolution (DE), and particle swarm optimization (PSO) for landslide spatial modelling. *Catena* **2017**, *157*, 310–324. [[CrossRef](#)]
15. Ouni, F.; Belloumi, M. Pattern of road traffic crash hot zones versus probable hot zones in Tunisia: A geospatial analysis. *Accid. Anal. Prev.* **2019**, *128*, 185–196. [[CrossRef](#)]
16. Xie, Z.; Yan, J. Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: An integrated approach. *J. Transp. Geogr.* **2013**, *31*, 64–71. [[CrossRef](#)]
17. Blazquez, C.A.; Celis, M.S. A spatial and temporal analysis of child pedestrian crashes in Santiago, Chile. *Accid. Anal. Prev.* **2013**, *50*, 304–311. [[CrossRef](#)]
18. Zhang, T.; Lin, G. On Moran's I coefficient under heterogeneity. *Comput. Stat. Data Anal.* **2016**, *95*, 83–94. [[CrossRef](#)]
19. Anselin, L. Local Indicators of Spatial Association-LISA. *Geogr. Anal.* **2010**, *27*, 93–115. [[CrossRef](#)]
20. Jana, M.; Sar, N. Modeling of hotspot detection using cluster outlier analysis and Getis-Ord  $G_i^*$  statistic of educational development in upper-primary level, India. *Model. Earth Syst. Environ.* **2016**, *2*, 60. [[CrossRef](#)]
21. Mitra, S. Spatial Autocorrelation and Bayesian Spatial Statistical Method for Analyzing Intersections Prone to Injury Crashes. *Transp. Res. Rec.* **2009**, *2136*, 92–100. [[CrossRef](#)]

22. Yuan, Y.; Cave, M.; Zhang, C. Using Local Moran's I to identify contamination hotspots of rare earth elements in urban soils of London. *Appl. Geochem.* **2018**, *88*, 167–178. [[CrossRef](#)]
23. Soltanolkotabi, M.; Candés, E.J. A geometric analysis of subspace clustering with outliers. *Ann. Stat.* **2012**, *40*, 2195–2238. [[CrossRef](#)]
24. Hashimoto, S.; Yoshiki, S.; Saeki, R.; Mimura, Y.; Ando, R.; Nanba, S. Development and application of traffic accident density estimation models using kernel density estimation. *J. Traffic Transp. Eng.* **2016**, *3*, 262–270. [[CrossRef](#)]
25. Anderson, T.K. Kernel density estimation and K-means clustering to profile road accident hotspots. *Accid. Anal. Prev.* **2009**, *41*, 359–364. [[CrossRef](#)] [[PubMed](#)]
26. Mawarni, M.; Machdi, I. Dynamic nearest neighbours for generating spatial weight matrix. In Proceedings of the 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Malang, Indonesia, 15–16 October 2016; IEEE: Malang, Indonesia, 2016; pp. 257–262. [[CrossRef](#)]
27. Qu, X.; Lee, L. Estimating a spatial autoregressive model with an endogenous spatial weight matrix. *J. Econom.* **2015**, *184*, 209–232. [[CrossRef](#)]
28. Ermagun, A.; Levinson, D. An introduction to the network weight matrix. *Geogr. Anal.* **2018**, *50*, 76–96. [[CrossRef](#)]
29. Obe, R.O.; Hsu, L.S. *PostgreSQL: Up and Running*, 2nd ed.; O'Reilly: Sebastopol, CA, USA, 2014; ISBN 978-1-4493-7319-1.
30. Bogorny, V.; Avancini, H.; de Paula, B.C.; Kuplich, C.R.; Alvares, L.O. Weka-STPM: A Software Architecture and Prototype for Semantic Trajectory Data Mining and Visualization. *Trans. GIS* **2011**, *15*, 227–248. [[CrossRef](#)]
31. Singh, P.S.; Lyngdoh, R.B.; Chutia, D.; Saikhom, V.; Kashyap, B.; Sudhakar, S. Dynamic shortest route finder using pgRouting for emergency management. *Appl. Geomat.* **2015**, *7*, 255–262. [[CrossRef](#)]
32. Scott, L.M.; Janikas, M.V. Spatial Statistics in ArcGIS. In *Handbook of Applied Spatial Analysis*; Fischer, M.M., Getis, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 27–41. ISBN 978-3-642-03646-0. [[CrossRef](#)]
33. Anselin, L.; Syabri, I.; Kho, Y. GeoDa: An Introduction to Spatial Data Analysis. *Geogr. Anal.* **2006**, *38*, 5–22. [[CrossRef](#)]
34. Wing, M.G.; Eklund, A.; Kellogg, L.D. Consumer-Grade Global Positioning System (GPS) Accuracy and Reliability. *J. For.* **2005**, *103*, 169–173. [[CrossRef](#)]
35. Khan, M.; Abdel-Rahim, A.; Williams, C.J. Potential crash reduction benefits of shoulder rumble strips in two-lane rural highways. *Accid. Anal. Prev.* **2015**, *75*, 35–42. [[CrossRef](#)]
36. Getis, A.; Aldstadt, J. Constructing the Spatial Weights Matrix Using a Local Statistic. *Geogr. Anal.* **2004**, *36*, 90–104. [[CrossRef](#)]
37. Seya, H.; Yamagata, Y.; Tsutsumi, M. Automatic selection of a spatial weight matrix in spatial econometrics: Application to a spatial hedonic approach. *Reg. Sci. Urban Econ.* **2013**, *43*, 429–444. [[CrossRef](#)]
38. Getis, A. Spatial interaction and spatial autocorrelation: A cross-product approach. *Environ. Plan. A Econ. Space* **1991**, *23*, 1269–1277. [[CrossRef](#)]
39. Liu, H.; Wang, J. Vulnerability assessment for cascading failure in the highway traffic system. *Sustainability* **2018**, *10*, 2333. [[CrossRef](#)]
40. Du, B.; Huang, R.; Chen, X.; Xie, Z.; Liang, Y.; Lv, W.; Ma, J. Active CTDaaS: A data service framework based on transparent IoD in city traffic. *IEEE Trans. Comput.* **2016**, *65*, 3524–3536. [[CrossRef](#)]
41. Tobler, W. On the First Law of Geography: A Reply. *Ann. Assoc. Am. Geogr.* **2004**, *94*, 304–310. [[CrossRef](#)]
42. Kumari, M.; Sarma, K.; Sharma, R. Using Moran's I and GIS to study the spatial pattern of land surface temperature in relation to land use/cover around a thermal power plant in Singrauli district, Madhya Pradesh, India. *Remote Sens. Appl. Soc. Environ.* **2019**, *15*, 100239. [[CrossRef](#)]
43. Haklay, M.; Weber, P. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Comput.* **2008**, *7*, 12–18. [[CrossRef](#)]
44. Aghajani, M.A.; Dezfoulian, R.S.; Arjroody, A.R.; Rezaei, M. Applying GIS to Identify the Spatial and Temporal Patterns of Road Accidents Using Spatial Statistics (case study: Ilam Province, Iran). *Transp. Res. Procedia* **2017**, *25*, 2126–2138. [[CrossRef](#)]
45. Estiri, H. Tracking Urban Sprawl: Applying Moran's I Technique in Developing Sprawl Detection Models. In Proceedings of the 43rd Annual Conference of the Environmental Design Research Association EDRA, Seattle, WA, USA, 30 May–2 June 2012; pp. 47–53.

46. Zheng, Z.; Ahn, S.; Chen, D.; Laval, J. Applications of wavelet transform for analysis of freeway traffic: Bottlenecks, transient traffic, and traffic oscillations. *Transp. Res. Part B Methodol.* **2011**, *45*, 372–384. [[CrossRef](#)]
47. Du, B.; Zhou, W.; Liu, C.; Cui, Y.; Xiong, H. Transit pattern detection using tensor factorization. *Inf. J. Comput.* **2019**, *31*, 193–206. [[CrossRef](#)]
48. Ma, X.; Luan, S.; Du, B.; Yu, B. Spatial copula model for imputing traffic flow data from remote microwave sensors. *Sensors* **2017**, *17*, 2160. [[CrossRef](#)] [[PubMed](#)]
49. Xu, B.; Song, G.; Mo, Y. Embedded piezoelectric lead-zirconate-titanate-based dynamic internal normal stress sensor for concrete under impact. *J. Intell. Mater. Syst. Struct.* **2017**, *28*, 2659–2674. [[CrossRef](#)]
50. Peng, J.; Hu, S.; Zhang, J.; Cai, C.S.; Li, L. Influence of cracks on chloride diffusivity in concrete: A five-phase mesoscale model approach. *Constr. Build. Mater.* **2019**, *197*, 587–596. [[CrossRef](#)]
51. Li, W.; Xu, C.; Ho, S.; Wang, B.; Song, G. Monitoring concrete deterioration due to reinforcement corrosion by integrating acoustic emission and fbg strain measurements. *Sensors* **2017**, *17*, 657. [[CrossRef](#)]
52. Peng, J.; Xiao, L.; Zhang, J.; Cai, C.S.; Wang, L. Flexural behavior of corroded HPS beams. *Eng. Struct.* **2019**, *195*, 274–287. [[CrossRef](#)]
53. Kong, Q.; Wang, R.; Song, G.; Yang, Z.J.; Still, B. Monitoring the soil freeze-thaw process using piezoceramic-based smart aggregate. *J. Cold Reg. Eng.* **2014**, *28*, 06014001. [[CrossRef](#)]
54. Wang, Y.; Tan, Y.; Guo, M.; Wang, X. Influence of Emulsified Asphalt on the Mechanical Property and Microstructure of Cement-Stabilized Gravel under Freezing and Thawing Cycle Conditions. *Materials* **2017**, *10*, 504. [[CrossRef](#)]
55. Mao, X.; Wang, J.; Yuan, C.; Yu, W.; Gan, J. A Dynamic Traffic Assignment Model for the Sustainability of Pavement Performance. *Sustainability* **2018**, *11*, 170. [[CrossRef](#)]
56. Du, B.; Huang, R.; Xie, Z.; Ma, J.; Lv, W. KID model-driven things-edge-cloud computing paradigm for traffic data as a service. *IEEE Netw.* **2018**, *32*, 34–41. [[CrossRef](#)]
57. García, F.; García, J.; Ponz, A.; de la Escalera, A.; Armingol, J.M. Context aided pedestrian detection for danger estimation based on laser scanner and computer vision. *Expert Syst. Appl.* **2014**, *41*, 6646–6661. [[CrossRef](#)]
58. García, F.; Jiménez, F.; Anaya, J.; Armingol, J.; Naranjo, J.; de la Escalera, A. Distributed pedestrian detection alerts based on data fusion with accurate localization. *Sensors* **2013**, *13*, 11687–11708. [[CrossRef](#)] [[PubMed](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).