

Identification and prioritization of accident-prone locations: A multi-criteria framework for analyzing traffic accidents in an urban environment

Fagner Sutel de Moura^{a*}, Lucas França Garcia^b, Tânia Batistela Torres^a, Leonardo Pestillo de Oliveira^b, Christine Tessele Nodari^a

^aIndustrial Engineering Graduate Program –Federal University of Rio Grande do Sul, Porto Alegre, Brazil

^bGraduate Program in Health Promotion – Cesumar University, Maringá, Brazil

Abstract

The identification of road traffic accidents (RTAs) and the prioritization of accident-prone locations is a consolidated practice in road safety analysis on highways; however, this type of approach requires improvements in the urban environment. This work proposes a framework for identifying and prioritizing accident-prone locations in urban areas through heuristic approaches. This framework adopts three heuristic approaches chained. The first approach consists of a clustering model through the Affinity Propagation Clustering (APC) method that seeks to generate units of analysis that best correspond to the spatial distribution of accidents. The second approach identifies candidates for APLs through the spatial association between neighboring UAs with a high frequency of RTAs provided by the Local Moran's Index. Finally, UAs are classified as APLs based on temporal dynamics by identifying Change Point Patterns (CPP) that describe different frequency distributions of RTAs over time. The results present the APC as a suitable approach for identifying the distribution pattern of RTAs and easy calibration; moreover, the spatial association metrics provided clusters of APL candidates. The CPP proved to be an efficient mechanism for identifying emerging APLs classified as APLs. This versatility provided an easy-to-apply framework that requires low-complexity information to execute.

Keywords: Accident-prone locations; urban; framework; heuristic; prioritization.

1. Introduction

Over time, displacement flows in cities change, the way people and goods do move advance, and the technological and social conditions transform traffic conditions. Although the mobility paradigm changes and evolves, an unwanted effect of this process is the occurrence of Road Traffic Accidents (RTAs). Considering that the patterns of the RTAs change, and the spatial distribution migrate according to the transformations of the environment, this problem requires monitoring actions and continuous intervention.

Such actions require instruments to identify accident distributions, monitor accident patterns, identify accident-prone locations (APLs), and prioritize critical situations. Identifying the distribution of accidents seeks to describe different spatial patterns of occurrences of RTAs. The prioritization of such locations elects them as candidates for mitigating actions and thus need accurate analysis for better application of resources. Although the treatment of APLs requires in any transport system, there is no consensus on the best way to group RTAs, identify, classify, and prioritize APLs (Moura et al., 2022).

The works related to identifying APLs to develop a pipeline that ranges from acquiring accident data to prioritizing critical locations. These studies present three limitations in some LPAs analysis steps. These limitations are discussed below, followed by new approaches proposed to address these issues.

First, many studies use arbitrary units such as sweep radii at intersections, fixed length of road sections in the accident grouping step (Moura et al., 2022). Although widely used, the segregation of the urban road network into homogeneous segments and fixed radius cannot properly represent the spatial distribution of phenomena associated with RTAs. These approaches present restrictions that condition the distribution of RTAs in equidimensional units of analysis that do not correspond to the natural distribution of events.

Second, although statistical tests are applied in the studies carried out, they are strongly related to the validation of the APL identification process and less associated with APL prioritization criteria (Moura et al., 2022). Statistical analysis restricted to this scope ignores the application of some criterion of statistical significance on decisions in the APLs prioritization stage.

Also, in the prioritization stage, there is still an incipient concern with the longitudinal assessment of the behavior of APLs elected as priorities (Moura et al., 2022). Thereby, works dedicated to evaluating APLs present characteristics of cross-sectional observational studies. The cross-sectional approach obstructs understanding the temporal dynamics of accidents and their relationship with APLs.

Therefore, this work proposes the adoption of three innovative approaches. The first approach consists of clustering accidents using a heuristic method to identify RTAs distribution. The second approach, which prioritizes APLs, consists of a longitudinal analysis through heuristic tools to identify RTA frequency patterns over time, and thus understand the historical dynamics of each APL. Furthermore finally, it proposes the application of confirmatory statistical tests over the results who elect the APLs as a priority.

By eliminating the arbitrariness of accident association to the topology of segments and intersections with equidimensional length, clusters better characterize RTAs distribution. Clusters represent areas of RTAs combined in segments, intersections, and transition areas, incorporating underlying arrangements capable of providing a more realistic representation of RTAs distribution (Moura and Nodari, 2022).

Longitudinal studies make it possible to monitor and describe the behavior of accidentality as a process. This strategy makes it possible to assess the growth or decline in the

number of accidents or severity of occurrences in the different locations chosen during the analysis. When mitigating measures adopt in APLs, the longitudinal analysis makes it possible to implement the after-before analysis. In this way, the longitudinal analysis can contribute to two stages. This contribution occurs initially in carrying out the diagnosis and later monitoring the effects of the adopted measures.

Considering the exposed issues about APLs analysis, this work proposes a new framework for APL identification delineated to deal with the urban context complexity. This model comprises the steps of grouping RTAs into Units of Analysis (UAs) by clustering techniques, identifying APLs, and prioritizing APLs using spatial associations and temporal variation criteria respectively. This work proposes the identification of RTAs distributions through the APC algorithm, identifying APLs with spatial associations through the Moran's Index, and measure temporal variation of statistical parameters of accident distributions using a heuristic tool called Change Point Pattern (CPP).

2. Methods

To develop a framework for grouping RTAs, identifying and prioritizing APLs, this work applies a pipeline of approaches presented in detail in Figure 1. This pipeline seeks to group accidents into clusters, identify critical locations, prioritize APLs and finally evaluate the elected APLs longitudinally. This process consists of a heuristic approach to grouping accidents, obtaining the degree of association of APLs, and prioritizing critical locations from criteria of spatial association and temporal distribution of RTAs. Finally, ANOVA and Mann-Whitney analysis do apply to confirm the longitudinal results.

Figure 1. Methodology Steps.



The proposed framework adopts the Affinity Propagation Clustering (APC) technique to group accidents into clusters defined by the spatial distribution pattern of RTAs. Once the RTA clusters obtain and their boundaries are defined, they are transformed into consolidated spatial polygons. The analysis of clusters as spatial UAs constitutes geostatistics that subsidizes second-order evaluation processes (Moura et al., 2022).

A database with 75102 RTAs registered between 2015 and 2020 in Porto Alegre, Brazil, was used to obtain the clusters. The RTAs data were submitted to the clustering process using the latitude-longitude of the RTAs. The dataset used has a large volume of records, which increases the computational cost of clustering. To reduce the computational cost, the APC application adopts the leveraged input approach of the *apcluster* package in the R language (Bodenhofer et al., 2011; Moura and Nodari, 2022). Leveraged affinity propagation reduces the dynamic and leveraged load for large datasets. For this process, the algorithm obtains a subset of data in multiple iterations, providing enough information about the grouping structure (Bodenhofer et al., 2011).

The APC requires a preference parameter as input to a vector containing the parameters applied in the model of partitioning data points into clusters. The leveraged learning function requires three parameters of the APC algorithm. The parameters are the input preference parameter (q); the fraction of sample used for optimized clustering ($frac$), and the number of leveraged clustering sweeps that change the subset of random samples used ($sweeps$) (Bodenhofer et al., 2011). These three parameters consist of the moderator of the quantile considered in the similarity matrix between data points, that keeps availability and accessibility updates moderate avoiding oscillations; the fraction of data at each iteration; and the number of scan iterations performed by the APC (Bodenhofer et al., 2011; Hennig et al., 2019). The parameters applied were $q=0.999$, $frac=0.1$, and $sweeps=5$, using 10% of the data points at each step for five iterations.

During this process, the clustering of RTAs provided UAs that constitute geographic areas that reproduce the different arrangements of RTAs distributed in space. After obtaining the UAs, the next step consisted of evaluating their distributions and comparing them with global values to identify unusual distribution patterns, and those places with denser distributions are considered APLs.

So that the distances in meters between UAs centroids could be obtained to evaluate the relationship between clusters, the data points do transform to the Mercator projection in WGS 84 coordinate system, EPSG:3857. The neighborhood matrix between UAs was calculated from the centroid distances, providing the distance in meters between centroids.

The normalized neighborhood relationships were calculated between UAs with Euclidean distances of up to 400 meters between centroids. This normalization simplifies the expression of Moran's Index to be calculated after. Such normalized relationships evaluate the weighted contribution of neighbors to each calculated UA. With the neighborhood matrix

calculated, Moran's global and local indexes were applied to obtain the relationships between UAs (Assunção and Reis, 1999).

The global Moran's index was applied to identify areas that represent different values than expected for a scenario of merely random occurrences. Local Moran's Index sought to measure the neighborhood relationship between UAs, describing how accidents in each location are associated with phenomena of the exact nature in the neighborhood.

Obtaining the neighborhood relations and identifying their spatial associations that present a local correlation significantly different from the rest of the data, the next step consisted of mapping regions that form conglomerates of UAs with a high occurrence of RTAs, characterizing APLs. This step performs by assigning the spatial associations between UAs to the Local Spatial Association Index (LISA) map from the Moran scatter diagram.

After mapping the relevant APLs with LISA, a CPP longitudinal analysis was applied (Gatrell et al., 1996). This analysis evaluated the change points in accident averages. Through this approach, the intervals between change points patterns provided homogeneous blocks about the average occurrence of RTAs.

The next step consists in obtaining the temporal variations in the average of accidents between the blocks. Subsequently, to describe the longitudinal behavior of the phenomenon, the accumulated variation over the period was calculated. The variations obtained were compared with the values expected globally. Those APLs with variation more critical than expected global variation were considered critical points.

Each block between CPPs constitutes a time interval with a stationary average of RTAs occurrences. The composite variation of averages informs whether, over time, this parameter is stationary or has some tendency to increase or decrease. The CPP assessment is applied globally and to each APL. The global and local analysis makes it possible to identify the global trend and compare it with the behavior of each UA.

Thus, when a negative variation in the global average is identified, those LPAs with a stationary average, increasing or with a downward trend below the global trend, are considered critical APLs because they do not follow the global downward trend. On the other hand, when the global variation between blocks is stationary, those UAs with increasing variation are elected as APLs. Finally, if the global variation shows an increasing trend, those UAs with more accentuated growth than the global trend are elected as APLs. Briefly, this comparison elects as APLs those UAs where the difference between the local and global variation is positive, as shown below.

$$\text{APL Election} = \begin{cases} \text{Var}(UA_i) - \text{Var}(\text{Global}) > 0 & \rightarrow \text{Elected.} \\ \text{Var}(UA_i) - \text{Var}(\text{Global}) \leq 0 & \neq \text{Not elected} \end{cases}$$

Finally, after obtaining the blocks of means for each cluster, a confirmatory analysis ANOVA and Mann-Whitney tests were applied to validate these results: one parametric and the other non-parametric. The confirmatory analysis consisted of comparing the first and last blocks in terms of means and medians.

As the CPP assumes a normal distribution, ANOVA compared means of the first and last blocks. As the blocks were not significant in the Shapiro-Wilk normality and Levene's homogeneity tests, the blocks were submitted to the Mann-Whitney median comparison. With these confirmatory tests, the tests sought to identify whether there is no difference in the initial and final means and medians for series with stationary mean, and if in the series with a tendency to increase or decrease, the means and medians differ.

For sites with a stationary mean, the ANOVA test confirms if the means of the first and last block are not different. By the Mann-Whitney test, locations with stationary mean should provide the statistic that confirms that the medians do not differ among the first and last blocks. For sites with a non-stationary trend, the comparison aims to identify whether the mean and median differ between blocks.

3. Background

The APC technique provides, through similarity/dissimilarity parameters, accident partitions that represent different distributions that are consolidated in spatial blocks, representing areas with distinct occurrence of RTAs (Moura and Nodari, 2022; Moura et al., 2022). These spatial blocks are transformed into spatial polygons that are later subjected to spatial analysis.

After obtaining objects with spatial representation, they can be subjected to spatial association analysis. first analysis to be applied consists of the neighborhood matrix. This matrix informs when a certain variable varies as a function of another, indicating neighborhood and dependence relationships. This association measure provides the notion of proximity as a measure of autocorrelation. This same pattern can also be weighted by the number of neighbors that each element has (Luzardo et al., 2017).

The Moran index provides spatial correlation metrics contained in the interval [-1, 1]. This interval presents the degree of association between the neighborhood elements. In this range, relationships that tend to -1 show inverse spatial correlation between neighbors. In the opposite direction, relationships between elements tending to 1 demonstrate a direct spatial association of the phenomenon of interest between neighbors. Finally, when the Global Moran Index is around zero, it indicates that there is no correlation between neighbors, featuring stochastic arrangements.

Under different significance levels, the Moran index allows determining when the data have a purely random distribution and when the data consist of different clusters with specific distribution patterns. The Moran index is widely used because it provides a z-score statistic that indicates the degree of significance between the identified associations. The Moran index is defined as:

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n W_{ij}^* (z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2}$$

where n is the number of regions, z_i contains the value of the attribute considered in element i , $\bar{z}$ the average value of the attribute in the region, and W_{ij}^* the elements of the spatial proximity matrix. The observed values z_i formally assume a normal distribution, which implies that the global Moran index assumes a normal distribution (Assunção and Reis, 1999; Fischer Wang, Jinfang., Springer Science+Business Media., 2011).

As Moran's global statistic provides a single significance value to attest to clusters with different distributions of the phenomenon of interest in the population, this statistic provides a global statistic for the entire area. For greater detail and obtaining information about local distribution patterns, the Local Moran's Index is applied together with the LISA map. This map provides local statistics and visualizations of the areas with occurrences of the phenomenon of interest, presenting its correlation with the neighborhood when attending significance levels established.

The higher the number of UAs, different association regimes can emerge, indicating sites with a strong association, weak association, or even stochastic distribution. Local Moran's Index allows investigating clusters characterized by neighborhood relations with characteristics in common, statistically associated.

The use of Local Moran's Index allows dealing with these different association regimes and pointing out the associative nature of each region through spatial autocorrelation of each element with its close neighbors. Moran's local index is defined as:

$$I_i = (z_i - \bar{z}) \sum_{j \in J_i}^n W_{ij}^* (z_j - \bar{z})^2$$

so that z_i provides the value of the attribute of interest in the region i , $\sum_{ij}^*$ provides the relationship of the elements of the spatial proximity matrix, J_i represents the neighborhood cells of area i , the sum gives the total of neighboring areas belonging to J_i and $\bar{z}$ the average of the neighboring areas J_i (Getis and Ord, 1992).

Based on the assumption that there is no spatial correlation, values neighboring each location exchange in their vicinity to calculate Local Moran's Index. When this hypothesis is rejected, the Local Moran's Index indicates the existence of conglomerates with distribution patterns different from the rest of the population. Such sites are identified as conglomerates that may require greater attention.

After these calculations, the scatter diagram obtained from Local Moran's statistics provides the spatial relationships distributed in four quadrants of the diagram. Such quadrants provide measures of spatial dependence between neighbors. The quadrants provide neighborhood relationships characterized as (H-H) High-High, (L-L) Low-Low, (H-L) High-Low, and (L-H) Low-High, of a parameter of interest contained into UAs.

When elements are in the H-H quadrant, the diagram shows that such UAs present high values of the parameter of interest, associated with neighbors also with high values of the parameter, forming locations with a high frequency of the parameter considered, characterized as LPAs. The opposite happens with elements contained in the L-L quadrant. This quadrant indicates that the UAs contained in its present low values of the parameter of interest, associated with neighbors that also present low values for the parameter, characterizing locations with a rare occurrence of the phenomenon of interest.

On the other hand, the H-L and L-H quadrants present, respectively, relations of UAs with high values of the parameter of interest associated with neighbors with low values of the parameter of interest, and relations of points with low values of the parameter of interest associated with neighbors with high values of the parameter of interest. The H-L and L-H relationships present scenarios in which the spatial context is characterized by randomness in distributing the parameter of interest among neighbors (Anselin, 1995).

Local indicators of spatial autocorrelation when contained in confidence interval established and confirmed by p-values of the local Moran's index, are associated to the criteria in quadrants of the scatterplot to generate a LISA map (Anselin, 1995). When the LISA map is obtained using Moran's local data, the data are sorted according to the quadrants of the Moran scatterplot. These data are inserted into the map when they meet the stipulated statistical significance criterion. Once the priority locations have been identified and mapped, the next challenge is to understand the temporal dynamics of the occurrence of RTAs in each cluster (Nie et al., 2015).

As time-series usually deal with long time paths, some applications have limited formulations to deal with short time series. This limitation leads to the option of replacing analyzes with the decomposition of trend, and periodicity by analysis of the statistical distribution of cohesive blocks. Such blocks have homogeneous parameter values that provide different probability distributions (Barry and Hartigan, 1993; Chen, 2019).

As a result, adopting approaches that evaluate time series in terms of CPP has grown. The use of CPP has wide application in industrial processes, as its nature consists of providing statistical summaries of a monitored process over time. Applying a CPP provides a more obvious way of identifying random fluctuations and imbalances in a process over time. (Chen, 2019).

To treat APLs as temporal processes constituted as cohesive blocks, CPP analysis is applied in the process analysis of intra-cluster RTA occurrences. This approach evaluates finite sequences of i.i.d. observations that present distinct partitions, identifying changes in statistical properties such as mean, variance, or distribution of the phenomenon over time. (Gallagher et al., 2013; Gurevich, 2006; Haynes et al., 2014; Killick et al., 2011). In this way, the CPP algorithm calculates the optimal placement and potentially the number of change points in the data distribution from different methods (Killick et al., 2011). The purpose of such approaches is to identify structural changes in the data indicating some change in the considered scenario (Dorcas Wambui, 2015).

When performing the disambiguation of time series, such analyses use non-parametric methods as a single point of change as At Most One Change method (AMOC), Binary Segmentation, and Segment Neighborhood. Among the applied parametric tests, approaches such as optimal partitioning and Pruned Exact Linear Time (PELT) stand out. Although they have their specificities, the different approaches consist of identifying the data points where the series change their statistical properties (Gurevich, 2006; Killick et al., 2011).

These approaches adopt a total cost function to be minimized. Therefore, the cost is obtained by the sum of the cost associated with the adjustment of an analyzed statistical property, such as:

$$C_t = \sum_{i=1}^{m+1} [C(y_{(\tau_{i-1}): \tau_i})] + \beta\gamma(m)$$

so that m gives the number of shift points; $[C(Y_{(t_{i-1}): \tau_1})]$ defines the cost function for each segment; $\beta\gamma(m)$ the penalty factor and C_t the total cost (Haynes et al., 2014). Assuming normal distribution with mean μ and variance σ^2 , the detection of change points in the mean statistical property uses the maximization of the logarithmic likelihood with respect to the mean. In this case the cost of the segment is obtained as twice the maximum negative likelihood:

$$[C(y_{(\tau_{i-1}): \tau_i})] = (t - s) \log(\sigma^2) \cdot \frac{1}{\sigma^2} \sum_{j=s+1}^t \left(y_j - \frac{1}{t - s} \sum_{i=s+1}^t y_i \right)^2$$

Changepoints are considered shocks that change the temporal pattern of a phenomenon. In this way, the nonparametric method AMOC searches for a single most significant shock, signaling to the point that separates a random block from a second stationary block on the time series (Gallagher et al., 2013; Gurevich, 2006; Shi, 2020).

The Binary Segmentation method is the most applied approach according to the specialized literature (Killick et al., 2011; Shi, 2020). By dynamic programming, this approach applies a single change point test statistic to all data through hierarchical segmentation applying AMOC. When a change point does locate, the two sections segmented from that point are re-evaluated for new change points. The segmentation and evaluation process is performed recursively until the data distribution stabilizes and no more change points are found in the partitioned segments. (Shi, 2020).

The segment neighborhood method uses a non-parametric framework to identify change points. It is characterized as an exact algorithm, but it requires considerable computational complexity (Chen, 2019; Dorcas Wambui, 2015). The PELT method is similar to the segment neighborhood algorithm in that it provides exact segmentation. However, due to its implementation model, it can be more efficient in computational terms, due to the use of dynamic methods used and applied pruning model (Dorcas Wambui, 2015; Killick et al., 2011).

The optimal partitioning PELT algorithm through dynamic programming optimizes each step based on the memory of previous optimal solutions. As each previous time point acts as a candidate for changepoint at the current instant, the optimization is obtained at each time t :

$$F(t) = \min_{0 \leq \tau < t} [F(\tau) + C(y_{\tau+1:t}) + \beta]$$

With initial value of $F(0)=0$ and with the proper choice of $\beta, F(t)$, all occurrences of t are calculated from that $F(n)$ gives the optimal cost and stores values of τ which minimize the above equation and the changepoints are obtained.

As in change point analysis processes, the algorithm is executed using the above equation about $t = 1:n$ seeking to minimize $\tau = 0:(t - 1)$. To make the process more efficient PELT prunes the number of candidates (Dorcas Wambui, 2015). For this purpose, pruning is performed at the current time when the current cost is less than the cost at candidate points plus the additional score of the segment. Thus, if at the current instant for:

$$F(\tau') + C(y_{\tau'+1:t}) < F(t)$$

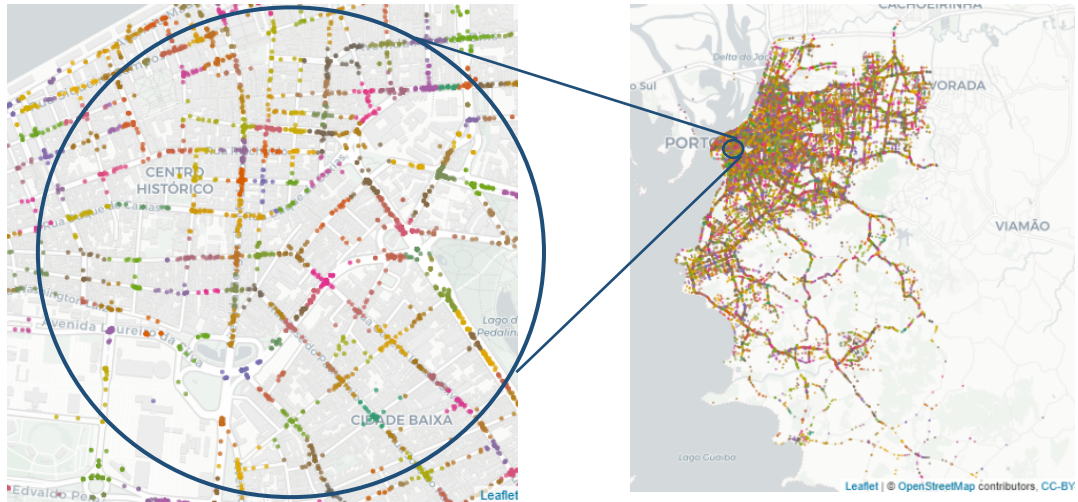
where $\tau' \leq (t - 1)$, τ' can be removed from the following steps, reducing the number of candidates. In this way, the PELT when executed with the true number of change points, being linearly distributed, works satisfactorily when compared to other methods. PELT has a computational cost of approximately $O(n)$, being more efficient when compared to other models with the same purpose (Killick et al., 2011). Given its characteristics, the PELT method has been adopted for producing consistent and visually identifiable results in the data series, in addition to reducing the computational cost (Dorcas Wambui, 2015).

4. Results

During the first step of the proposed model, the APC algorithm discovered 3639 clusters of RTAs partitioned according to the underlying spatial distribution pattern. The spatial distribution pattern from spatial coordinates provided clusters that describe similar distribution patterns of RTAs. The Figure 2 shows the different clusters found, and the colors are randomly applied to contrast the different clusters generated. This grouping pattern that considers the

most significant features to identify similarities provided arrangements that delimit the ‘pool strength’ of each location. The pool strength describes the influence imposed by neighboring locations on the UA under analysis (Aguero-Valverde and Jovanis, 2008; Cheng et al., 2018; Huang et al., 2017; Lee et al., 2019).

Figure 2. Clusters distribution in Porto Alegre, Brazil.



After the clusters were identified, they were transformed into a Spatial Polygons Data Frame (SPDFs) object. SPDFs objects provide GIS data that consolidates spatial polygon information (Moura and Nodari, 2022). These objects allow different information contained within the boundaries of the clusters to be treated as spatially associated data. In the SPDFs structure, cluster data such as types, outcome, date, location, density, date, and severity index of the RTAs associated with each cluster were stored. This approach allows working with usual information from RTAs without requiring more detailed data or exposure factors such as AADT.

As a result, the 3639 clusters provided UAs, of which 99% of them contained up to 101 RTAs. After the clusters are generated and consolidated as Spatial UAs, they were submitted to spatial analysis to identify the neighborhood relationships between the different items.

The normalized neighborhood matrix provided proximity relationships between centroids up to 318 meters apart, with this distance corresponding to the average distance between intersections, when considering the ten intersections closest to each other. The weighted influence of neighbors on each UA were obtained using a normalized neighborhood matrix.

After calculating the neighborhood relations, the Global Moran's Index, using the normality test, confirmed the existence of clusters with the parameters: Moran's I statistic standard deviate = 14,498, and p-value < 2.2e-16 presented in detail in Table 1.

Table 1. Global Moran's Index.

Moran I test under normality		
weights: nb2listw(dnb, style = "W")		
Moran I statistic sd = 14.498, p-value < 0.000000000000000022		
alternative hypothesis: greater		
sample estimates:		
Moran I statistic	Expectation	Variance
0.1548239001	-0.0003921569	0.0001146223

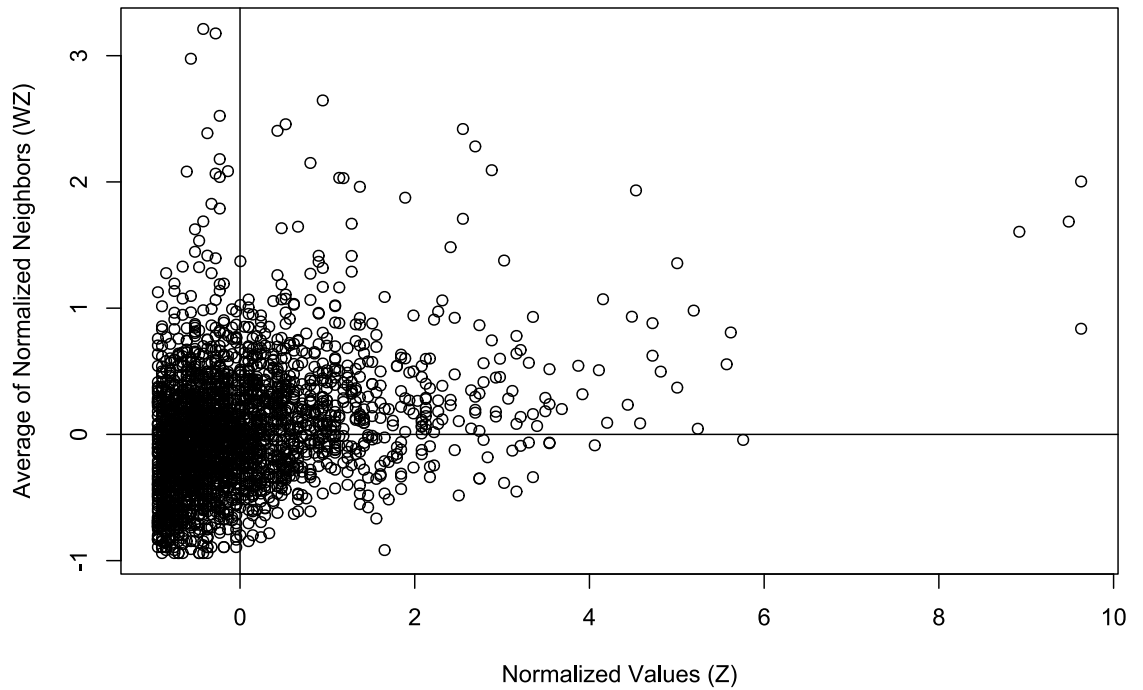
The local Moran index provided the neighborhood relationship pattern for each cluster Table 2. The results of Local Moran's Index gave rise to the scatterplot. This graph made it possible to compare the relationship between normalized values of RTAs and the normalized mean of the neighbors identified by the neighborhood matrix.

Table 2. Local Moran Index (first six results).

	Ii	E.Ii	Var.Ii	Z.Ii	Pr(z != E(Ii))
1	0.76574385	-0.00336747335	0.493546528	1.0947757	0.2736149100
2	0.33676286	-0.00028834075	0.072265975	1.2538020	0.2099139952
3	-0.05439107	-0.00035554355	0.074193068	-0.1983799	0.8427478220
4	-0.15637082	-0.00022817117	0.095467694	-0.5053512	0.6133121581
5	-0.10609446	-0.00002295739	0.007199831	-1.2500794	0.2112705585
6	0.31585368	-0.00003049868	0.006947920	3.7896619	0.0001508525

According to the arrangement of data in the scatter plot presented in Figure 3, 576 clusters were allocated in quadrant one, indicating H-H relationships. In quadrant four, 983 clusters with L-H relationships do allocate. Regarding quadrant two, 357 clusters with L-L relationships do identify. Finally, in quadrant three, 635 clusters with H-L relationship were allocated.

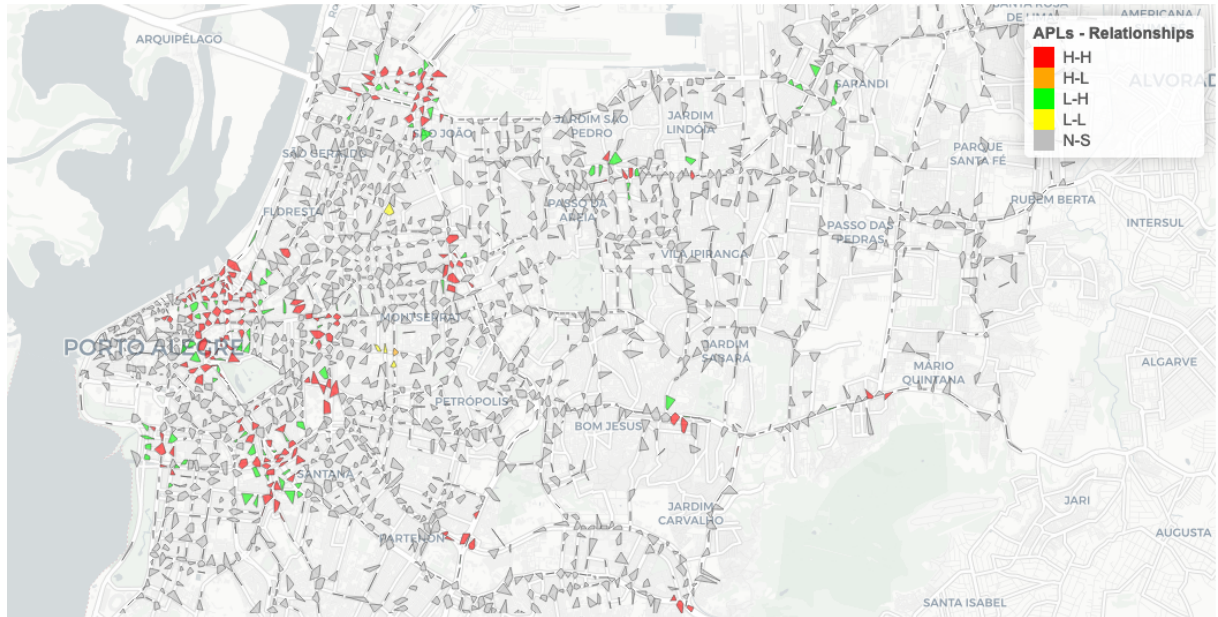
Figure 3. Local Moran scatter plot.



The local Moran's index data combined in the scatterplot when contained within the 95% confidence interval provided the map of the neighborhood relationships identified for each UA and the association relationships between neighbors. This map provides the local indicators of spatial autocorrelation (LISA). In this way, the different regions with statistically significant spatial correlations are presented in the spatial association map, indicating the existence of conglomerates with correlated UAs.

This map presented in Figure 4, makes it possible to identify regions with UAs with a high frequency of RTAs associated with neighboring units with the same parameter behavior and therefore identified with a High-High (H-H) ratio. In this way, the different regions with statistically significant spatial correlations are presented in the spatial association map, indicating the existence of conglomerates with correlated UAs that have a high frequency of RTAs. On the other hand, UAs presented as L-L correspond to regions with a low frequency of RTAs associated with neighbors that present the same behavior. Finally, regions with H-L or L-H relationships form regions that intersperse areas with high and low frequency patterns, indicating randomness in the neighborhood relationship process, and therefore not characterized as pool strength areas. Sites identified as N-S represent non-statistically significant neighborhood relationships.

Figure 4. Clusters with Neighborhood Relations at 95% confidence interval.

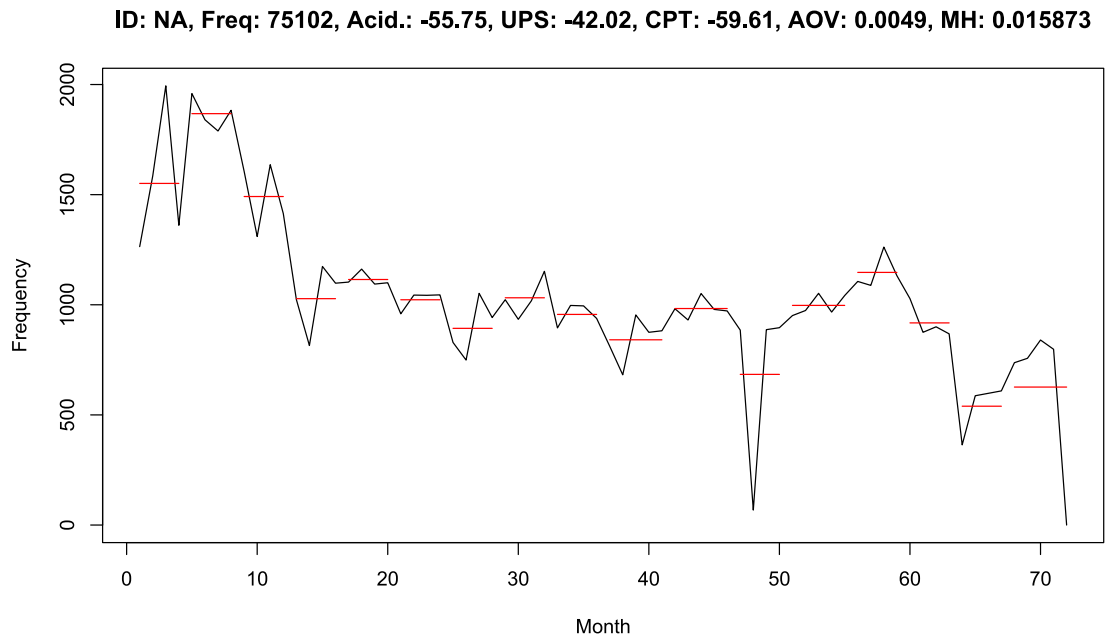


After identifying the neighborhood relations and selecting clusters with a high frequency of RTAs associated with neighbors and a high number of occurrences of RTAs, these locations classify as APLs. APLs with such characteristics totaled 188 UAs (4.3%) from a universe of 3639 clusters. Once the APLs identify, the CPP provides the temporal distribution of the RTAs over the six years. Initially, the CPPs on the global dynamics of RTAs are identified to obtain the global pattern of variation in the average of accidents between blocks.

As shown in Figure 5, seventeen blocks were identified between CPPs during the global analysis. Each block represents a stationary mean interval, and the composite variation of the averages showed a reduction of 59.6%, indicating a strong downward trend in the average of RTAs over the period evaluated.

After obtaining global variation, it was calculated from each AU. The difference between the variation of each AU concerning the global variation was obtained. Those UAs with a positive difference maintained the classification as APLs because they did not follow the global reduction trend.

Figure 5. Global dynamics of the time series of RTAs.

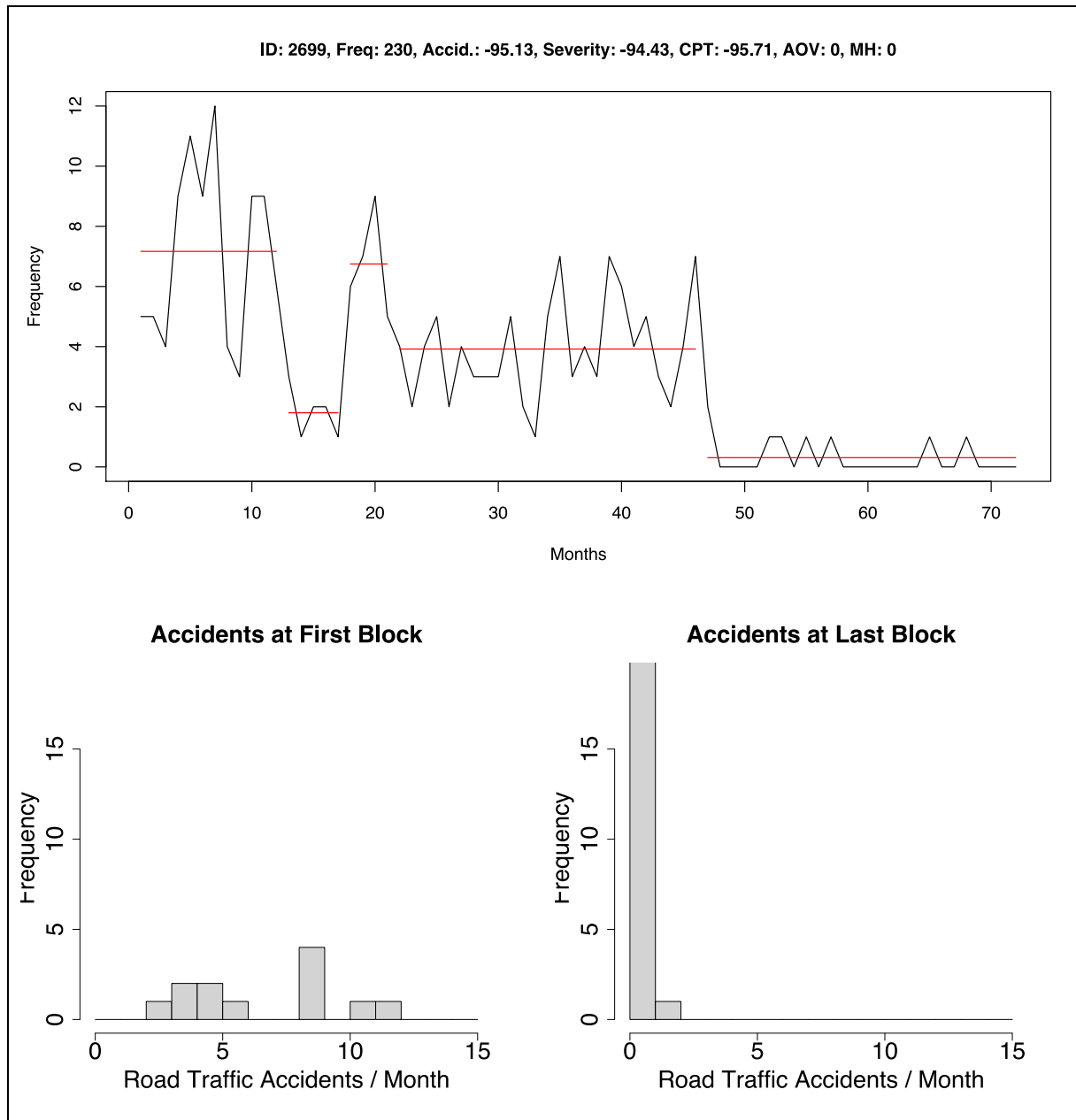


When the CPP tests is applied to the different clusters, series patterns with stationary and variable average are found. Among the 157 clusters, the CPP found 121 clusters with a stationary mean, and 36 with a varying mean, 33 with decreasing variation and 3 with increasing variation.

ANOVA and Mann-Whitney tests validated these findings. In stationary series, it is expected that there is no significant difference between the means and median of the initial and final blocks. During this work, when working with a series of stationary averages is expected that the averages and medians do not show differences between the initial and final blocks. The results of the block-average tests and the p-values of the ANOVA and Mann-Whitney test are exemplified in Figure 6, with the respective frequency distributions in the time series and in the first and last blocks. In this example, an APL has five blocks with stationary averages and a composite variation of -95%; this reduction implies a temporal dynamic that describes the first block with a monthly frequency of RTAS between 2 and 12, while in the last block the frequency varies between zero and one RTA per month.

This figure presents the graphs of a UA, showing a negative variation in the average of RTAs between blocks. As can be seen, for the non-stationary AU the time series represents the decay and the histograms present the monthly frequencies of RTAs in the first and last blocks, showing the differences to be evaluated by the ANOVA and Mann-Whitney tests. In turn, a stationary UA presents an AU with a stationary trend, which gives very similar frequency distributions between the blocks.

Figure 6. Stationary and non-stationary mean series.



When observing the set of stationary clusters, the ANOVA analysis confirmed that in 79% of the UAs, the means are not different, and the Mann-Whitney test confirmed that in 72% of the UAs, the medians do not differ. The same tests were applied to the set of non-stationary clusters. For these cases, the ANOVA analysis confirmed that the means are different in 92% of the UAs, and the Mann-Whitney test confirmed that the medians differ in 92% of the UAs.

Based on these results, contingency matrix statistics were applied to measure accuracy, sensitivity, and specificity. When cross tests were applied, the stationarity and non-stationarity results compared with the ANOVA results provided an accuracy of 84%, a sensitivity of 100%,

and specificity of 59%. When cross-table tests were applied to the Mann-Whitney median results, an accuracy of 98%, the sensitivity of 98%, and specificity of 100% were verified Table 3. Although the CPP assumes a normal distribution, the non-parametric confirmatory test proved to be more adherent, which is justified by the fact that the distribution of RTAs follows a Poisson distribution and the rejection of the hypothesis of homogeneity and normality by the Levene and Shapiro-Wilk tests respectively.

When the CPP was applied, 157 UAs were identified with APLs. After applying the confirmatory ANOVA and Mann-Whitney tests, the screening selected 134 UAs as APLs. With these results from the universe of 3639, 134 (3.6%) were characterized APLs.

Table 3. Confusion Matrix and Statistics oa ANOVA and Mann-Whitney.

Matrix and Statistics of ANOVA	Matrix and Statistics of Mann-Whitney
Actual Predicted 0 1 0 38 0 1 33 117 Accuracy: 0.8245 95% CI: (0.7624, 0.876) P-Value: 0.000000001295 Sensitivity: 0.5352 Specificity: 1.0000	Actual Predicted 0 1 0 35 3 1 0 150 Accuracy: 0.984 95% CI: (0.9541, 0.9967) P-Value: 0.0000000000002122 Sensitivity: 1.0000 Specificity: 0.9804

5. Discussion

The proposed roadmap presented a set of heuristic models to identify APLs. Initially, the model of clusters of RTAs as UAs uses distributions identified with the APC method. Subsequently, the UAs were submitted to a spatial analysis process to identify UAs with a high frequency of RTAs associated with neighbors with the same characteristic. Finally, the analysis of the time series of RTAs was applied to discover homogeneous temporal blocks about the average occurrence of RTAs, obtain the composite variation in the frequency of RTAs and compare with the globally expected value.

Grouping RTAs into clusters with APC using resampling and a single calibration parameter allowed to modulate the number of clusters generated. This number of clusters consequently defined the pattern and distribution of UAs in the geographic space in which RTAs occur. The contribution of this approach, the APC allowed to find homogeneous spatial

data partitions. Such partitions delimit regions with specific accident distributions related to underlying situations that leverage the number of RTA occurrences.

Global and Local Moran's Index provided the necessary tools to identify neighborhood relationships. Such relationships characterize the influence of neighboring clusters with centroids up to 318 meters apart. Neighborhood relations and the frequency of occurrence of RTAs provided the intensity of the relationship of influence between neighbors and the level of significance of the associations verified through the Moran's statistics.

The first selection of APLs was performed with the data grouped in UAs. For this selection, the significant neighborhood relationships between UAs with a high frequency of RTAs were maintained for subsequent analysis of the time series of RTAs.

In this step, elected UAs were submitted to temporal analysis of the pattern of occurrence of RTAs. In this phase, the trend of the series was evaluated, considering the average occurrence of accidents. As the global average of RTAs in the evaluated scenario shows a decrease, local APLs with a stationary average over the period was considered. Such locations with stationary mean constitute regions not susceptible to the trend of global reduction of accidents. Thus, as such locations configure regions of high occurrence of RTAs, inflexible to the variation in the temporal pattern of occurrences of RTAs, they were maintained in the analysis as APLs. This process consists of the second step of data reduction, which discards APLs with a tendency to decay in RTAs over time and maintains APLs with stationary mean as candidates.

The last step of selecting APLs consisted of testing to confirm the CPP results. At this stage, ANOVA and Mann-Whitney tests apply to ensure that the data identified as stationary or with a temporal trend present this characteristic. When the ANOVA and Mann-Whitney tests were applied, they confirmed whether the first and last blocks of the CPP present stationarity by comparing means and median.

The ANOVA and Mann-Whitney tests were compared with the CPP stationarity indicator through contingency tables. The accuracy data does obtain using such tables. Such tests provided the indices of accuracy, sensitivity, and specificity of the CPP test.

6. Conclusions

This work proposed a framework for grouping RTAs, analyzing neighborhood relations, identifying APLs, and prioritizing criteria in urban contexts. The results demonstrate that it was possible to analyze APLs in urban areas with a reduced number of features easily obtainable.

With an integrated algorithm application model, this work developed a roadmap of heuristic applications capable of detecting different underlying patterns of the phenomenon considered. The patterns detected consist of the discovery of i) similarities and dissimilarities related to the distribution pattern of the RTAs; ii) the identification of neighborhood relations between the UAs; and finally, iii) the identification of homogeneous blocks of statistical parameters that describe patterns of distribution of the average of accidents over time in the different UAs.

This effort made it possible to provide an integrated solution using publicly accessible information of RTAs. In this way, the results obtained can be reproduced and replicated in new locations, contributing to the improvement of the proposed model. Although this work analyzes accident frequencies, the results can be applied to other metrics such as severity index, saving costs, or density. It is enough to replace frequency values with the adopted weighted values and monitor the spatial association and variation relationships.

In this way, the present work contributes to using of different metrics in the same framework, making the model replicable, reproducible, and comparable. This versatility provides an easy-to-apply framework using an open-source tool with requires low-complexity information available in any RTAs inventory.

References

- Aguero-Valverde, J., Jovanis, P.P., 2008. Analysis of Road Crash Frequency with Spatial Models. *Transp. Res. Rec. J. Transp. Res. Board* 2061 1 , 55–63. doi:10.3141/2061-07
- Anselin, L., 1995. Local Indicators of Spatial Association—LISA. *Geogr. Anal.* 27 2 , 93–115. doi:10.1111/J.1538-4632.1995.TB00338.X
- Assunção, R.M., Reis, E.A., 1999. A new proposal to adjust Moran's I for population density - *PubMed. Stat. Med.* 18 16 , 2147–62. doi:10.1002/(sici)1097-0258(19990830)18:16<2147::aid-sim179>3.0.co;2-i
- Barry, D., Hartigan, J.A., 1993. A Bayesian Analysis for Change Point Problems. *J. Am. Stat. Assoc.* 88 421 , 309. doi:10.2307/2290726
- Bodenhofer, U., Kothmeier, A., Hochreiter, S., 2011. APCluster: an R package for affinity propagation clustering. *Bioinformatics* 27 17 , 2463–2464. doi:10.1093/bioinformatics/btr406
- Chen, H., 2019. Sequential change-point detection based on nearest neighbors. *Ann. Stat.* 47 3 , 1381–1407.
- Cheng, W., Gill, G.S., Ensich, J.L., Kwong, J., Jia, X., 2018. Multimodal crash frequency modeling: Multivariate space-time models with alternate spatiotemporal interactions. *Accid. Anal. Prev.* 113, 159–170. doi:10.1016/J.AAP.2018.01.034
- Dorcas Wambui, G., 2015. The Power of the Pruned Exact Linear Time(PELT) Test in Multiple Changepoint Detection. *Am. J. Theor. Appl. Stat.* 4 6 , 581. doi:10.11648/j.ajtas.20150406.30

- Fischer Wang, Jinfang., Springer Science+Business Media., M.M., 2011. Spatial data analysis : models, methods and techniques. Springer, Heidelberg.
- Gallagher, C., Lund, R., Robbins, M., 2013. Change point detection in climate time series with long-term trends. *J. Clim.* 26 14 , 4994–5006. doi:10.1175/JCLI-D-12-00704.1
- Gatrell, A.C., Bailey, T.C., Diggle, P.J., Rowlingson, B.S., 1996. Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology. *Trans. Inst. Br. Geogr.* 21 1 , 256. doi:10.2307/622936
- Getis, A., Ord, J.K., 1992. The Analysis of Spatial Association by Use of Distance Statistics. *Geogr. Anal.* 24 3 , 189–206. doi:10.1111/J.1538-4632.1992.TB00261.X
- Gurevich, G., 2006. Nonparametric AMOC Change point Tests for Stochastically Ordered Alternatives. *Commun. Stat. - Theory Methods* 35 5 , 887–903. doi:10.1080/03610920500501411
- Haynes, K., Eckley, I.A., Fearnhead, P., 2014. Efficient penalty search for multiple change point problems.
- Hennig, C., Viroli, C., Anderlucci, L., 2019. Quantile-based clustering. *Electron. J. Stat.* 13 2 , 4849–4883. doi:10.1214/19-EJS1640
- Huang, H., Zhou, H., Wang, J., Chang, F., Ma, M., 2017. A multivariate spatial model of crash frequency by transportation modes for urban intersections. *Anal. Methods Accid. Res.* 14, 10–21. doi:10.1016/J.AMAR.2017.01.001
- Killick, R., Fearnhead, P., Eckley, I.A., 2011. Optimal detection of changepoints with a linear computational cost. *J. Am. Stat. Assoc.* 107 500 , 1590–1598. doi:10.1080/01621459.2012.737745
- Lee, A.S., Lin, W.H., Gill, G.S., Cheng, W., 2019. An enhanced empirical bayesian method for identifying road hotspots and predicting number of crashes. *J. Transp. Saf. Secur.* 11 5 , 562–578. doi:10.1080/19439962.2018.1450314
- Luzardo, A.J.R., Filho, R.M.C., Rubim, I.B., 2017. Análise espacial exploratória com emprego do Índice de Moran. *GEOgraphia* 19 40 , 161–179. doi:10.22409/GEOGRAPHIA2017.V19I40.A13807
- Moura, F. sutel de, Nodari, C.T., 2022. Application of the Affinity Propagation Clustering Technique to obtain traffic accident clusters at macro, meso, and micro levels. *arxiv.org preprint*.
- Moura, F.S. de, Garcia, L.F., Torres, T.B., Oliveira, L.P. de, Nodari, C.T., 2022. Identification and Prioritization Techniques Applied to Urban Accident-Prone Locations: A Scoping Review. *OSF Prepr.* doi:10.31219/OSF.IO/HUQ58
- Nie, K., Wang, Z., Du, Q., Ren, F., Tian, Q., 2015. A network-constrained integrated method for detecting spatial cluster and risk location of traffic crash: A case study from Wuhan, China. *Sustain.* 7 3 , 2662–2677. doi:10.3390/su7032662
- Shi, X., 2020. A Survey of Change point Techniques for Time Series Data. *Clemson Universit.*