# Zonal Statistics to Identify Hot-regions of Traffic Accidents

**Ömer M. Soysal[1,2,3], Helmut Schneider[1,2], Asim Shrestha[1,2], Christy D. Guempel[2], Pei Li[2], Harisha Donepudi[2,3], Naveen K. Kondoju[2,3], Kazim Sekeroglu[2]**

{[1]Department of Information Systems and Decision Sciences, [2]Highway Safety Research Group, [3]Department of Computer Science}, Louisiana State University, Baton Rouge, LA, USA

**Abstract** - *This paper presents how to utilize ArcGIS zonal statistic tool for identification of hot-regions. For this purpose, we used two descriptive statistical indexes Sum and Max over normalized raster data. Our effort was to get a better understanding of the results generated by the zonal statistics tool to identify the traffic accidents hot-spot zones. Our approach makes it possible for comparison of statistical parameters across different zone areas within the same scale. This helps decision makers to pinpoint locations with higher crash index and take preventive measures.*

**Keywords:** Hot-spot, Zonal statistics, traffic accident, raster.

## 1 INTRODUCTION

Spatial data like weather data, demographics, land management data, and crash data need to be analyzed at different levels for analysis of trends and patterns. In the analysis of crash data, it is crucial to identify hot-spot regions of traffic accidents. For such identification, a comparison of different zones of interest is required based upon statistical parameters.

A difficulty in this type of analysis is that aggregate spatial data (like sum, maximum) exists at different levels of scale on the map, for example, sum or maximum of crime incidents across the nation cannot be directly compared to that of a county or a city. For such type of comparative analysis, we propose to normalize the results obtained from the zonal analysis in ArcGIS. We use these normalized values to identify the hot-spots of traffic accidents over different levels of administrative zones. Furthermore, the use of two different statistics parameters ensures that different hot-spot regions can be identified according to the nature of analysis.

Hot-spot Analysis is a key area for research not just for traffic accidents but a wide range of other topics such as environmental studies, spread of diseases, biological studies, migration studies, etc. Different methodologies have been proposed in the literature to study the patterns of occurrence of spatial data. The methodologies can be cross-domain; meaning that methods applied to find out hot-spots in one area of research can also be applied to another. [1] has used Kernel Density Estimation and K-means clustering to study the spatial patterns of injury related road accidents and to create a classification of road accident hot-spots.

There is no universal definition of accidental hot-spots and researchers have used different sophisticated methods to quantify hot-spots. [2] has provided a review of the following three different hot-spot detection techniques:

- Kernel density estimation
- Network analysis
- Census Output Area estimation

An extension to the Kernel density estimation in which the network space is represented with basic linear units of equal network length, known as lixe (linear pixel) and related network topology is proposed by [3]. They argue that the use of lixe facilitates the systematic selection of a set of regularly spaced locations along a network for density estimation and makes the practical application of the network Kernel density estimation feasible by significantly improving the computation efficiency.

[4] has used a process called kernel smoothing for hot-spot analysis. This process creates local estimates of the measure of the spatial intensity using the count of frequency of points within a given distance of each point, relative to symmetric distribution. The author has used a Gaussian kernel of 0.5 kilometer bandwidth for the hot-spot analysis.

Zonal statistics which we employed in this paper follows a raster-based method. Raster-based methods are widely used in environmental and geophysical studies [5], geographic analysis [6], surface modeling [7], planning and design [8], biological studies [9].

In this paper, we have used the sum values of different zones (of the same scale) to identify hot-spots with a greater number of events (crashes) per unit area. The Max values, on the other, hand are used to identify hot-spot zones that have high number of events at sub-zonal units. We have used two statistical parameters instead of just one so that different zones can be identified as hot-spots according to the objective of the analysis. The usage of zonal statistics and the visualization of normalized results obtained from it will help to draw meaningful analysis and interpretations. It serves as a visual aid in knowing where hot-spots occur and helps in finding out the reason behind high number of un-expected crashes in specific locations.

## 2 METHOD

We first used the zonal statistics tool in ArcGIS to get the sum and max index of each zone for three different zonal

levels. The statistical parameters thus obtained were normalized and hot-spots were identified based upon these values.

### 2.1 Raster Based Approach for Computation of Sum and Max

In this approach a grid is drawn such that it encompasses the area over which hot-spot analysis is to be performed. An area under consideration would contain a finite number of grid cells. Sum is computed as the total number of events (crashes in our case) contained by all the grid cells enclosed within a particular zone. Similarly, Max is computed as the highest number of events enclosed by a single grid cell of a zone. In Figure 1 four different zones marked by boundaries with color red, orange, blue, and green are shown. The Sum and Max values for the four zones are (7, 5), (8, 2), (3, 1), and (1, 1) respectively. If the hot-spot zones are identified according to Sum value then the orange zone would rank at the top, while the red zone is identified as the 'hottest' zone if Max value is chosen as the criteria. Thus, the use of both Sum and Max values ensures that all the hot-spot zones are identified depending upon the objective of the ranking. High values of Sum indicate a high overall total number of events in a zone which may or may not be influenced by local high values. High values of Max on the other hand indicate the presence of a local region within a zone with a high density of events.

### 2.2 Normalization

The results obtained from zonal statistics are not sufficient for insightful comparisons. Zones with high values of Sum may have those values only because their area is bigger. Also, one zone may have only a single small area with high number of accidents and show a very high Max value. In order to have a better interpretation of crash patterns, we normalize the results by first converting the zonal areas which are in square decimal degrees to square miles, and then getting the Sum for one square mile of area in each of the zones. The total sum per unit square mile area is computed for each zone level. The same process is followed for Max value of each of the three zonal levels.

Since the normalized results of the sum value shows the number of accidents within a one mile square area, comparisons can be made among the different zones within the same scale, i.e. a direct inference can be drawn about whether one parish/census tract/census block group has more accidents per square mile than the other. Sum values suggest which zones have the highest number of total accidents per unit area. Max value gives us an idea about where the peak values (sub-zonal unit areas with highest number of crashes) are located and about their peak value. When interpreting the Max value, it should be remembered that the peak value is within one mile area somewhere in the corresponding zone unit and comparisons are logical only within the same zonal scale level. Figure 4 shows the results obtained by using zonal statistics tool and the normalized results based on our calculations.

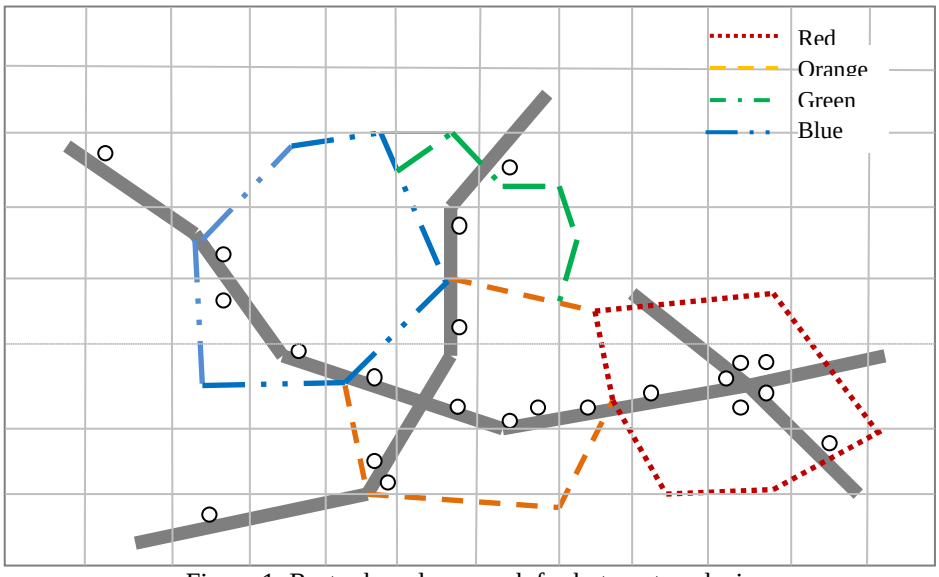


Figure 1: Raster based approach for hot-spot analysis

### 2.3 Hot-spot Identification

Use of only one statistical parameter can be inconclusive for the identification of the hot-spot zones. If Sum is used to rank the hot-spots then zones with high peak values may be left out. Conversely, if Max is used to rank the hot-spot regions then zones with a high total number of events per unit area could be ignored. This point is illustrated in the Figure 2 and explained below.

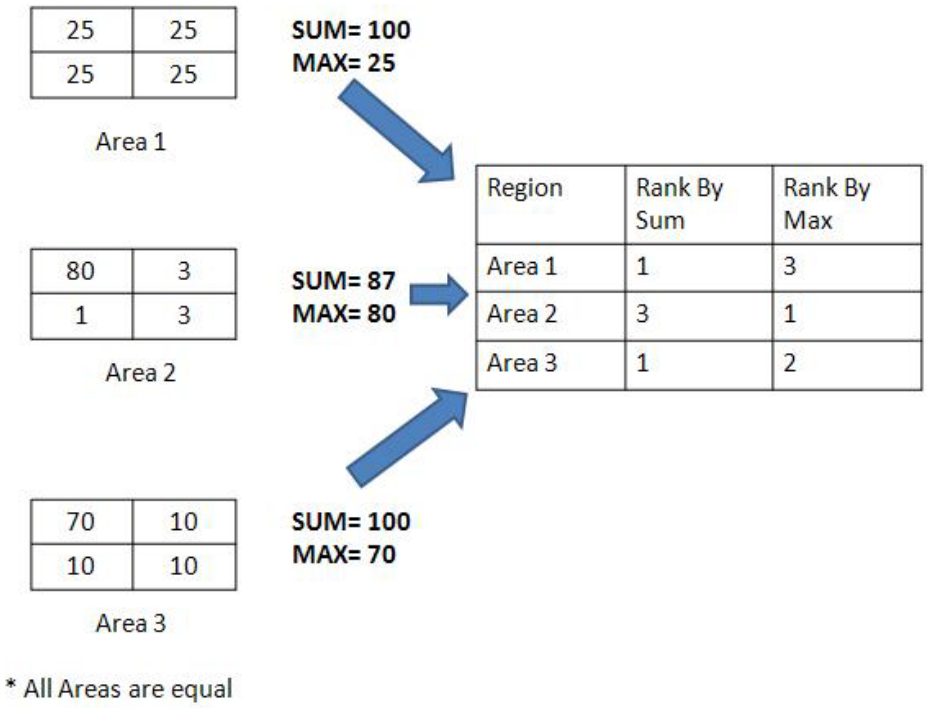



Figure 2: Hot-spot Identification by using Sum and Max

We can see from Figure 2 that Area 1 is composed of consistently high valued sub-zones, and its rank would be 1 if the sum value is considered. Area 2 on the other hand has one single sub-zone with a very high Max value. Area 2 would be ranked number 1 if Max value is to be considered. In Area 3 there is a high Max value as well as other sub-zonal area units with significant counts. It is therefore necessary not to rely upon a single parameter for the identification of the hot-spots. Depending on the objectives of the analysis, a list of different zones could be identified as hot-spots.

# 3 RESULTS

We have used a zonal statistics tool in ArcGIS to get the sum and max values of crash data at three different zonal scale levels: 1) Parish level, 2) Census Tract level, and 3) Census Block Group level. Sum and Max of the crash data from different zonal scale levels cannot be compared directly for any insightful interpretations. For each zonal scale level, we normalize the Sum and Max values generated by the zonal statistics to get the total/highest number of accidents within a one mile square area for each zone. The Sum and Max indexes are visualized per unit area of the zone. The map visualization of the normalized results using Sum and Max values are presented in Figure 5 and 6.

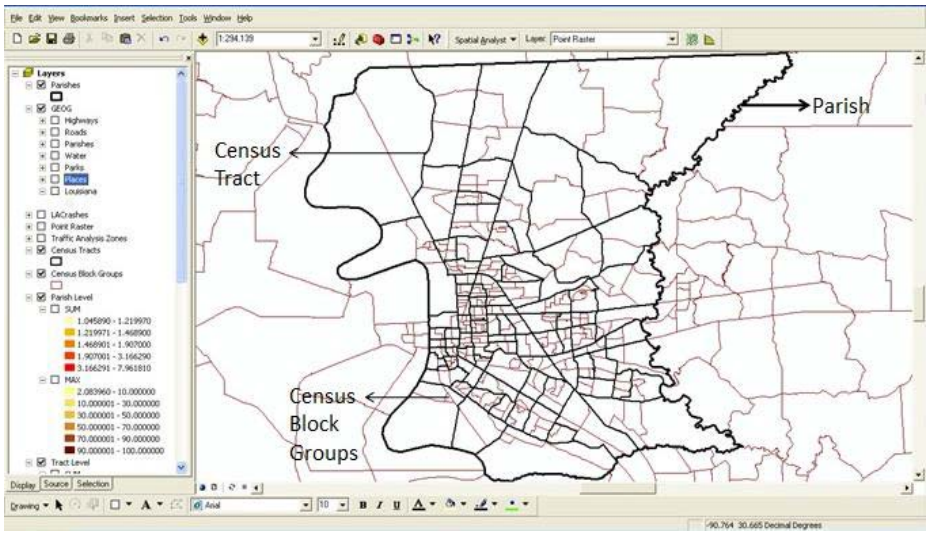
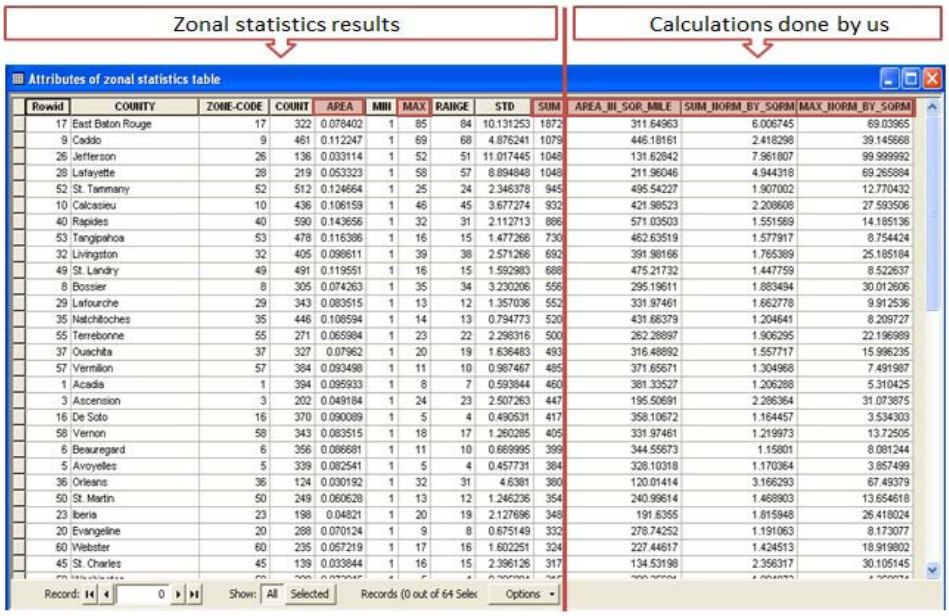



Figure 3: Three Different Zonal Levels




| Rowid | COUNTY | ZONE-CODE | COUNT | AREA | MIN | MAX | RANGE | STD | SUM | AREA_III_SQR_MILE | SUM_NORM_BY_SQRM | MAX_NORM_BY_SQRM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 17 | East Baton Rouge | 17 | 322 | 0.078402 | 1 | 85 | 84 | 10.131253 | 1872 | 311.64963 | 6.006745 | 69.03965 |
| 9 | Caddo | 9 | 461 | 0.112247 | 1 | 69 | 68 | 4.876241 | 1079 | 446.18161 | 2.418298 | 39.145668 |
| 26 | Jefferson | 26 | 136 | 0.033114 | 1 | 52 | 51 | 11.017445 | 1048 | 131.62842 | 7.961807 | 99.999992 |
| 28 | Lafayette | 28 | 219 | 0.053323 | 1 | 58 | 57 | 8.894848 | 1048 | 211.96046 | 4.944318 | 69.265884 |
| 52 | St. Tammany | 52 | 512 | 0.124664 | 1 | 25 | 24 | 2.346378 | 945 | 495.54227 | 1.907002 | 12.770432 |
| 10 | Calcasieu | 10 | 436 | 0.106159 | 1 | 46 | 45 | 3.677274 | 932 | 421.98523 | 2.208608 | 27.593506 |
| 40 | Rapides | 40 | 590 | 0.143656 | 1 | 32 | 31 | 2.112713 | 886 | 571.03503 | 1.551589 | 14.185136 |
| 53 | Tangipahoa | 53 | 478 | 0.116386 | 1 | 16 | 15 | 1.477266 | 730 | 462.63519 | 1.577917 | 8.754424 |
| 32 | Livingston | 32 | 405 | 0.098611 | 1 | 39 | 38 | 2.571268 | 692 | 391.98166 | 1.765389 | 25.185184 |
| 49 | St. Landry | 49 | 491 | 0.119551 | 1 | 16 | 15 | 1.592983 | 688 | 475.21732 | 1.447759 | 8.522637 |
| 8 | Bossier | 8 | 305 | 0.074263 | 1 | 35 | 34 | 3.230206 | 556 | 295.19611 | 1.883494 | 30.012606 |
| 29 | Lafourche | 29 | 343 | 0.083515 | 1 | 13 | 12 | 1.357036 | 552 | 331.97461 | 1.662778 | 9.912536 |
| 35 | Natchitoches | 35 | 446 | 0.108594 | 1 | 14 | 13 | 0.794773 | 520 | 431.66379 | 1.204641 | 8.209727 |
| 55 | Terrebonne | 55 | 271 | 0.065984 | 1 | 23 | 22 | 2.298316 | 500 | 262.28897 | 1.906295 | 22.196989 |
| 37 | Ouachita | 37 | 327 | 0.07962 | 1 | 20 | 19 | 1.636483 | 493 | 316.48892 | 1.557717 | 15.996235 |
| 57 | Vermilion | 57 | 384 | 0.093498 | 1 | 11 | 10 | 0.987467 | 485 | 371.65671 | 1.304968 | 7.491987 |
| 1 | Acadia | 1 | 394 | 0.095933 | 1 | 8 | 7 | 0.593844 | 460 | 381.33527 | 1.206288 | 5.310425 |
| 3 | Ascension | 3 | 202 | 0.049184 | 1 | 24 | 23 | 2.507263 | 447 | 195.50691 | 2.286364 | 31.073875 |
| 16 | De Soto | 16 | 370 | 0.090089 | 1 | 5 | 4 | 0.490531 | 417 | 358.10672 | 1.164457 | 3.534303 |
| 58 | Vernon | 58 | 343 | 0.083515 | 1 | 18 | 17 | 1.260285 | 405 | 331.97461 | 1.219973 | 13.72505 |
| 6 | Beauregard | 6 | 356 | 0.086681 | 1 | 11 | 10 | 0.669995 | 399 | 344.55673 | 1.15801 | 8.081244 |
| 5 | Avoyelles | 5 | 339 | 0.082541 | 1 | 5 | 4 | 0.457731 | 384 | 328.10318 | 1.170364 | 3.857499 |
| 36 | Orleans | 36 | 124 | 0.030192 | 1 | 32 | 31 | 4.6381 | 380 | 120.01414 | 3.166293 | 67.49379 |
| 50 | St. Martin | 50 | 249 | 0.060628 | 1 | 13 | 12 | 1.246236 | 354 | 240.99614 | 1.468903 | 13.654618 |
| 23 | Iberia | 23 | 198 | 0.04821 | 1 | 20 | 19 | 2.127696 | 348 | 191.6355 | 1.815948 | 26.418024 |
| 20 | Evangeline | 20 | 288 | 0.070124 | 1 | 9 | 8 | 0.675149 | 332 | 278.74252 | 1.191063 | 8.173077 |
| 60 | Webster | 60 | 235 | 0.057219 | 1 | 17 | 16 | 1.602251 | 324 | 227.44617 | 1.424513 | 18.919802 |
| 45 | St. Charles | 45 | 139 | 0.033844 | 1 | 16 | 15 | 2.396126 | 317 | 134.53198 | 2.356317 | 30.105145 |

Figure 4: Zonal Statistics and Normalized Results

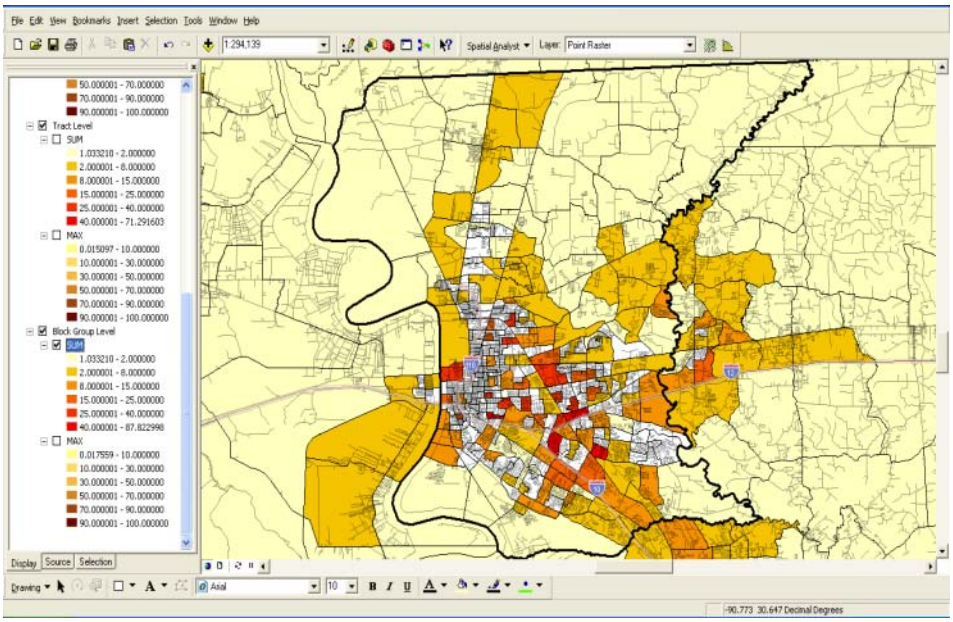


Figure 5: SUM (Census Block Groups Scale, zoomed at a single parish)

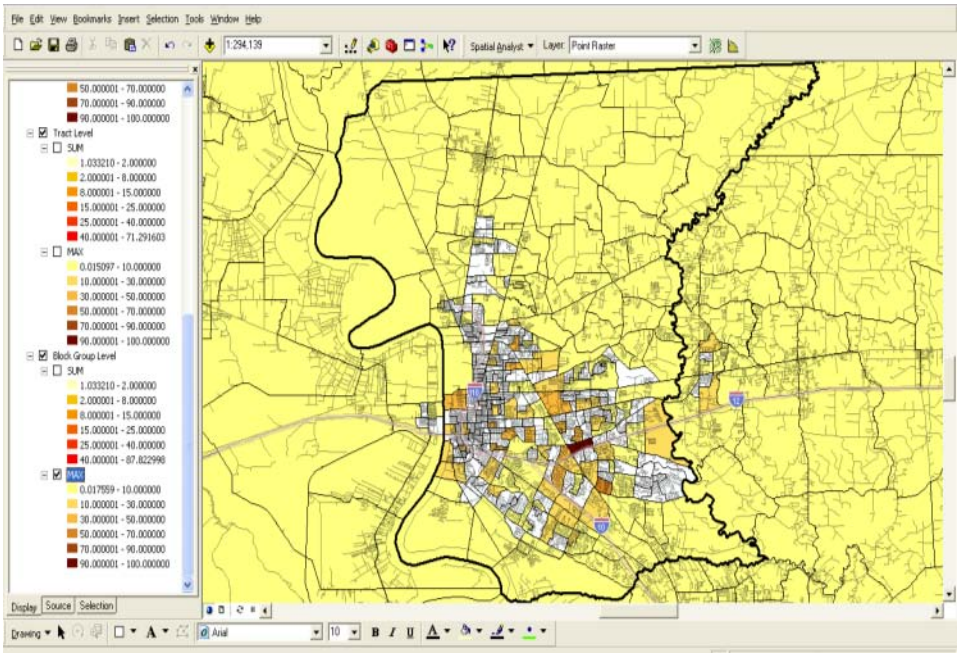


Figure 6: MAX (Census Block Groups Scale zoomed at a single parish)

# 4 CONCLUSION

We have presented a way to improvise the results generated by the zonal statistics in ArcGIS. The normalization step that we propose following the zonal statistics computation is used for hot-spot analysis to find out patterns and/or trends of crashes. Our effort was to get a better understanding of the results generated by the zonal statistics tool to identify the traffic accident hot-spot zones. Our approach makes it possible for comparison of statistical parameters across different zone areas within the same scale. This helps decision makers to pinpoint locations with a higher crash index and take preventive measures. The road condition can be improved, proper signs can be placed and, traffic laws can be imposed more strictly in the areas determined as having high risk by the analysis. This would not only help to prevent loss of millions of dollars' worth of property that are damaged due to crashes, but more importantly save thousands of innocent lives each year.

**ACKNOWLEDGEMENT**: The authors of this paper would like to thank LADOTD for supporting this work.

# 5 REFERENCES

[1] T. K. Anderson, "Kernel density estimation and K-means clustering to profile road accidenthotspots," *Accident Analysis & Prevention, 41(3),* pp. 359-364, 2009.

[2] T. Anderson, "Comparison of spatial methods for measuring road accident 'hotspots': a case study of London," *Journal of Maps, 3 (1),* 2007.

[3] Z. Xie and J. Yan, "Kernel Density Estimation of trafficaccidents in a network space," *Computers, Environment and Urban Systems, 32 (5),* pp. 396-406, 2008.

[4] L. Schweitzer, "Environmental justice and hazmat transport: A spatial analysis in southern California," *Transportation Research Part D: Transport and Environment, 11 (6),* pp. 408-421, 2006.

[5] P. D. Bates and A. P. J. De Roo, "A simple raster-based model for flood inundation simulation," *Journal of Hydrology, 236 (1-2),* pp. 54-77, 2000.

[6] S. K. Jenson and J. O. Domingue, "Extracting Topographic Structure from Digital Elevation Data for Geographic Information System Analysis," *Photographic Engineering and Remote Sensing,* pp. 1593-1600, 1988.

[7] P. Y. Julien, B. Saghafian and F. L. Ogden, "Raster-Based Hydrological Modeling of Spatially-varied Surface Runoff," *Journal of the American Water Resources Association,* pp. 523-536, 1995, 31 (3).

[8] C. Ratti and P. Richens, "Raster Analysis of Urban Form," *Environment and Planning B: Planning and Design,* pp. 297-309, 2004, Vol 31.

[9] I. S. Sotheran, R. L. Foster-Smith and J. Davies, "Mapping of marine benthic habitats using image processing techniques within a raster-based geographic information system," *Estuarine, Coastal and Shelf Science,* pp. 25-31, 1997, 44 (1).

[10] ESRI, "An Overview of the Zonal tools," [Online]. Available: http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=An_overview_of_the_Zonal_tools. [Accessed 16 May 2012].

[11] ESRI, "How Zonal Statistics Works (ArcGIS Desktop Help)".

# 6 Appendix

## 6.1 Zonal Statistics

Zonal analysis is an important analysis tool in ArcGIS under its spatial analyst extension. It is useful for several types of GIS-related analysis or studies such as environmental monitoring, demographic studies, land management, traffic data analysis and so on. Zonal analysis is the creation of an output raster (or statistics table) in which the desired function is computed on the cell values from the input value raster that intersect or fall within each zone of a specified input zone dataset. The zonal tools in which the zones are defined by a single input value raster either calculate statistics or quantify the characteristics of the geometry of the input zones.

Zonal tools are categorized as zone shapes (Zonal Area, Zonal Centroid, Zonal Perimeter), zone attributes (Zonal Max, Zonal Min, Zonal Sum), and determine specified zones (Zonal Fill). The Zonal Statistics tool records the specified statistic of the values of all cells in the value Dataset that belong to the same zone as the output cell in each output cell [10]. An input zone for the zonal statistics tool can be a vector file or an integer raster file. Similarly, an input value raster includes any raster file that contains values that can be analyzed visually and statistically. The input zone dataset is only used to define the size, shape, and location of each zone, while the input value raster contains the values to be used in the evaluations within the zones.

**Maximum**: For this statistic, the zone input must be an integer. The data type of the output will be the same as the value input.

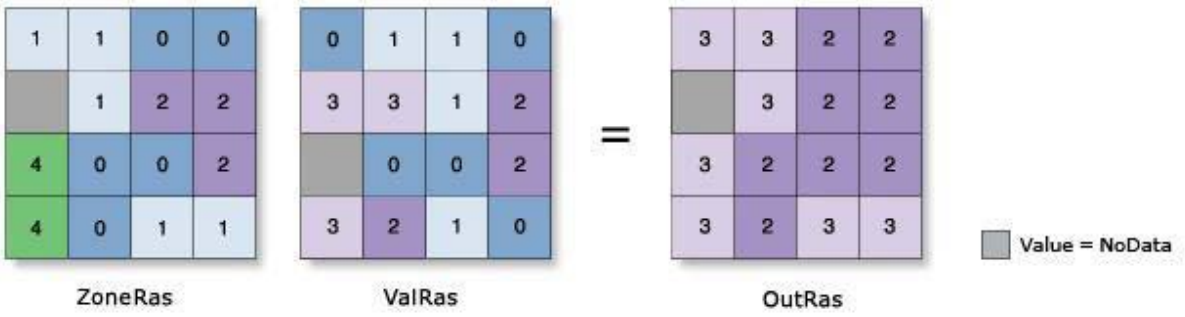



Figure 7: Maximum Calculation (Zonal Statistics) [11]

In our work, Maximum value gives an idea about where the peak values are located and what the peak value is. During interpretation of the maximum value, it should be remembered that the peak value is within a one mile area somewhere in the corresponding zone unit.

**Sum**: For this statistic, the zone input must be an integer. The data type of the output raster is a floating point. This is because the value for the Sum tends to be quite large, and may not be possible to represent with an integer value.

For example, if a zone has 2500 rows and columns of cell in size, and the value of each cell is 1000, the sum for that zone would be 2500*2500*1000 = 6.25 billion [11]. If an integer output is required and the range is within ± 2.147 billion, then the INT function can be applied [11].

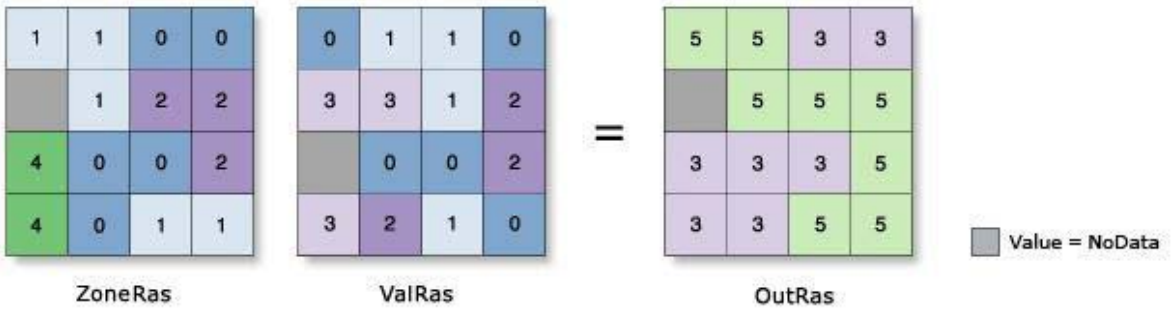



Figure 8: Sum Computation (Zonal Statistics) [11]

In our work Sum value shows the number of accidents within a one mile square area.