Utilizing ArcGIS Pro to Analyze Collision Locations and Contributing Factors to Collision Rates in Corvallis, Oregon 2007-2017

Amanda Riley

GIS in Transportation Engineering

Final Project

Introduction

Collision analysis, also called crash analysis, is commonly used in Transportation Safety Planning (TSP) to identify collision trends and contributing factors. The Oregon Department of Transportation (ODOT) references the statewide Crash Data System (CDS) for standard collision analysis. This system is proprietary to ODOT and may not be available to all analysts. This project will attempt to show that collision analysis may be completed for the state of Oregon using statistical analysis in R and ArcGIS Pro analysis software.

Statistical data analysis in transportation is commonly done by using regression methods. (Zhu, Yu, Wang, B., & Tang, 2018) For this project the binary logit model of logistical regression will be used to determine the statistical significance of collision factors. The binary logit model allows for factors of different sample volumes to be compared across studies and avoid some oversimplification in linear regression models.

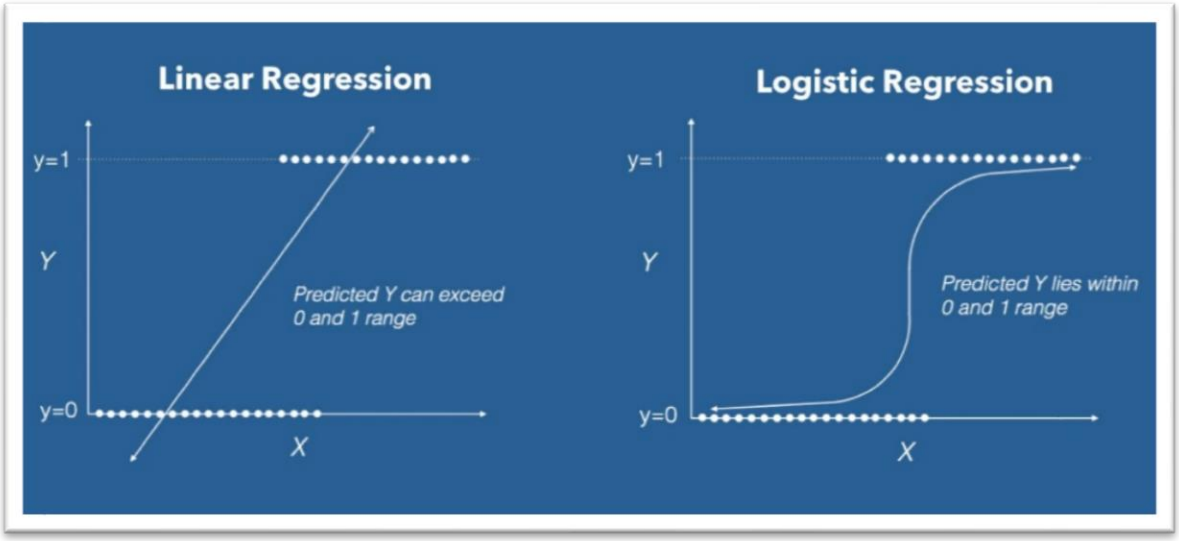


Figure 1.Examples of linear and logistic regression (Le, 2019)

https://www.datacamp.com/community/tutorials/logistic-regression-R

Background

Corvallis is located in central western Oregon in Benton County. With a current population of 54,462 (United States Census Bureau, 2020) and home to Oregon State University (OSU), the city experiences approximately 500 collisions each year. Three state routes run through the city: Highway 20 and Highway 34 run east/west and Highway 99W runs north/south, with one interchange for these highways. Multiple bridges over the Willamette River and the Marys River create constriction points in the traffic flow. These conditions are exacerbated by a large amount of traffic during home football game at OSU.

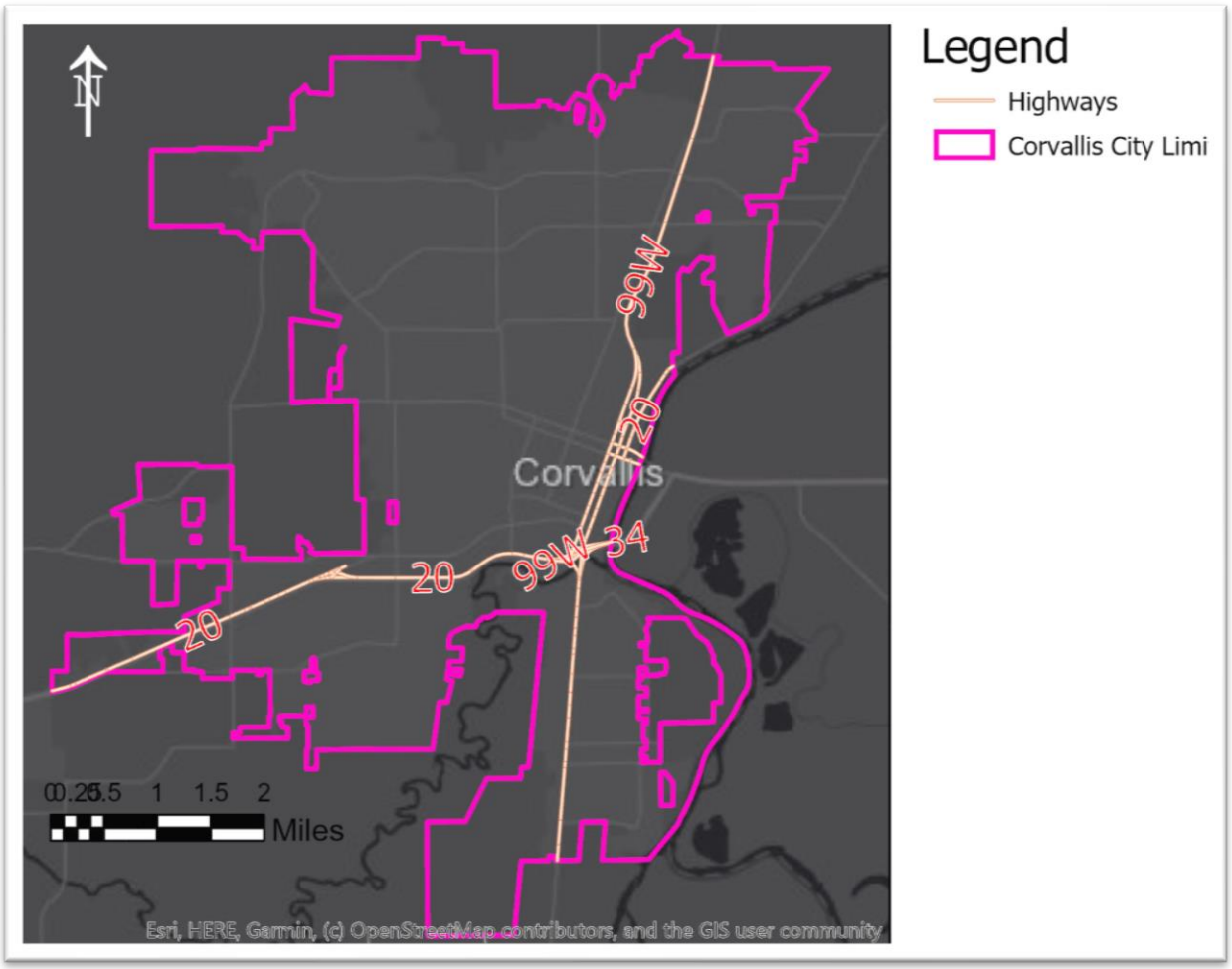


Figure 2. Corvallis, Oregon

Data Acquisition

All data acquired for use in ArcGIS Pro was sourced from municipalities and is Geographic Information System (GIS) mapping compliant. The Oregon Department of Transportation (ODOT) provided all transportation related data. This data included the geometry of the roadways, crash data, and city limits. The data was gathered from the source location for TransGIS, the ODOT interactive GIS mapping website (Oregon Department of Transportation, 2020). Of the crash data gathered crash data for eleven years was used from 2007-2017. Crash data was broken down into multiple unique variables. All data used in ArcGIS Pro was geocoded using accepted surveying practices (Office of Planning, Federal Highway Administration, 2012). Additional data was acquired from Wikipedia for the home dates of OSU home football games (Wikipedia, 2020).

Data Manipulation

GIS data is coded to comply with geographic coordinates used by the original survey source. This is called the datum. The datum used in these analyses is the geographic coordinate system: GCS North American 1983. This datum takes Geodetic Reference Systems (GRS) 1980 spheroid and both survey and satellite observations into account (ESRI, 2020). Two manipulations are made to this original datum to enable its use in the state of Oregon. Due to the spherical nature of GCS North American 1983, it does not translate well into flat maps. Therefore, the Lambert Conformal Conic projection is used as is standard in the state of Oregon. This adjusts the spherical coordinate system to a cone shape which provides a more accurate representation of the data on a flat surface as seen in figure 3.

Figure 3. Conic Projection (GISGeography, 2020)

https://gisgeography.com/conic-projection-lambert-albers-polyconic/

The second manipulation occurs due to standard coordinate systems differing from state to state. The GCS North American 1983 is projected into the Oregon state standard system NAD 1983 Oregon Statewide Lambert Feet International. This system reduces the distortion created by the conic projection. Various layers of data were standardized to these two projections for these analyses.

Corvallis Transportation Overview

Both spatial data and temporal data were used in these analyses. ODOT provided Roadway data which included all geometries used in the highway, collector and public roads data. Crash data used included 47 unique variables including but not limited to: time of day, day of week,

KABCO injury severity and roadway conditions. The city limits of Corvallis which were used as the boundary for the first set of analyses was also provided by ODOT.

Collision Trends

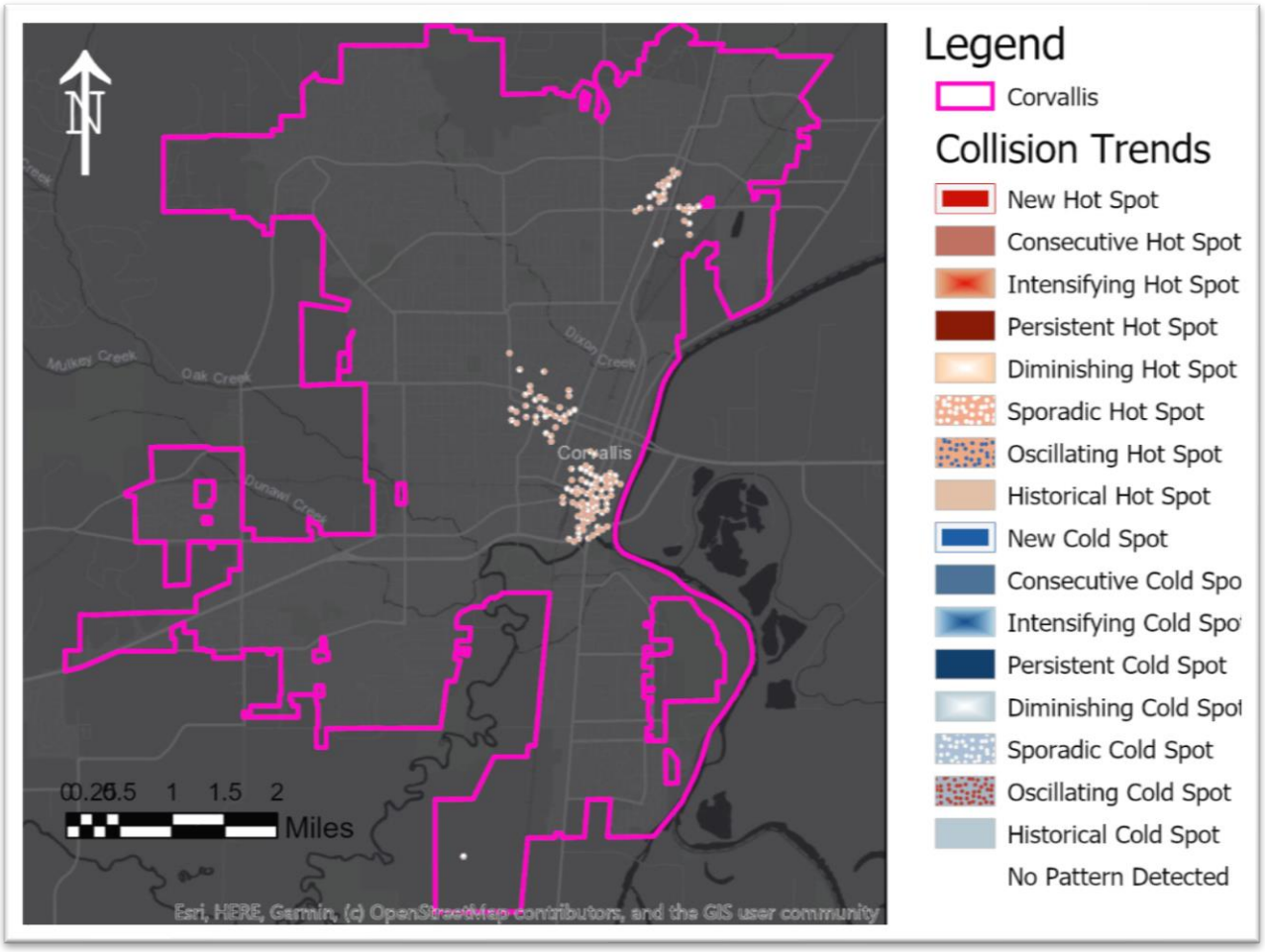


Figure 4. Corvallis, Oregon - Collision Trends

Collision Trends for Corvallis were determined by using the Emerging Hot Spot Analysis tool in ArcGIS Pro. All significant hot spots found are sporadic hot spots, seen above in figure 4. The Emerging Hot Spot Analysis tool uses count data to identify trends in the clustering of point densities (ESRI, 2020). These clusters are placed in a time series bin called a Space Time Cube.

Space Time Cubes call be either two dimensional or three dimensional. Two-dimensional Space Time Cubes were used for this analysis. Neighboring cubes use spatial relationships of a fixed distance to determine adjacency. Each bin is analyzed in comparison to the neighboring cubes to measure the intensity of clustering. Trends on increasing or decreasing intensity are then identified by determining the z-score and p-score of the null hypothesis of complete spatial randomness. These trends are determined to three confidence levels: 90%, 95%, and 99% and then categorized. The general categories are either increasing, decreasing or no pattern detected. The findings for the project show sporadic hot spots near downtown, northeast of OSU, and near NW 9th Street and NW Circle Boulevard.

Collision Rate Changes

Collision rate changes in Corvallis were then determined by finding Hot Spots at which the collision trends are changing. Hot spots are areas were statistically significant change in rate are seen. Increases are seen in red and decreases are seen in blue in figure 5 below. These are determined by using the Hot Spot Analysis (Getis-Ord Gi*) geoprocessing tool. This tool allows for a weight set of features to be used to determine statistically significant hot and cold spots (ESRI, 2020). Like the Emerging Hot Spot Analysis tool, z-scores and p-scores and confidence levels are used. Unlike the previous tool, the Getis-Ord statistic is more complex than complete spatial randomness due to spatial weight being employed. The statistical equations used are as shown in figure 6.

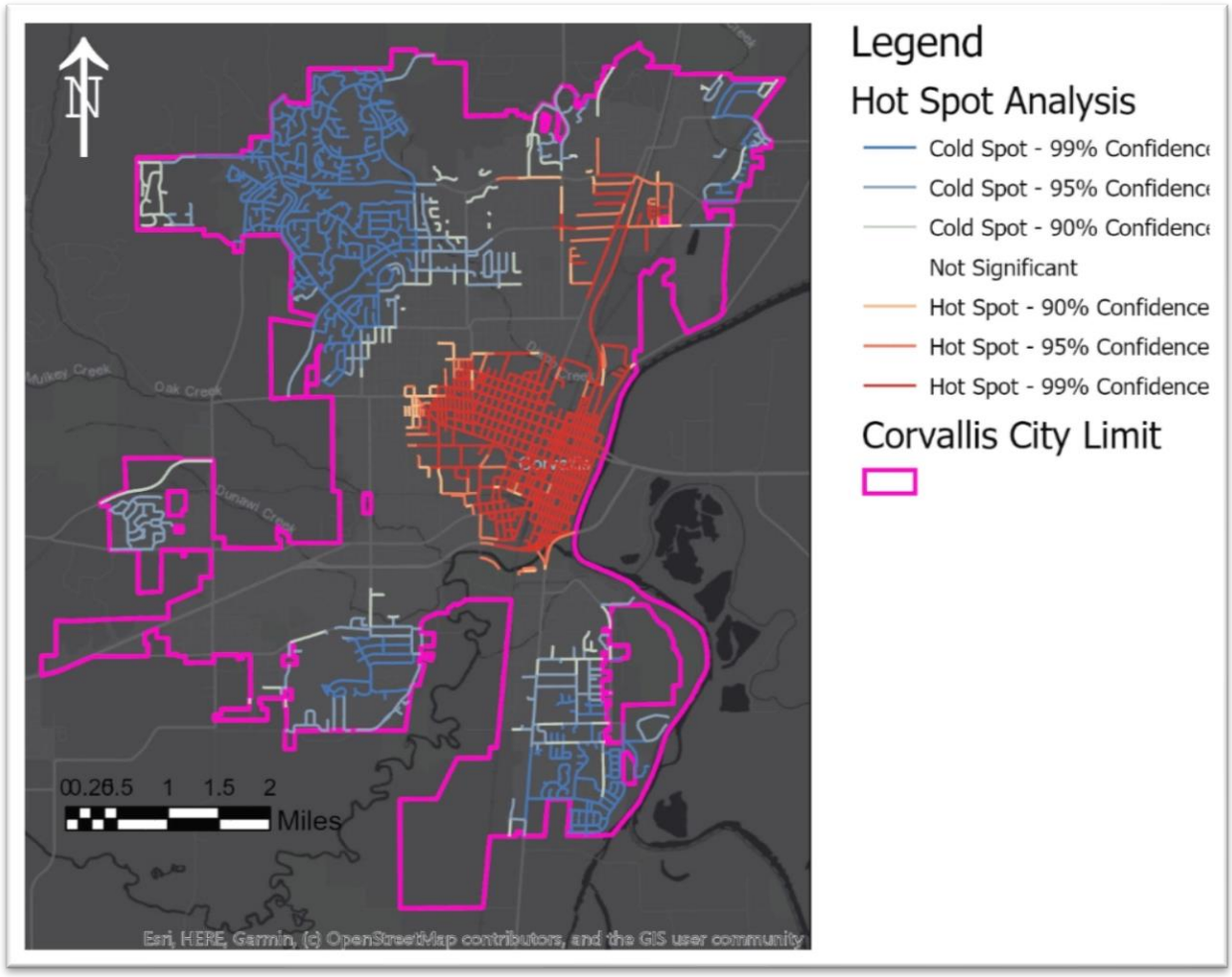
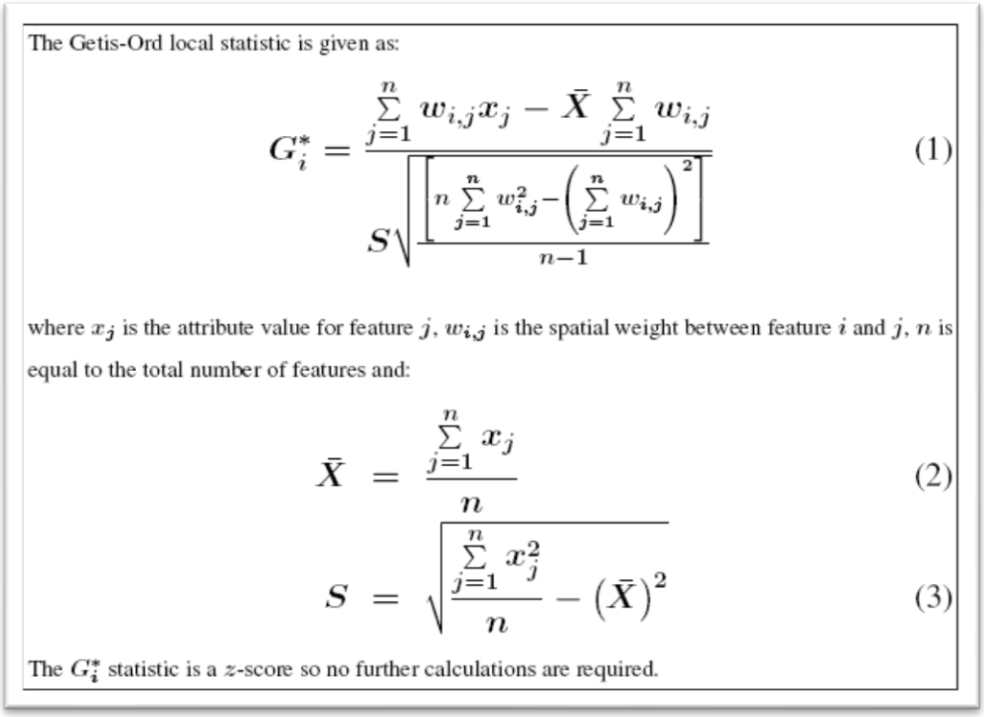


Figure 5. Corvallis, Oregon - Collision Rate Changes

A false discovery rate correction is then applied to reduce the p-value determining the cofidence levels. This will help to account for multiple testing and spatial dependence. Segments of road are determined to provide smaller resonable lengths. This creates a valueable analysis that allows for the determination of where change is occuring. Segments used in this analysis were determined by ODOT and were within the GIS roadway data. A weights matrix file was used for this analysis based on a euclidian distance of 15 feet. This provides the distance that the tool may

use to determine the significance of a collision to the segment of road. These spatial references and collision trends determine the increasing and decreasing rates along roadway segments.



The Getis-Ord local statistic is given as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}} \quad (1)$$

where x_j is the attribute value for feature j, $w_{i,j}$ is the spatial weight between feature i and j, n is equal to the total number of features and:

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n} \quad (2)$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\bar{X}\right)^2} \quad (3)$$

The G_i^* statistic is a z-score so no further calculations are required.

Figure 6. Getis-Ord local statistics (ESRI, 2020)

https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-hot-spot-analysis-getis-ord-gi-spatial-stati.htm

Collision Clusters

Using the ESRI Incident Analysis toolbox, a count of clustered collisions was determined. Shown in figure 7 below. “The cluster analysis tool identifies spatial clusters and labels each cluster with the number of incidents it contains” (ESRI, 2020). This is a simple count of collisions located in a general given area. The area used for this analysis was 100 feet. No geographical boundary was used for this analysis and this limits its usefulness.

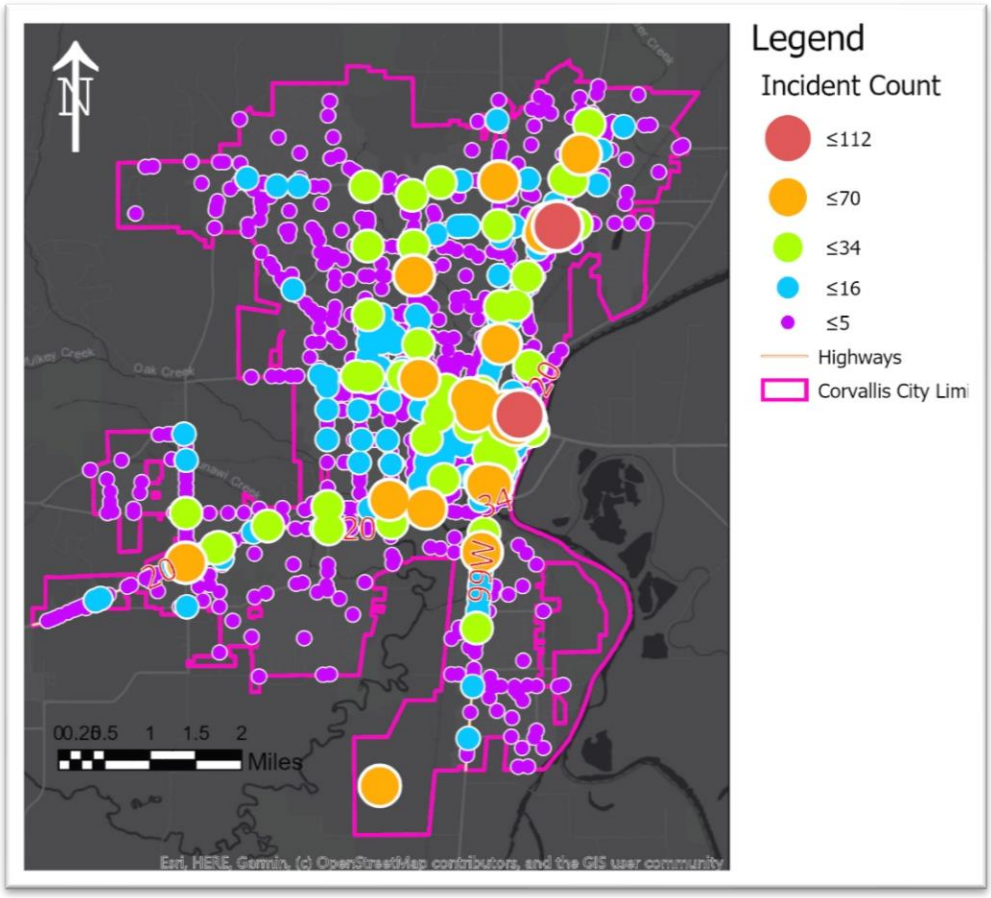


Figure 7. Corvallis, Oregon - Collision Clusters

Collision Clusters by Location

Using the ESRI Incident Analysis toolbox, a count of collisions by lines of communication (LOC) was determined. Shown in figure 8 below. The highways, collector roads, and local roads were used to determine the count on each segment of the facility. The tool uses a search radius to measure how near a collision is to a line of communication. A distance of 100 feet was used. This results in a corridor representation of frequency, very useful in determining what segments of roadway may need remediation measures.

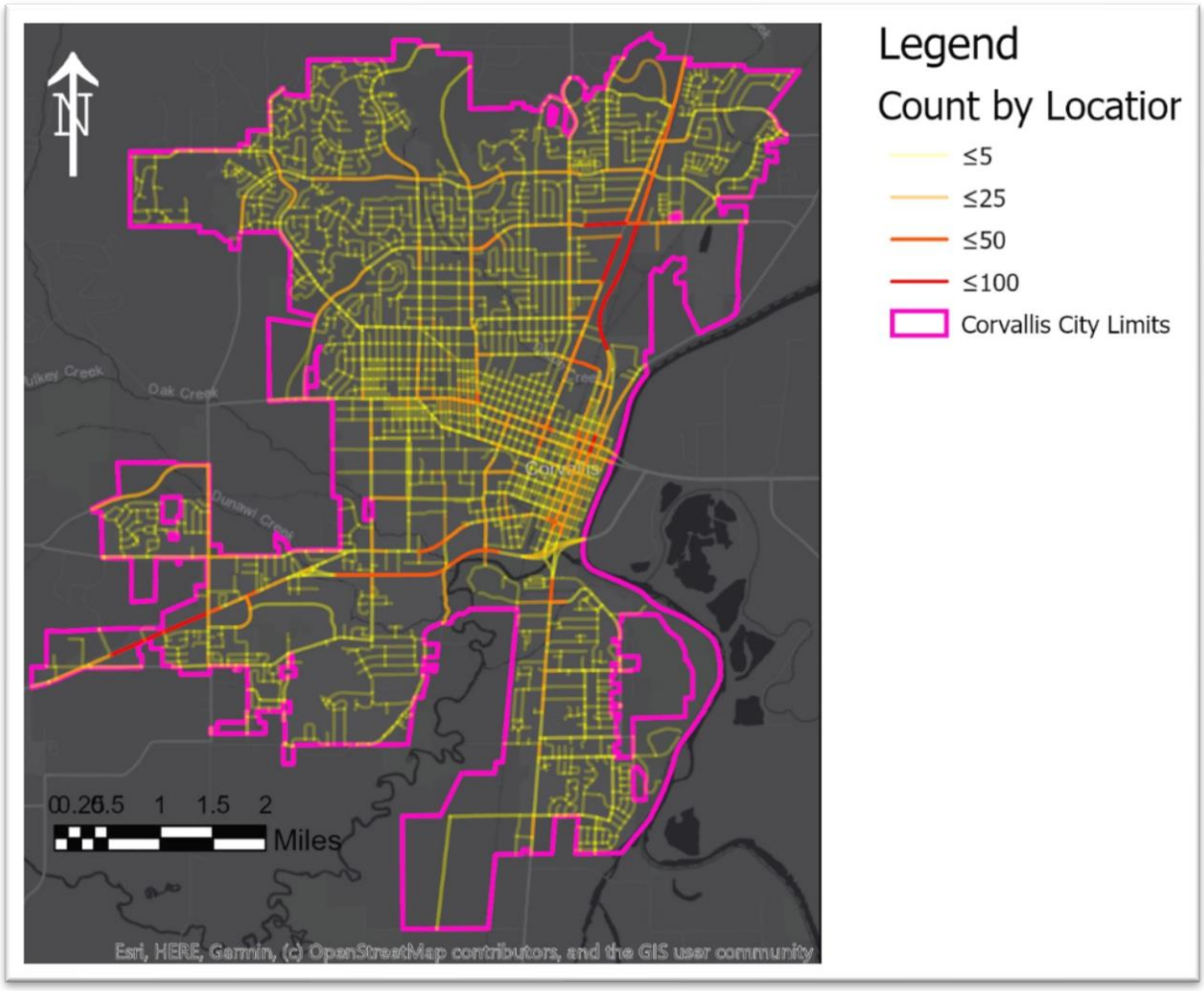


Figure 8. Corvallis, Oregon - Count by Lines of Communication

Analysis of Collision Factors

The large number of variables in the crash dataset allowed for the analysis of statistical significance of specific factors. Time factors are known to be significant in Corvallis, as with most cities. Thus, two specific temporal variables: time of day and day of week were explored. Two temporal analyses were completed: a collision percent change and a weekday hour of the day analysis.

Collision Percent Change

Using the ESRI Incident Analysis toolbox, the percent change of collisions 2007-2017 was determined. The tool required two time periods. By breaking up the eleven years of collision data into two equal halves: January 2007 to June 2012 and July 2012 to December 2017, the percent change for Corvallis during these periods was found. This is shown in figure 9 below. The areas shown are the Corvallis Emergency Medical Service (EMS) Areas. This was chosen to illustrate the change in demand by collisions for each respective area. All changes were increases, with the OSU area and South Corvallis seeing the greatest increases.

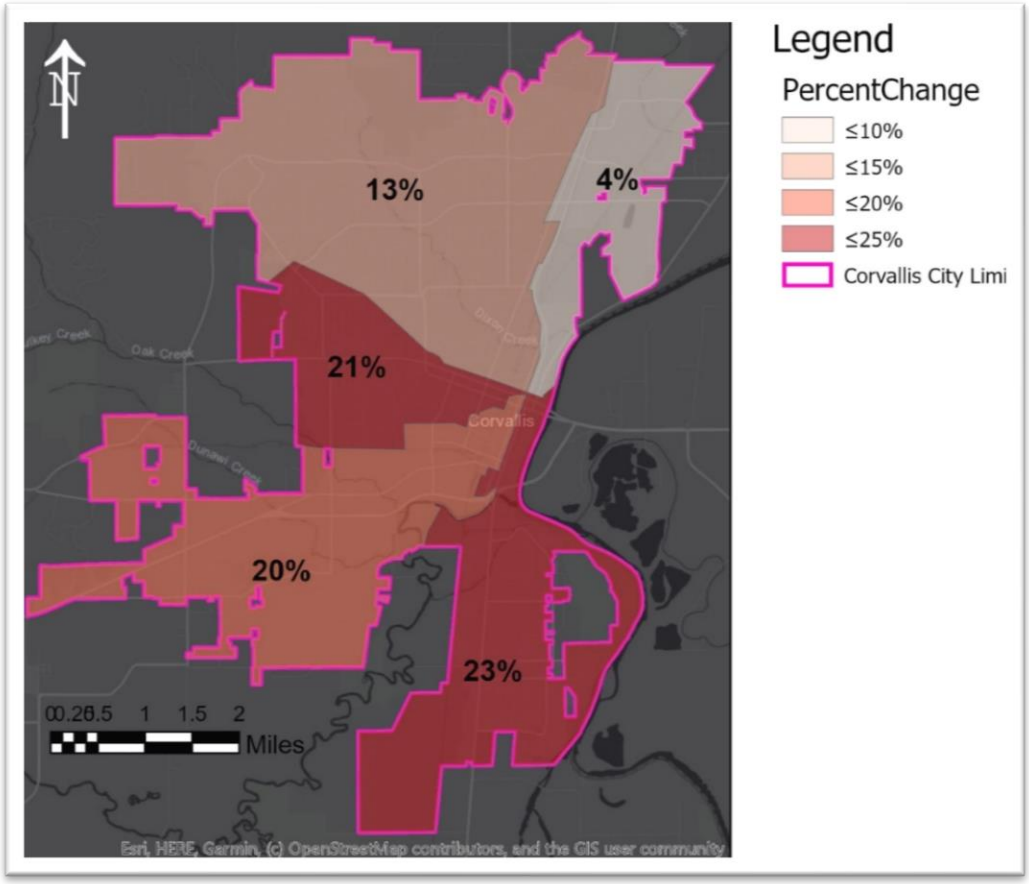


Figure 9. Corvallis, Oregon - Collision Percent Change 2007-2017

Time of Day Analysis

To determine the frequency of weekday collisions in Corvallis, the dataset needed to be filtered. The were 5,812 collisions in total in the dataset. 5,785 collisions had date and time associated to the collision, roughly 99.5% of all collisions. From this subset, only collisions occurring on weekdays was used to show general traffic leveling in the city. A simple count was completed by day of the week. This is shown in figure 10 below. The analysis spikes in collision frequency near morning and evening commutes and the pickup and drop off times for the schools.

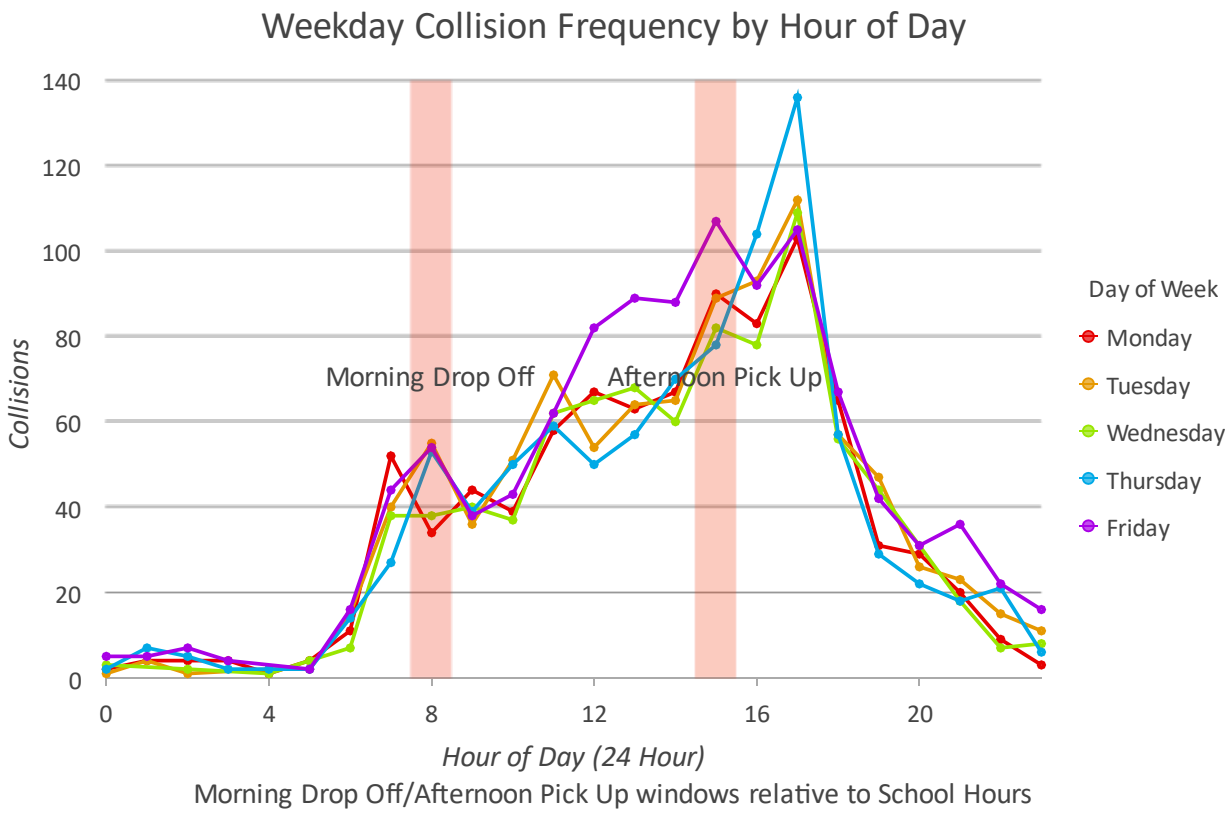


Figure 10. Collision Frequency in the city of Corvallis, Oregon 2007-2017

Determining Statistical Significance of Collision Factors

A binary logit model was run in R, an opensource programming language and software environment for statistical computing and graphics. No graphical representation of the statistics were used. Factors were broken down into five categories: weather, events, impairment, geometry, and special factors. Percent of total collision records are given in parentheses. Fisher Scoring iterations gives the number of times the model was iterated to reach a conclusion.

Weather factors modeled were: clear days with no weather impacts on the road (66%), inclement days with weather impacts on the road (6.9%), and days with weather impacts on the road (24.6%). Of these factors, any day with weather impacts on the road was found to be significant as shown in figure 11 below. 78 records could not be used in this model due to unknown weather conditions.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2195  -0.0513  -0.0513  -0.0513   3.6423

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -21.566   1204.502  -0.018  0.98571
Clear[T.1]     14.934   1204.502   0.012  0.99011
Impacted[T.1]  12.518   1204.502   0.010  0.99171
Wet[T.1]        2.918      1.107   2.637  0.00836 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.45  on 5811  degrees of freedom
Residual deviance: 128.56  on 5808  degrees of freedom
AIC: 136.56

Number of Fisher Scoring iterations: 20
```

Figure 11. R binary logit model - weather factors

The only events considered for this project were the OSU football home games. Dates for each home game was found (Wikipedia, 2020) and cross-referenced with the crash database. The entire date was considered as a game day, with no distinction may for time of day. Game days (2.1%) were found to be mildly significant in this model as shown below in figure 12.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1273  -0.0531  -0.0531  -0.0531   3.6240

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.5653     0.3538 -18.556   <2e-16 ***
Game          1.7531     1.0646   1.647   0.0996 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.45  on 5811  degrees of freedom
Residual deviance: 132.69  on 5810  degrees of freedom
AIC: 136.69

Number of Fisher Scoring iterations: 9
```

Figure 12. R binary logit model - event factors

Impairment factors considered were drugs (0.3%) and alcohol (2.0%). The legal limit for alcohol in the state of Oregon is 0.8 blood alcohol content (BAC). Determination that a collision occurred under the influence of drugs may be dependent on officer observation. Alcohol was found to be strongly significant in this model as shown below in figure 13.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2468  -0.0421  -0.0421  -0.0421   4.6010

Coefficients:
                    Estimate Std. Error z value    Pr(>|z|)
(Intercept)          -7.0304     0.4475 -15.710     < 2e-16 ***
Alcohol.Involved.Flag   3.5540     0.6772   5.248 0.000000154 ***
Drug.Involved.Flag     -2.6666     6.0795  -0.439       0.661
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.45  on 5811  degrees of freedom
Residual deviance: 116.24  on 5809  degrees of freedom
AIC: 122.24

Number of Fisher Scoring iterations: 9
```

Figure 13. R binary logit model - Impairment factors

Geometry factors considered were: arterial road type (73.3%), collector road type (14.5%), local road type (11.8%), intersection (56.8%), driveway/alley (10.3%), and straight roadway section (30.2%). This model may have been influenced by correlation error. Intersection was the only factor found to be significant as shown below in figure 14.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1538  -0.0401  -0.0401   0.0000   3.7755

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            3.545e+08  1.497e+13   0.000   1.0000
Arterial[T.1]         -3.545e+08  1.497e+13   0.000   1.0000
Collector[T.1]        -3.545e+08  1.497e+13   0.000   1.0000
Driveway.Alley[T.1]   -1.860e+01  3.066e+03  -0.006   0.9952
Intersection[T.1]     -2.696e+00  1.230e+00  -2.192   0.0284 *
Local[T.1]            -3.545e+08  1.497e+13   0.000   1.0000
Straight.Roadway[T.1] -9.552e-01  1.086e+00  -0.880   0.3791
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.45  on 5811  degrees of freedom
Residual deviance: 120.04  on 5805  degrees of freedom
AIC: 134.04

Number of Fisher Scoring iterations: 22
```

Figure 14. binary logit model - geometry factors

Special factors considered in this model were: hit and run collisions (1.5%), school zone (0.2%), and work zone (0.4%). None of these factors were found to be significant as shown below in figure 15.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1299  -0.0772  -0.0772  -0.0772   3.4108

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -5.8137     0.3338 -17.415   <2e-16 ***
Hit.and.Run.Flag            -1.0434     2.5768  -0.405    0.686
School.Zone.Indicator[T.1] -11.5157  1571.8758  -0.007    0.994
Work.Zone.Indicator[T.1]   -11.5788  1371.0022  -0.008    0.993
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 122.93  on 3065  degrees of freedom
Residual deviance: 122.67  on 3062  degrees of freedom
  (2746 observations deleted due to missingness)
AIC: 130.67

Number of Fisher Scoring iterations: 16
```

Figure 15. binary logit model - special factors

The final model used the significant factors found in each category to determine the overall significance of the factors. The result was that alcohol is the most significant factor in the model shown in figure 16 below. Further analysis of alcohol, OSU game days and weather impacted roads was completed. Due to the variable nature of the latter two factors, both may be considered to be significant when present in Corvallis.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7085  -0.0475  -0.0357  -0.0247   4.3169

Coefficients:
                     Estimate Std. Error z value   Pr(>|z|)
(Intercept)           -6.7859     0.5715 -11.873    < 2e-16 ***
Alcohol.Involved.Flag  3.2685     0.6927   4.719 0.00000237 ***
Game                   1.5267     1.1295   1.352      0.176
Intersection[T.1]     -1.3104     0.8141  -1.610      0.107
Wet[T.1]               0.7366     0.6887   1.070      0.285
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.45  on 5811  degrees of freedom
Residual deviance: 110.56  on 5807  degrees of freedom
AIC: 120.56

Number of Fisher Scoring iterations: 10
```

Figure 16. binary logit model - final model factors

Factor Counts by Lines of Communication

Alcohol collisions make up 2.0% of all collisions in Corvallis. These collisions are centralized on Monroe Avenue north of OSU and Highway 99W north of downtown and south of NW Circle Boulevard as shown below in figure 17.

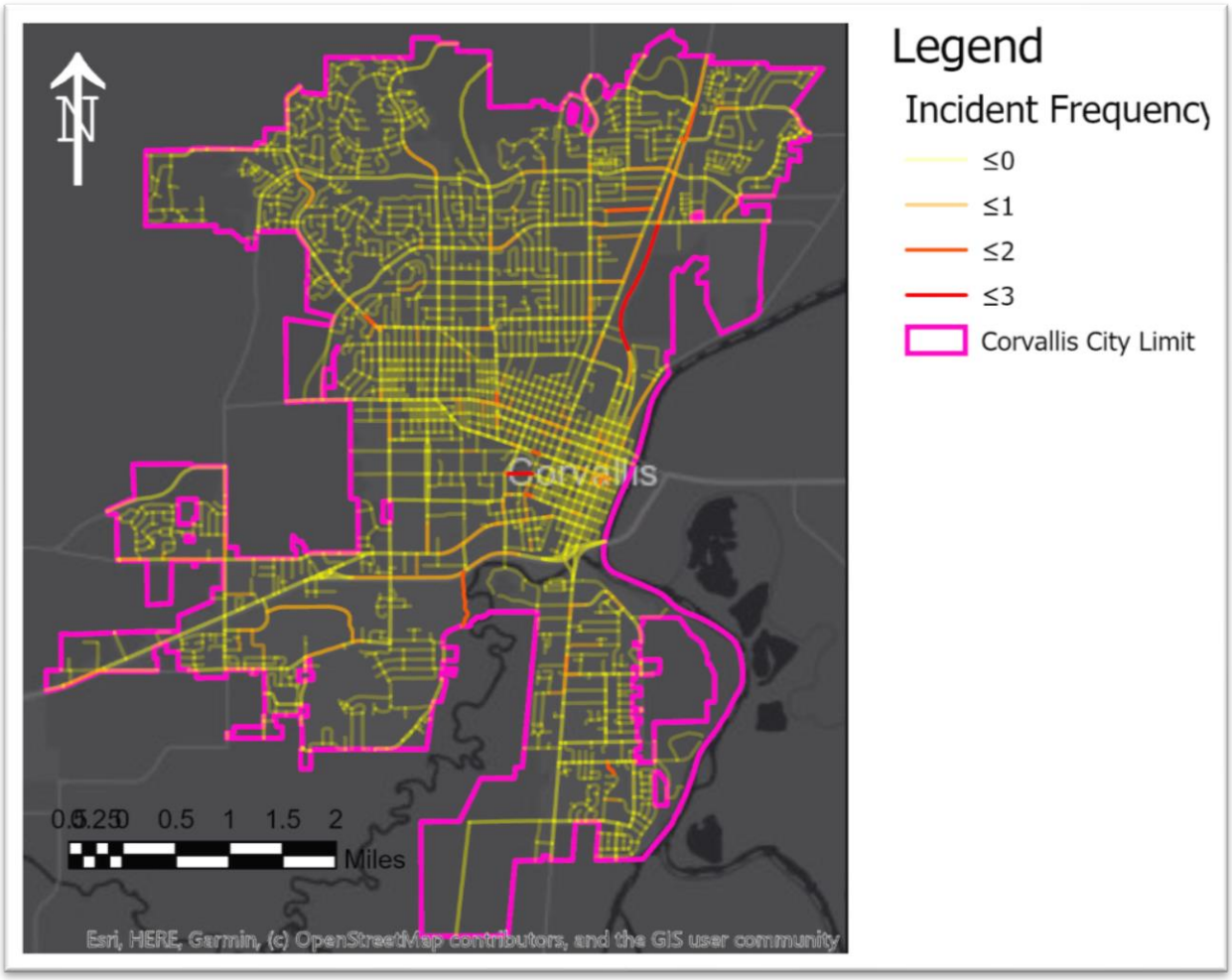


Figure 17. Alcohol Related Lines of Communication

Collisions where the roadway is impacted by weather make up 24.6% of all collisions in Corvallis. Significant locations are shown at Highway 99W near the Van Buren Bridge and south of the Highway 20/Highway 99W interchange as shown in figure 18 below.

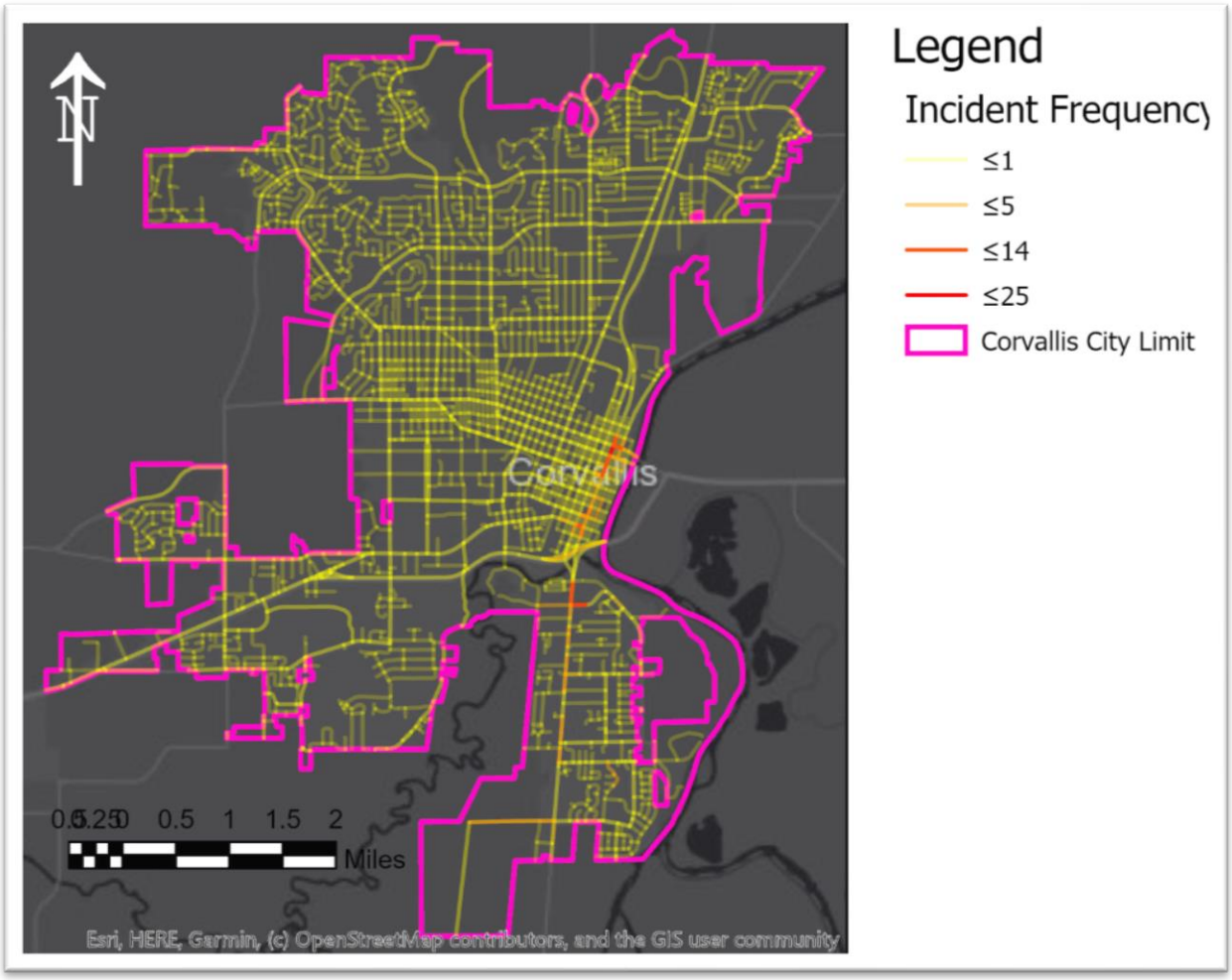


Figure 18. Weather Related Lines of Communication

OSU football home game day collisions make up 2.1% of all collisions in Corvallis. Significant locations are at Highway 20 near Avery Park and west of SW 53rd Street, and NW 9th Street near NW Circle Boulevard. These are shown in figure 19 below. The relatively higher number of collisions in these areas may be due to limited travel options to enter of leave Corvallis. This project does not consider data exterior to the Corvallis city limits. Significant data to the east of Corvallis near Van Buren Bridge may have been neglected in this model.

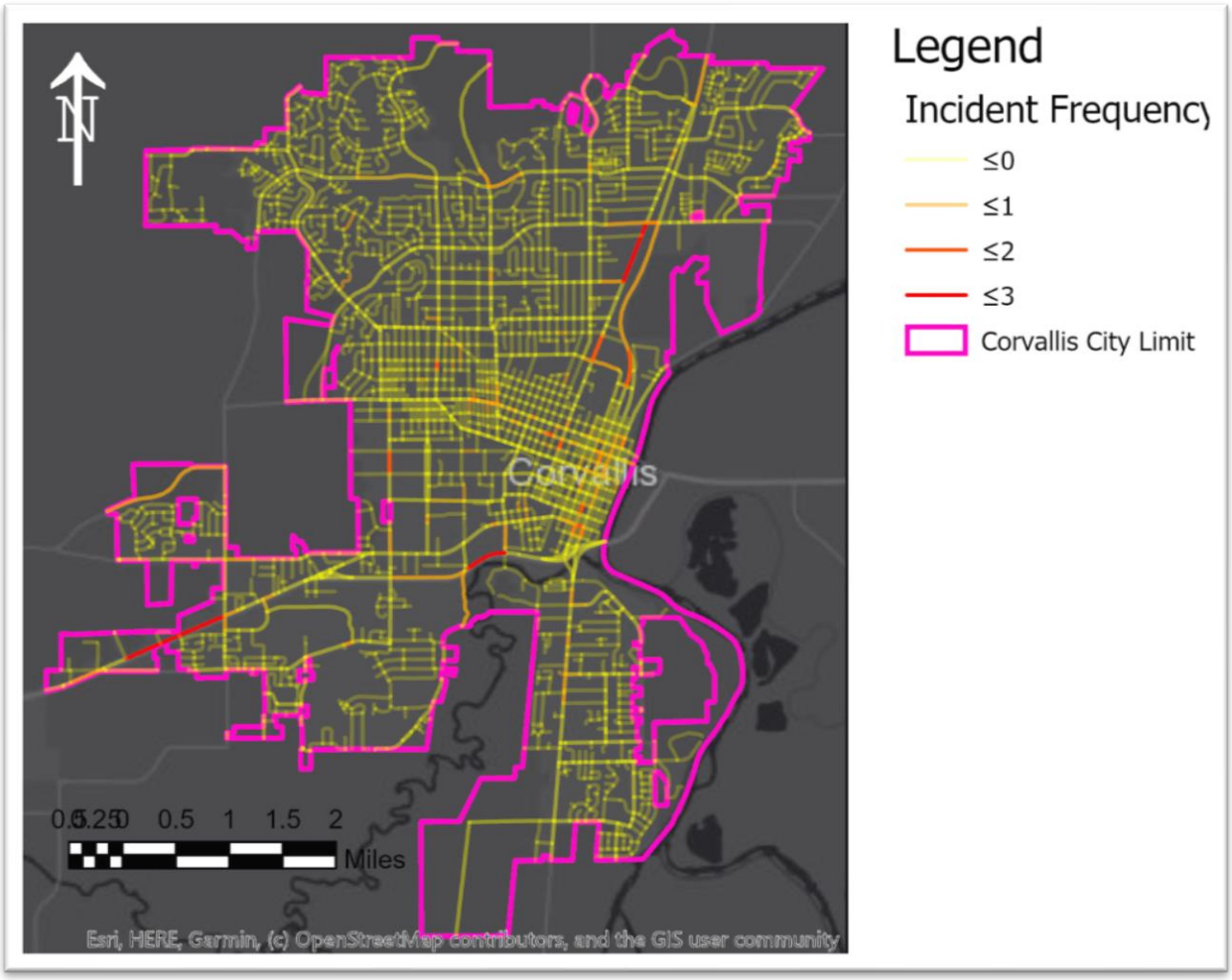


Figure 19. OSU football home game days lines of communication

Potential Errors and False Positives

Potential errors begin in the data used for the analysis. There is no way to confirm the validity of the data entries. Crash data in particular has as wide variability when entered from in the field accounts. Some data was excluded from the time of day analysis due to the lack of a known time of day. Some data was excluded from the weather analysis due to unknown conditions. One location in particular is a likely false positive in the City of Corvallis Collision Trends. This is due to the geolocation listed not agreeing with the latitude and longitude. Thus, it appears on the

map in a location that does not agree with its description. This falsely increases the statistical significance of this point due to the location never seeing other crashes.

Analyses in ArcGIS Pro may also be prone to errors due to oversimplification of key statistics and relationships. Emerging Hot Spot Trends uses the assumption of neighboring bins which will not consider any geographical barriers. This is especially egregious when used in transportation applications as the vast majority of collisions occur on the roadway which is geographically bound to a corridor. Hot Spot Analysis (Getis-Ord Gi*) uses a weight matrix file, which may distort the significance of collisions based on the given weights. In this analysis the euclidian distance used suffers from the same failing as the Emergin Hot Spot Trend. This is remediated by using a distance of 15 feet which is only slightly wider than the given roadway. This reduces error, but also reduces the significance of the analysis by not allowing it to take corridors into account.

Binary logit models make multiple assumptions which can result in errors. The variables must be discrete and dichotomous. Either they exist or do not. If the factors used do not comply with this assumption, then the results may be invalid. Regression models are vulnerable to outlier data such as may be present due to large sporting events that occur regularly in Corvallis. Models can suffer from over fitting if too many variables are used at the same time. This will result in an inaccurate result that distorts the best fit model. This also may result in R^2 values that are suspect. No confusion matrix was used in this model. This would reduce correlation errors which may occur when correlated factors are demonstrated in the same model. Correlation errors may result in invalid models and results.

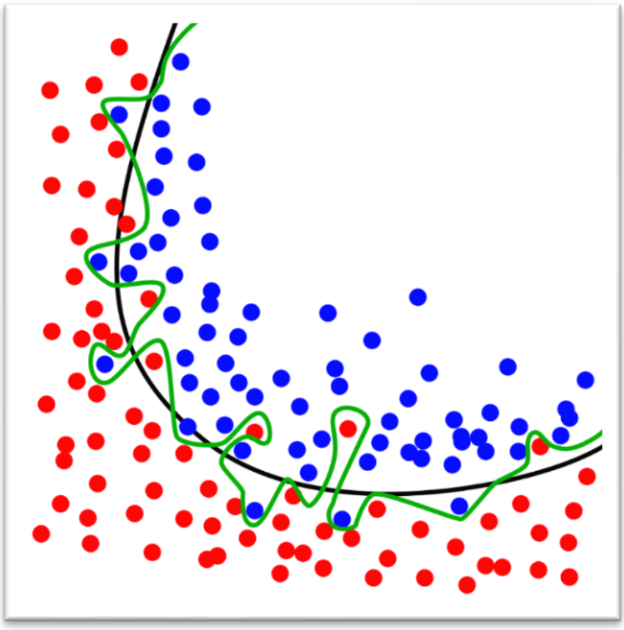


Figure 20. Example of Over Fitting (Elite Data Science, 2019)

https://elitedatascience.com/overfitting-in-machine-learning

Conclusion

Collision analysis may be completed for the state of Oregon using statistical analysis in R and ArcGIS Pro analysis software. The collision analysis for Corvallis, Oregon has resulted in reasonable conclusions that fit with expected results. Increases in the collision rate north of OSU and in south Corvallis may require remediation to reduce the impact of this increase. Alcohol related collisions on Monroe Avenue pinpoint a key location which may benefit from increased enforcement in the area. Weather events which may result in higher collision rates on 99W could be reduced by increasing enforcement, or considering a variable speed limit based on conditions. OSU football home game days may benefit from alternate routes and temporary traffic control measures. These recommendations would be appropriate to incorporate into the TSP plan for Corvallis, Oregon.

References

Elite Data Science. (2019, March 12). *Overfitting in Machine Learning: What It Is and How to Prevent It*. Retrieved from Elite Data Science: https://elitedatascience.com/overfitting-in-machine-learning

ESRI. (2020, Februrary 28). *Emerging Hot Spot Analysis (Space Time Pattern Mining)*. Retrieved from ArcGIS Pro: https://pro.arcgis.com/en/pro-app/tool-reference/space-time-pattern-mining/emerginghotspots.htm

ESRI. (2020, February 28). *Hot Spot Analysis (Getis-Ord Gi*)*. Retrieved from ArcGIS Desktop: https://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/hot-spot-analysis.htm

ESRI. (2020, February 28). *North American datums.* Retrieved from ArcGIS for Desktop: https://desktop.arcgis.com/en/arcmap/10.3/guide-books/map-projections/north-american-datums.htm

GISGeography. (2020, February 28). *Conic Projection: Lambert, Albers and Polyconic*. Retrieved from GISGeography: https://gisgeography.com/conic-projection-lambert-albers-polyconic/

Le, J. (2019, March 12). *Logistic Regression in R Tutorial*. Retrieved from DataCamp: https://www.datacamp.com/community/tutorials/logistic-regression-R

Office of Planning, Federal Highway Administration. (2012). *BEST PRACTICES IN GEOGRAPHIC INFORMATION SYSTEMS-BASED.* Washington, D.C.: U.S. Department of Transportation.

Oregon Department of Transportation. (2020, March 12). *Index of /tdb/transdata/GIS_data/*. Retrieved from ODOT TransGIS Data: ftp://ftp.odot.state.or.us/tdb/trandata/GIS_data/

United States Census Bureau. (2020, March 12). *Population and Housing Occupancy Status 2010*. Retrieved from American Fact Finder: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=DEC_10_PL_GCTPL2.ST10&prodType=table

Wikipedia. (2020, March 12). *Oregon State Beavers football*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Oregon_State_Beavers_football

Zhu, L., Yu, F., Wang, Y., B., N., & Tang, T. (2018). Big Data Analytics in Intelligent Transportation Systems: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 383-398.