Exploratory Spatial Data Analysis within a GIS to Identify Statistically Significant Clusters of Childhood Pedestrian Collisions

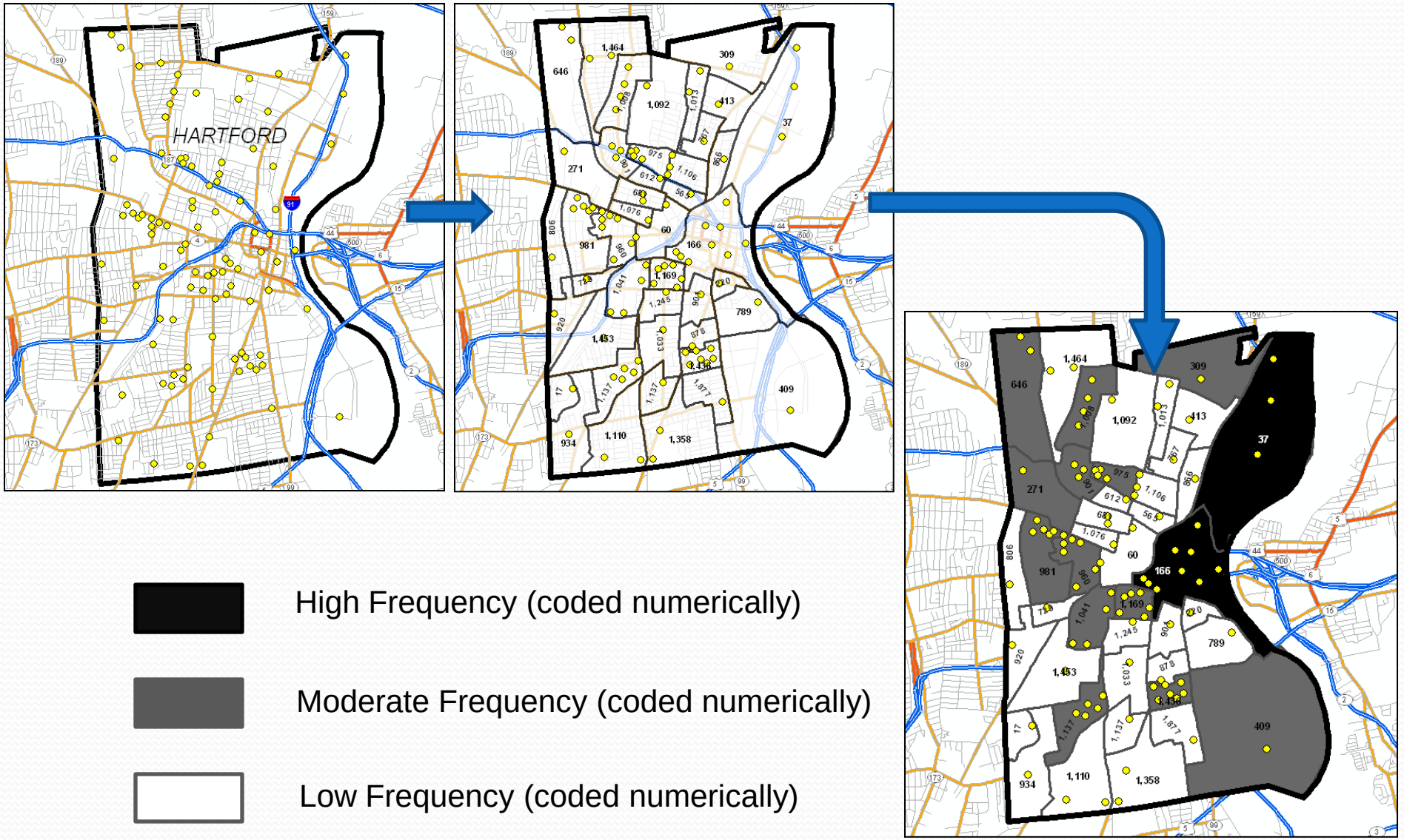
Peter Hayward, PhD

Kevin Borrup, JD, MPA
Steven Rogers, MD
Garry Lapidus, PA-C, MPH

Annual Meeting and Exposition of the American Public Health Association (APHA), November 9, 2009

A Common Approach…

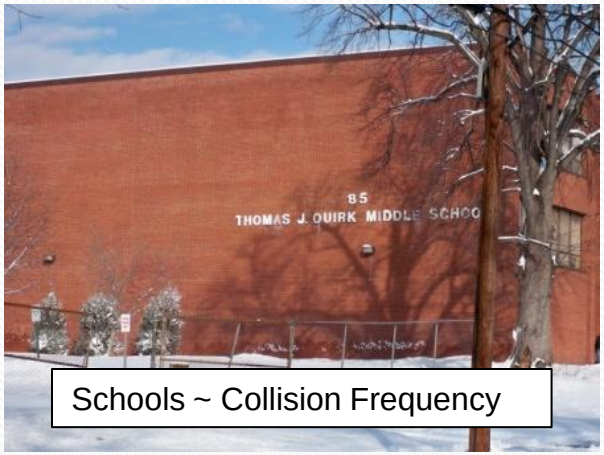
- Identify Locations and Frequencies of Collisions



A Common Approach…

• Identify Correlating Factors and Formulate Hypotheses

Poverty ~ Collision Frequency



Schools ~ Collision Frequency

Parks ~ Collision Frequency

Police ~ Collision Frequency

Housing ~ Collision Frequency

➢ Context of high frequency neighborhoods versus moderate and low frequency areas.

A Common Approach…

- Perform Regression Analysis and Evaluate Results

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} + e_i$$

WHERE:

y_i	Frequency of collision
b_0	Estimate of numerical constant
b_k	Numerical coefficients
x_{ik}	Values of k independent variables
e_i	Estimate of error (residual) in predicting frequency of collision

A Common Approach…

- Questions?

- Is this an appropriate investigation method into the spatial context of childhood pedestrian collisions?

- What problems may arise from using this method?

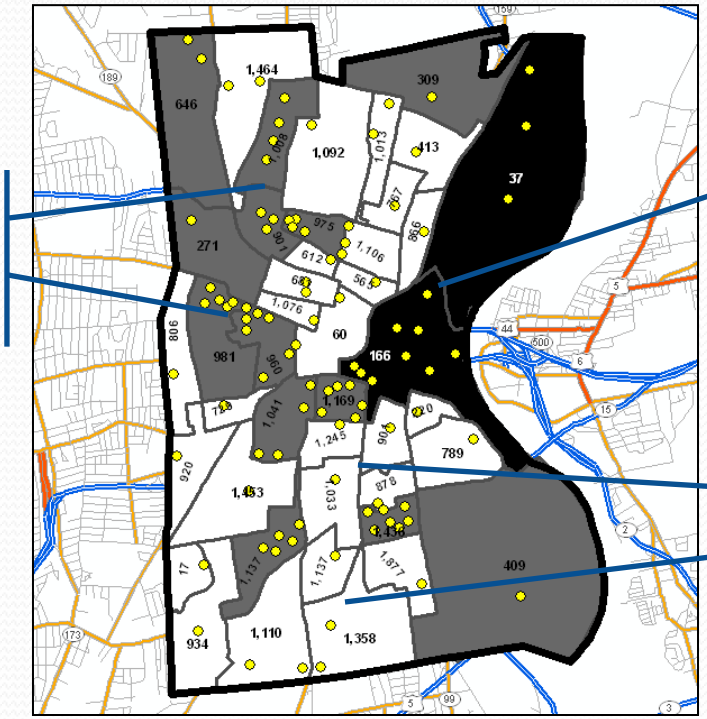
- What can be done to resolve these issues?

A Common Approach…

• Problem and Solution: Spatial Autocorrelation

- Problem: Grouping neighborhoods based on frequency of collision and neighborhood context begins to take the “space” out of a “spatial” analysis.

Problem Spots? Large number of collisions and large populations.



Problem Spots? Large number of collisions but small populations.

Ideal locations? Few collisions despite large populations.

- Solution: Taking underlying population values into account, identify adjacent locations (clusters) that have statistically significant high or low values.

A Common Approach…

• Problem and Solution: Spatial Autocorrelation

- Problem: The value of the dependent variable at one location can be impacted by the values of the independent variables at that location and at neighboring locations (spatial lag). The residuals across spatial units may also be autocorrelated (spatial error). These forms of spatial autocorrelation may lead to inflated coefficients and/or a better model fit.

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} + e_i$$

Inaccurate R^2

Inflated

- Solution: First, detect any spatial lag or spatial error and then account for spatial autocorrelation in the regression analysis.

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} + e_i$$

Logit [Probability(Outcome)]

Add more variables

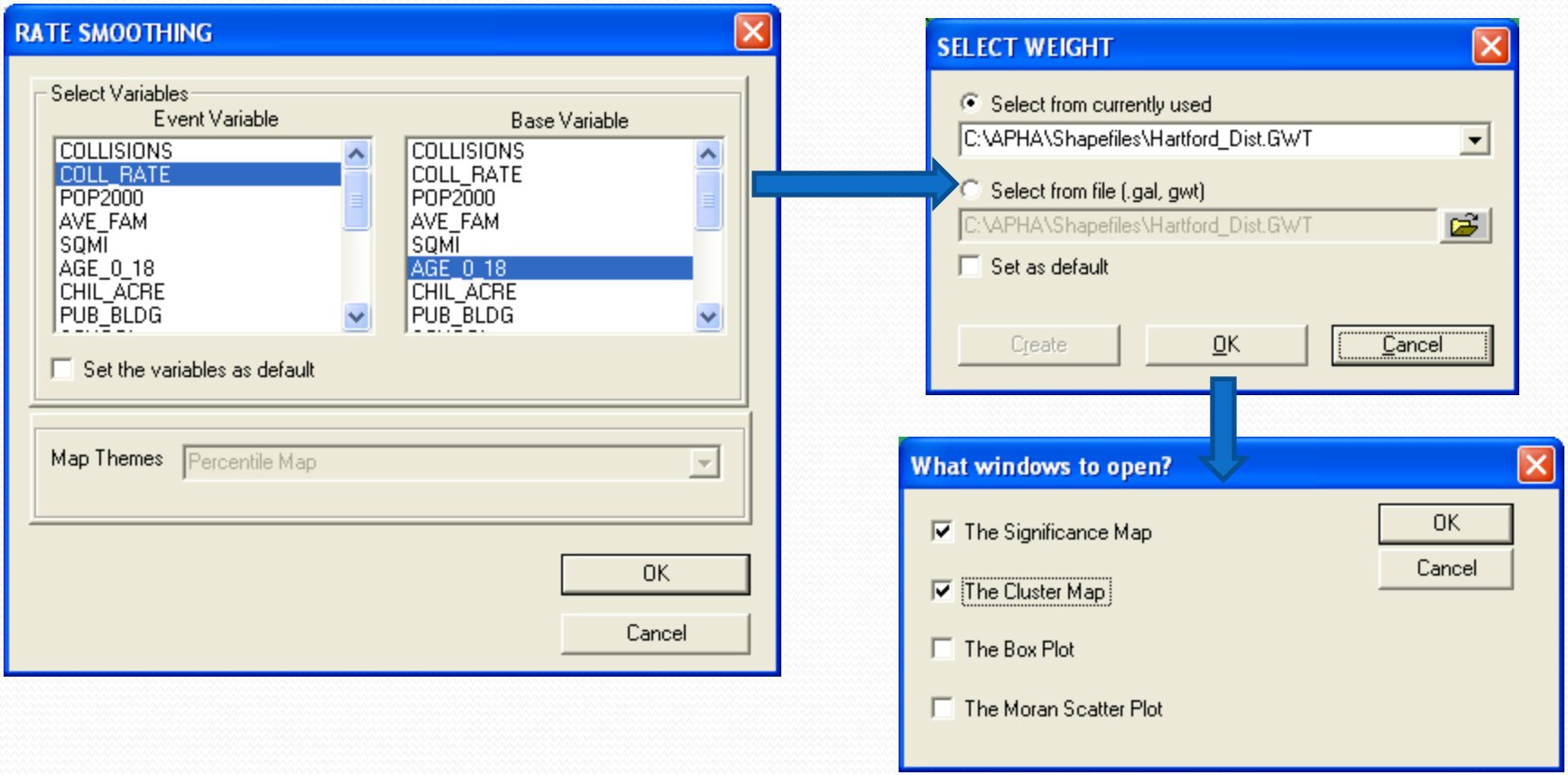
Example

• ESDA and Childhood Pedestrian Collisions in Hartford, CT

- ESDA
  - Exploratory spatial data analysis (ESDA) is a sophisticated method for detecting spatial patterns in data (clusters) meant for incorporation into statistical models.
- Objective
  - Demonstrate the utility of ESDA in exploring the impact of specific variables on childhood pedestrian collisions in Hartford, CT.
- Data
  - Dependent Variable: frequency of childhood pedestrian collisions determined by Hartford census tract, 2005 – 2006.
  - Independent Variables: contextual variables by census tract (poverty, education, schools, public buildings, housing characteristics, children per square mile).
- Methods
  - Model statistically significant clusters using GeoDA.
  - Perform regression analysis when accounting for spatial autocorrelation using GeoDA.

Example

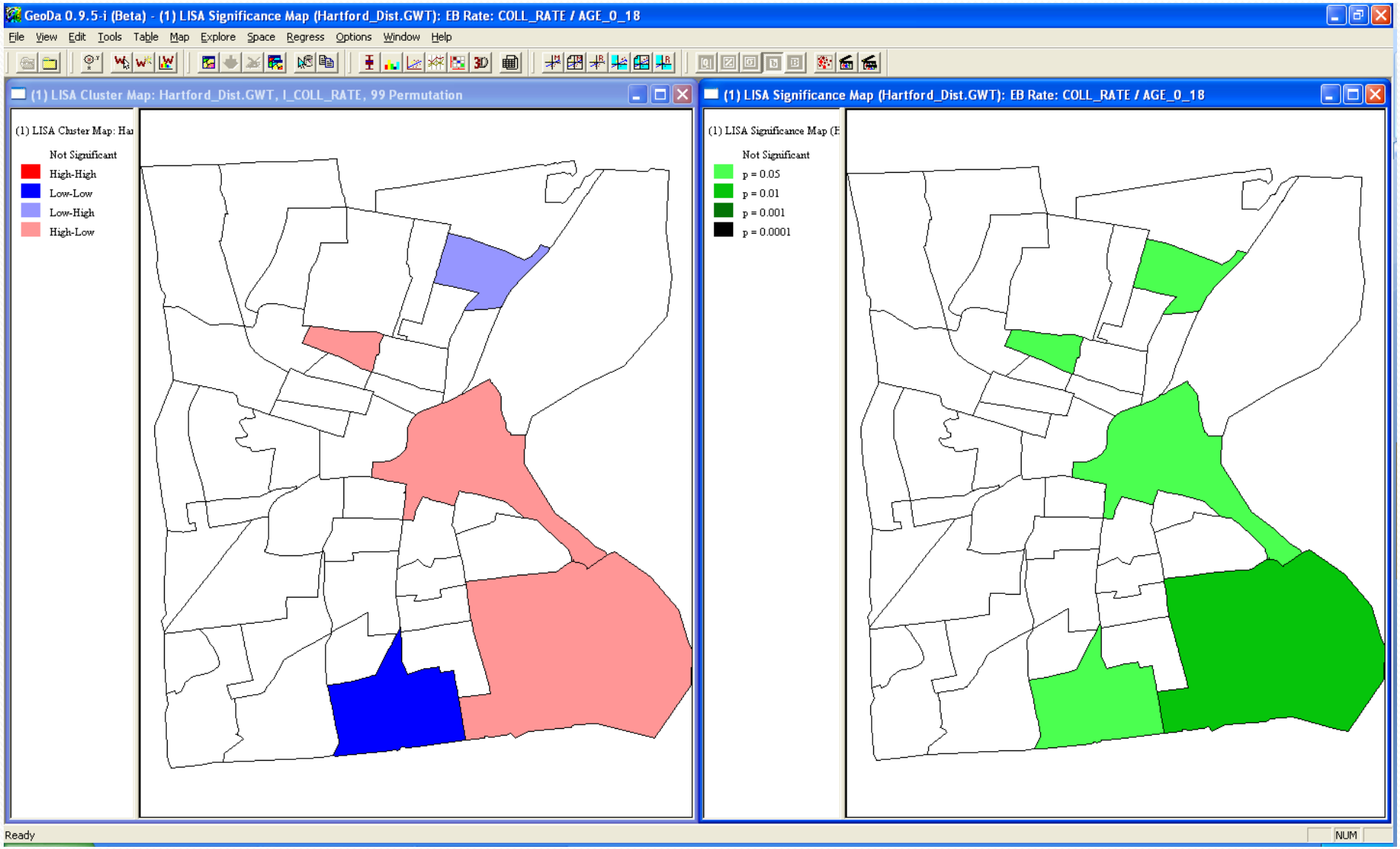
• ESDA and Childhood Pedestrian Collisions in Hartford, CT



- Taking underlying population values into account, use a local indicator of spatial autocorrelation (LISA) to determine statistically significant clusters.

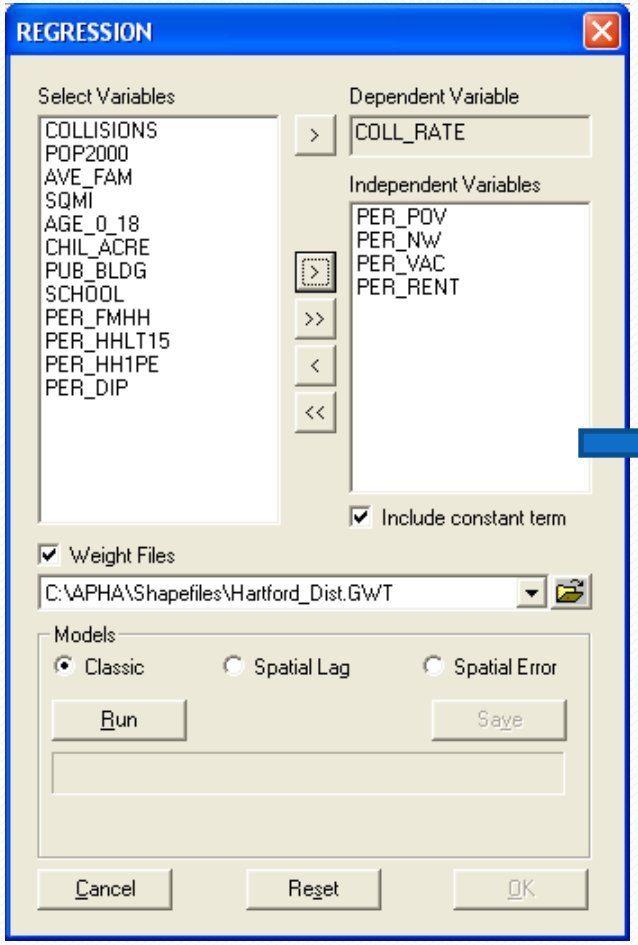
Example

- ESDA and Childhood Pedestrian Collisions in Hartford, CT



Example

• ESDA and Childhood Pedestrian Collisions in Hartford, CT

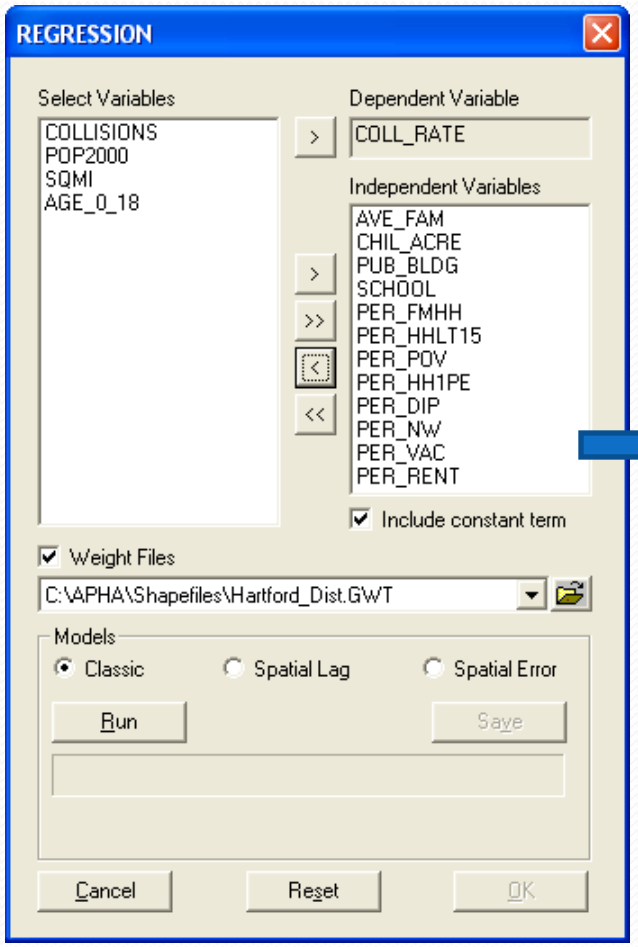



```
REGRESSION--SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Dependent Variable  :    COLL_RATE  Number of Observations:     43
Mean dependent var  :     0.473939  Number of Variables    :      5
S.D. dependent var  :      1.43551  Degrees of Freedom     :     38

R-squared           :     0.331996  F-statistic            :   4.72147
Adjusted R-squared  :     0.261680  Prob(F-statistic)      :  0.0034245
Sum squared residual:      59.1919  Log likelihood         :   -67.8854
------------------------------------------------------------------------
    Variable     Coefficient      Std.Error    t-Statistic    Probability
------------------------------------------------------------------------
    CONSTANT        4.838322       1.123474       4.306571     0.0001126
     PER_POV      0.02044269      0.02637677      0.7750264     0.4431200
      PER_NW     -0.02805238       0.0133631      -2.099241     0.0424925
     PER_VAC      0.04840837      0.03804521       1.272391     0.2109653
    PER_RENT     -0.05356087      0.01721634      -3.111049     0.0035290
------------------------------------------------------------------------


DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : Hartford_Dist.GWT  (row-standardized weights)
TEST                              MI/DF        VALUE           PROB
Moran's I (error)              0.169948    2.0130461      0.0441096
Lagrange Multiplier (lag)             1    8.3820018      0.0037895
Lagrange Multiplier (error)           1   10.3944463      0.0012639
```

- Perform regression analysis; diagnostics indicate need to account for spatial autocorrelation.

Example

• ESDA and Childhood Pedestrian Collisions in Hartford, CT




```
REGRESSION--SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Dependent Variable  :   COLL_RATE  Number of Observations:   43
Mean dependent var  :     4.7394  Number of Variables   :   13
S.D. dependent var  :    14.3551  Degrees of Freedom    :   30

R-squared           :   0.838345  F-statistic           :    12.9651
Adjusted R-squared  :   0.773684  Prob(F-statistic)     :9.18221e-009
Sum squared residual:    1432.42  Log likelihood        :    -136.392
```

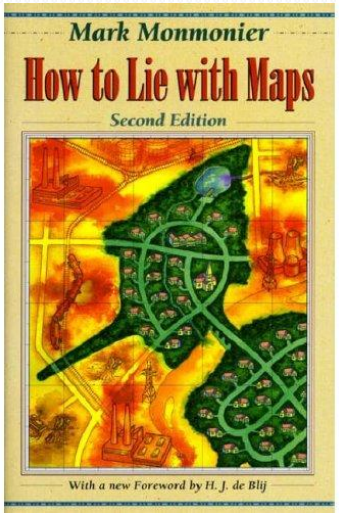
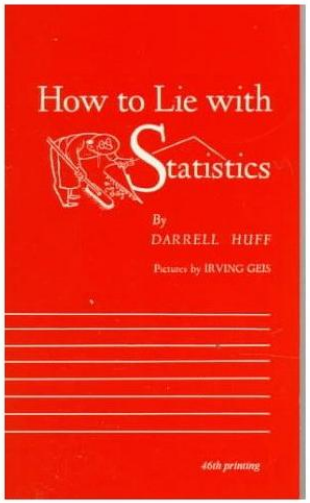
Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	88.35835	14.17169	6.234847	0.0000007
AVE_FAM	-10.42038	2.726431	-3.821986	0.0006213
CHIL_ACRE	0.520912	0.4133146	1.260328	0.2172663
PUB_BLDG	1.429644	0.297601	4.803895	0.0000405
SCHOOL	-1.771134	1.117257	-1.585252	0.1233952
PER_FMHH	-0.495804	0.1911551	-2.593726	0.0145416
PER_HHLT15	-0.2523631	0.1935513	-1.303857	0.2021990
PER_POV	0.4080984	0.2111391	1.932841	0.0627441
PER_HH1PE	0.1390817	0.3353025	0.4147947	0.6812439
PER_DIP	-0.3686721	0.18302	-2.014381	0.0530112
PER_NW	0.106922	0.1248146	0.8566461	0.3984321
PER_VAC	-0.124795	0.2996131	-0.4165204	0.6799950
PER_RENT	-0.4287531	0.1105455	-3.87852	0.0005326

```
DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : Hartford_Dist.GWT  (row-standardized weights)
```

TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.072868	1.3474983	0.1778199
Lagrange Multiplier (lag)	1	0.5704315	0.4500875
Lagrange Multiplier (error)	1	0.3700903	0.5429550

- Perform regression analysis; diagnostics indicate no need to account for spatial autocorrelation.

Conclusion

• ESDA and Childhood Pedestrian Collisions in Hartford, CT

- Spatio-Temporal Analysis
  - ESDA could be used to investigate the changes in the number and location of high versus low clusters over time.
  - ESDA could be used to investigate the spatio-temporal characteristics of childhood pedestrian collisions, which has important ramifications for injury prevention.
- Conclusions
  - ESDA provides validity to spatial patterns using statistical significance.
  - ESDA provides a more robust examination of statistical relationships.