Multivariate Regression Analysis to Determine the Influence of Factors on Land Value at the Neighborhood Scale

Alan Dittmar
Aidan Padilla
Mark Cooney

INTRODUCTION

- Rule #1 in Real Estate:

  Location, Location, Location!

- What does this mean? Is location still a factor when looking in the same region, city, down to the neighborhood level?

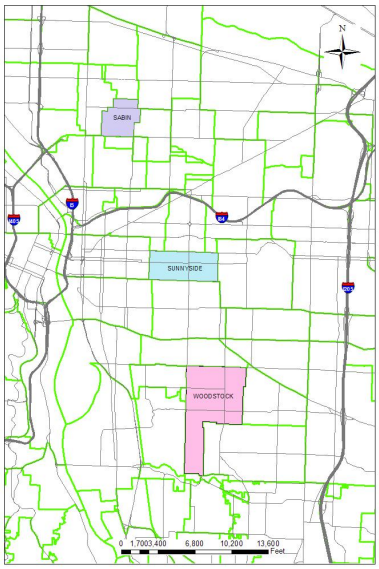


OBJECTIVE

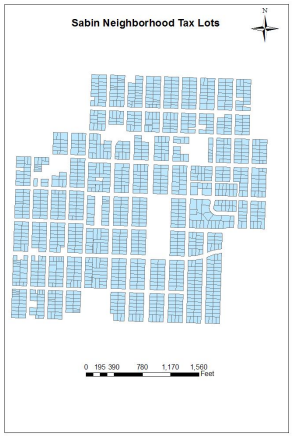
- What variables influence the differences in price of **LAND VALUE** of tax lots in the same general locale (neighborhood)?
- Should lots of the same size be the same price? If not, why?
- Using land values from tax lot data, we are interested in the value differences from individual locations (lots) and not buildings (houses).

- Using regression analysis, could we construct neighborhood models that are representative of the influence of different variables on land value?

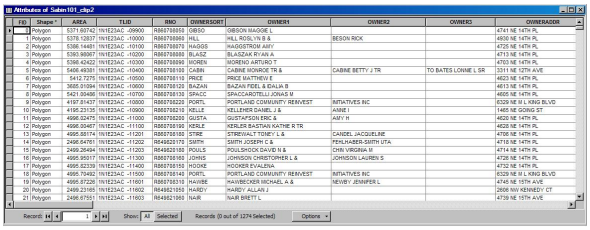
Study Area



Utilizing Tax Lot Data



Sabin Tax Lots Attribute Table

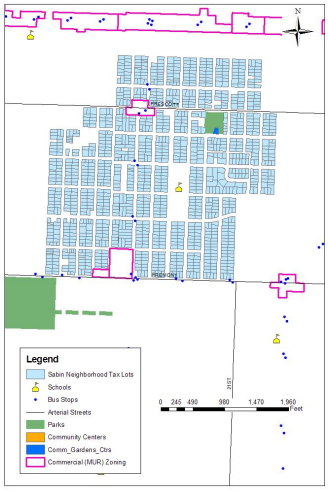


-Property Code
-Area
-Appraised Value (Land Value, Building Value, Total Value)
-Owner
-Address
-Tax Lot ID
-Sale Price
-Etc.

Determining Proximity to Factors

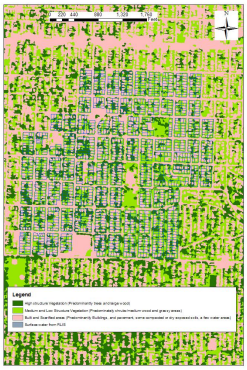
Distance to:
-Bus stops (point)
-Schools (point)
-Arterial Streets(line)
-Commercial Areas (polygon)
-Parks (polygon)
-Community Centers (polygon)
-Community Gardens (polygon)

Near Tool Bug:
-Reports distance squared when calculating distance to point features

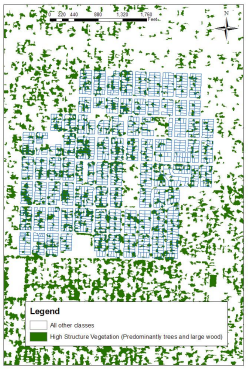


Determining Percent Canopy Cover (Tall Trees)

- Metro's 2007 Vegetation Layer (3 ft resolution)
  - 1) Clip Layer to Lesser Extent
  - 2) Reclassify



Reclassify



Calculating Percent Canopy Cont.

3) Focal Statistics → Focal Mean raster (5x5)

4)Rasterize Tax Lots

5) Zonal Statistics → Zonal Sum raster

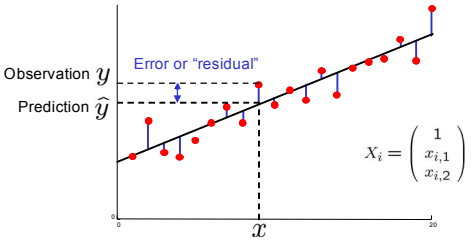
6) Field Calculator →Canopy Cover Ratio (Sum / Count)

7) Join Zonal Sum table to Tax Lots layer

Multivariate Linear Regression (Ordinary Least Squares)

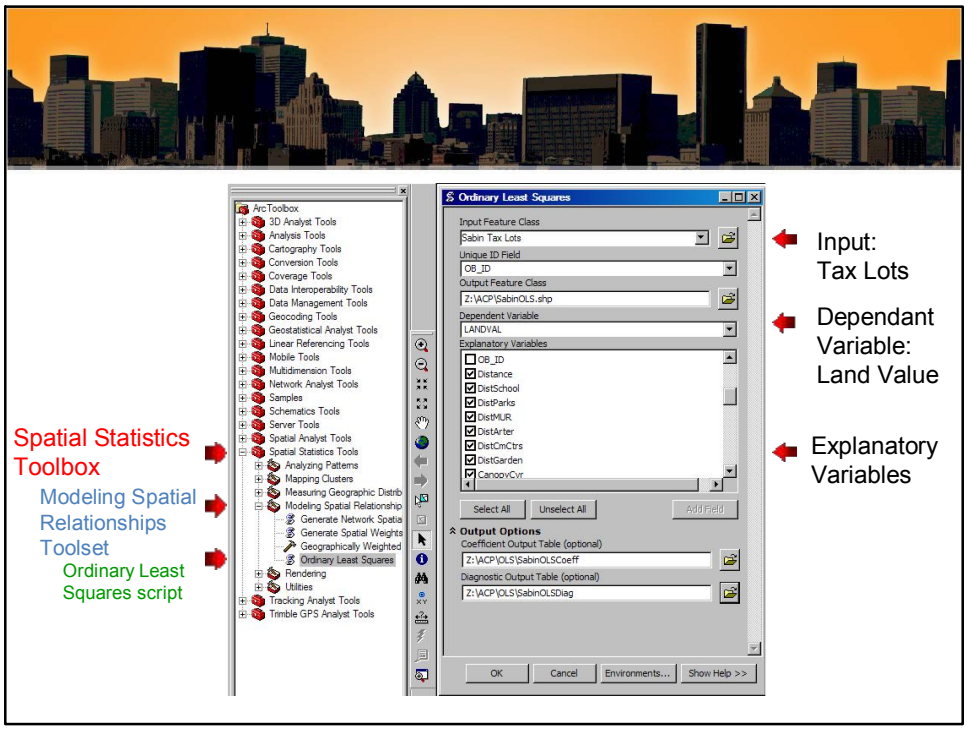
ArcGIS 9.3 Webhelp:
Ordinary Least Squares (OLS) is the best known of all regression techniques. **It is also the proper starting point for all spatial regression analyses.**
It provides a global model of the variable or process you are trying to understand or predict; it creates **a single regression equation** to represent that process.

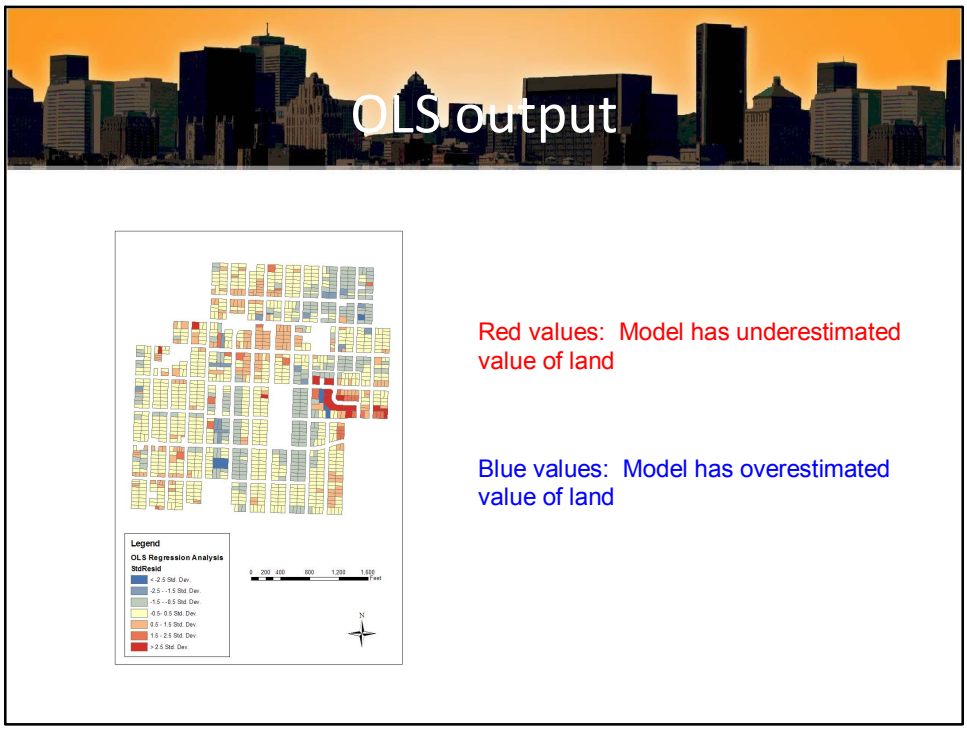
Ordinary Least Squares (OLS)



Sum squared error $\sum_i (X_i^\top w - y_i)^2$

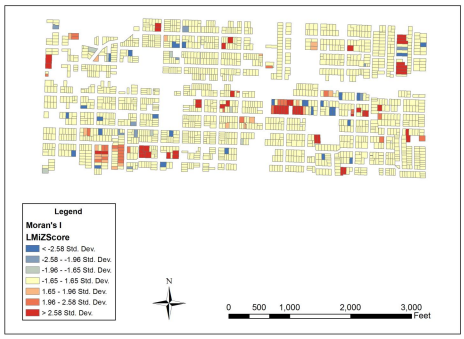
www.eecs.berkeley.edu/~asimma


Spatial Statistics Toolbox
Modeling Spatial Relationships Toolset
Ordinary Least Squares script
ArcToolbox
3D Analyst Tools
Analysis Tools
Cartography Tools
Conversion Tools
Coverage Tools
Data Interoperability Tools
Data Management Tools
Geocoding Tools
Geostatistical Analyst Tools
Linear Referencing Tools
Mobile Tools
Multidimension Tools
Network Analyst Tools
Samples
Schematics Tools
Server Tools
Spatial Analyst Tools
Spatial Statistics Tools
Analyzing Patterns
Mapping Clusters
Measuring Geographic Distrib
Modeling Spatial Relationship
Generate Network Spatia
Generate Spatial Weights
Geographically Weighted
Ordinary Least Squares
Rendering
Utilities
Tracking Analyst Tools
Trimble GPS Analyst Tools
Ordinary Least Squares
Input Feature Class
Sabin Tax Lots
Unique ID Field
OB_ID
Output Feature Class
Z:\ACP\SabinOLS.shp
Dependent Variable
LANDVAL
Explanatory Variables
OB_ID
Distance
DistSchool
DistParks
DistMLR
DistArter
DistCmCtrs
DistGarden
Select All
Unselect All
Add Field
Output Options
Coefficient Output Table (optional)
Z:\ACP\OLS\SabinOLSCoeff
Diagnostic Output Table (optional)
Z:\ACP\OLS\SabinOLSDiag
OK
Cancel
Environments...
Show Help >>
Input: Tax Lots
Dependant Variable: Land Value
Explanatory Variables


OLS output
Legend
OLS Regression Analysis
StdResid
Red values: Model has underestimated value of land
Blue values: Model has overestimated value of land

Examining Evidence of Clustering and Outliers

- Local Moran's I: Uses residual values of OLS output to determine clustering and outliers



Red areas: Residuals are similar - clustering

Blue areas: Residuals are different - outliers

Sabin

Attributes of AlaDiagTable

OID	Field1	Diag_Name	Diag_Value	Definition
0	0	AIC	28962.176837	Akaike's Information Criterion: A relative measure of performance used to compare models; the smaller AIC indicates the superior model.
1	0	R2	0.532409	R-Squared, Coefficient of Determination: The proportion of variation in the dependent variable that is explained by the model.
2	0	AdjR2	0.528707	Adjusted R-Squared: R-Squared adjusted for model complexity (number of variables) as it relates to the data.
3	0	F-Stat	143.807793	Joint F-Statistic Value: Used to assess overall model significance.
4	0	F-Prob	0	Joint F-Statistic Probability (p-value): The probability that none of the explanatory variables have an effect on the dependent variable.
5	0	Wald	1476.518163	Wald Statistic: Used to assess overall robust model significance.
6	0	Wald-Prob	0	Wald Statistic Probability (p-value): The computed probability, using robust standard errors, that none of the explanatory variables have an effect on the dependent variable.
7	0	K(BP)	198.298202	Koenker's studentized Breusch-Pagan Statistic: Used to test the reliability of standard error values when heteroskedasticity (non-constant variance) is present.
8	0	K(BP)-Prob	0	Koenker (BP) Statistic Probability (p-value): The probability that heteroskedasticity (non-constant variance) has not made standard errors unreliable.
9	0	JB	44351.86586	Jarque-Bera Statistic: Used to determine whether the residuals deviate from a normal distribution.
10	0	JB-Prob	0	Jarque-Bera Probability (p-value): The probability that the residuals are normally distributed.
11	0	Sigma2	433234203.344	Sigma-Squared: OLS estimate of the variance of the error term.

Sunnyside

Attributes of OLS_all_diag

OID	Field1	Diag_Name	Diag_Value	Definition
0	0	AIC	40398.109835	Akaike's Information Criterion: A relative measure of performance used to compare models; the smaller AIC indicates the superior model.
1	0	R2	0.446427	R-Squared, Coefficient of Determination: The proportion of variation in the dependent variable that is explained by the model.
2	0	AdjR2	0.443948	Adjusted R-Squared: R-Squared adjusted for model complexity (number of variables) as it relates to the data.
3	0	F-Stat	180.139658	Joint F-Statistic Value: Used to assess overall model significance.
4	0	F-Prob	0	Joint F-Statistic Probability (p-value): The probability that none of the explanatory variables have an effect on the dependent variable.
5	0	Wald	350.685184	Wald Statistic: Used to assess overall robust model significance.
6	0	Wald-Prob	0	Wald Statistic Probability (p-value): The computed probability, using robust standard errors, that none of the explanatory variables have an effect on the dependent variable.
7	0	K(BP)	473.961616	Koenker's studentized Breusch-Pagan Statistic: Used to test the reliability of standard error values when heteroskedasticity (non-constant variance) is present.
8	0	K(BP)-Prob	0	Koenker (BP) Statistic Probability (p-value): The probability that heteroskedasticity (non-constant variance) has not made standard errors unreliable.
9	0	JB	8954.391925	Jarque-Bera Statistic: Used to determine whether the residuals deviate from a normal distribution.
10	0	JB-Prob	0	Jarque-Bera Probability (p-value): The probability that the residuals are normally distributed.
11	0	Sigma2	342062503.182	Sigma-Squared: OLS estimate of the variance of the error term.

Record: 1 Show: All Selected Records (0 out of 12 Selected) Options

Woodstock

Attributes of diag

OID	Field1	Diag_Name	Diag_Value	Definition
0	0	AIC	80227.103684	Akaike's Information Criterion: A relative measure of performance used to compare models; the smaller AIC indicates the superior model.
1	0	R2	0.441975	R-Squared, Coefficient of Determination: The proportion of variation in the dependent variable that is explained by the model.
2	0	AdjR2	0.440555	Adjusted R-Squared: R-Squared adjusted for model complexity (number of variables) as it relates to the data.
3	0	F-Stat	311.268959	Joint F-Statistic Value: Used to assess overall model significance.
4	0	F-Prob	0	Joint F-Statistic Probability (p-value): The probability that none of the explanatory variables have an effect on the dependent variable.
5	0	Wald	1019.296946	Wald Statistic: Used to assess overall robust model significance.
6	0	Wald-Prob	0	Wald Statistic Probability (p-value): The computed probability, using robust standard errors, that none of the explanatory variables have an effect on the dependent variable.
7	0	K(BP)	951.048283	Koenker's studentized Breusch-Pagan Statistic: Used to test the reliability of standard error values when heteroskedasticity (non-constant variance) is present.
8	0	K(BP)-Prob	0	Koenker (BP) Statistic Probability (p-value): The probability that heteroskedasticity (non-constant variance) has not made standard errors unreliable.
9	0	JB	4742.121358	Jarque-Bera Statistic: Used to determine whether the residuals deviate from a normal distribution.
10	0	JB-Prob	0	Jarque-Bera Probability (p-value): The probability that the residuals are normally distributed.
11	0	Sigma2	388423121.426	Sigma-Squared: OLS estimate of the variance of the error term.

Sabin

OID	Field1	Variable	Coef	StdError	t_Stat	Prob	Robust_SE	Robust_t	Robust_Pr
0	0	Intercept	55029.090244	16406.310916	3.354142	0.000836	14034.487571	3.92099	0.000102
1	0	AREA	10.966429	0.484094	22.653514	0	1.362712	8.047504	0
2	0	DISTANCE	8.759252	3.174045	2.75965	0.005871	2.501647	3.501394	0.000494
3	0	DISTSCHOOL	-18.490455	3.819396	-4.841199	0.000002	3.142842	-5.883356	0
4	0	DISTPARKS	-2.312238	2.211329	-1.045633	0.295921	1.756417	-1.316452	0.18827
5	0	DISTMUR	5.611391	3.151506	1.780543	0.075233	2.014929	2.784907	0.005436
6	0	DISTARTER	-3.065227	2.434089	-1.259291	0.208163	2.563991	-1.19549	0.232121
7	0	DISTCMCTRS	5.942349	2.646453	2.245401	0.024899	1.866431	3.183803	0.001504
8	0	DISTGARDEN	13.289919	4.073894	3.262216	0.001151	2.967023	4.47921	0.000011
9	0	CANOPYCVR	-553.160166	3077.592412	-0.179738	0.85738	3048.84998	-0.181432	0.85605
10	0	DISTALAMED	-13.753479	1.359674	-10.115275	0	1.294655	-10.62328	0

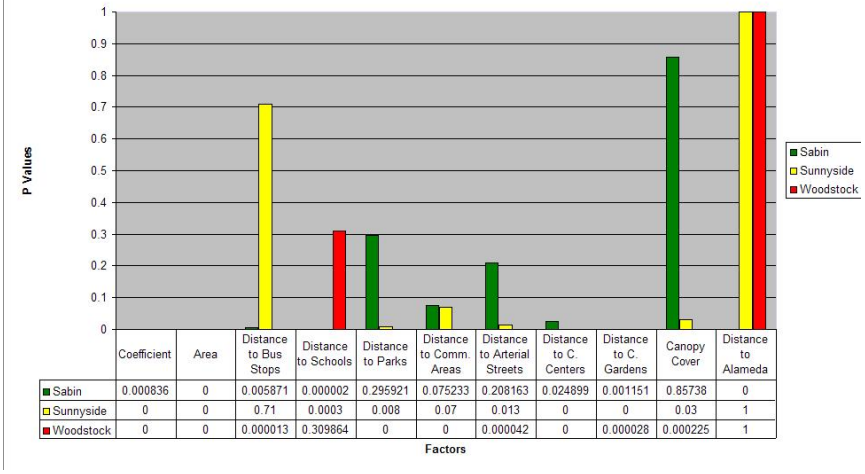
Sunnyside

OID	Field1	Variable	Coef	StdError	t_Stat	Prob	Robust_SE	Robust_t	Robust_Pr
0	0	Intercept	99510.588993	2154.550501	46.186241	0	2977.930222	33.416024	0
1	0	AREA	11.331552	0.321823	35.210513	0	0.795133	14.251145	0
2	0	DIST_ART	10.177584	4.085838	2.490942	0.012819	3.488808	2.917209	0.003584
3	0	DIST_BUS	1.260943	3.435279	0.367057	0.713635	3.351204	0.376266	0.706779
4	0	DIST_PARKS	3.063237	1.153061	2.656613	0.007959	1.01164	3.027991	0.002509
5	0	DIST_MUR	-9.210249	5.083288	-1.811868	0.070178	4.430204	-2.078967	0.037751
6	0	DIST_SCHL	-3.862745	1.081958	-3.570143	0.000381	1.064937	-3.627204	0.000308
7	0	CAN_COV	-4932.82811	2300.157062	-2.144561	0.032108	2500.191812	-1.97298	0.048645
8	0	DIST_GARD	-8.544976	0.593356	-14.401102	0	0.648345	-13.179681	0

Woodstock

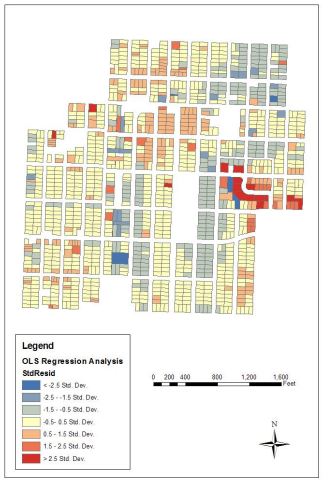
OID	Field1	Variable	Coef	StdError	t_Stat	Prob	Robust_SE	Robust_t	Robust_Pr
0	0	Intercept	54960.570677	2016.931901	27.249592	0	2615.267405	21.015278	0
1	0	AREA	8.059593	0.164943	48.862928	0	0.397289	20.286472	0
2	0	CNPY	5798.012236	1563.597478	3.708123	0.000225	1709.663595	3.391318	0.00072
3	0	DIST_BUS	4.484313	1.01201	4.431097	0.000013	1.049813	4.271534	0.000024
4	0	DIST_SCHL	-0.687784	0.6772	-1.01563	0.309864	0.6368	-1.080063	0.28018
5	0	DIST_COMCE	-3.496669	0.317873	-11.000204	0	0.329454	-10.613523	0
6	0	DIST_CMGAR	1.617186	0.381423	4.239876	0.000028	0.384487	4.206085	0.000032
7	0	DIST_PARK	-3.979842	0.694032	-5.734377	0	0.763098	-5.215378	0
8	0	DIST_ZONE	6.528417	0.901905	7.238478	0	0.869453	7.508652	0
9	0	DIST_ARTL	2.474471	0.598581	4.133898	0.000042	0.539131	4.589739	0.000006

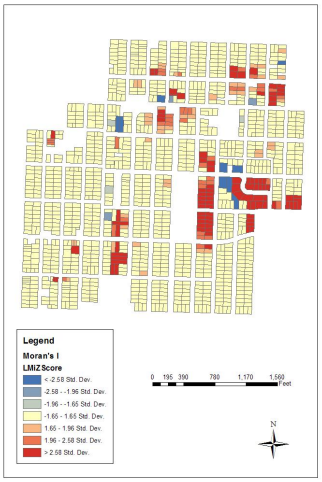
Prediction = a0+ a1x1 + a2x2 + a3x3 + ...



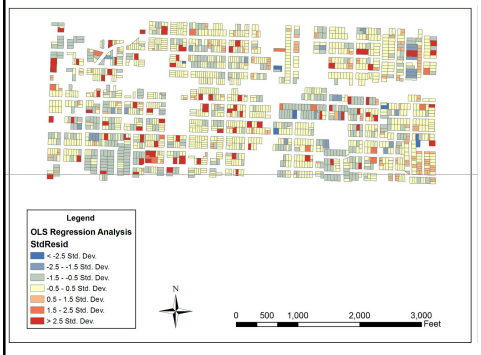
	Coefficient	Area	Distance to Bus Stops	Distance to Schools	Distance to Parks	Distance to Comm. Areas	Distance to Arterial Streets	Distance to C. Centers	Distance to C. Gardens	Canopy Cover	Distance to Alameda
Sabin	0.000836	0	0.005871	0.000002	0.295921	0.075233	0.208163	0.024899	0.001151	0.85738	0
Sunnyside	0	0	0.71	0.0003	0.008	0.07	0.013	0	0	0.03	1
Woodstock	0	0	0.000013	0.309864	0	0	0.000042	0	0.000028	0.000225	1

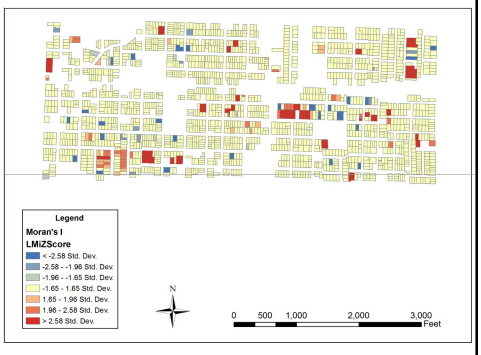
Sabin



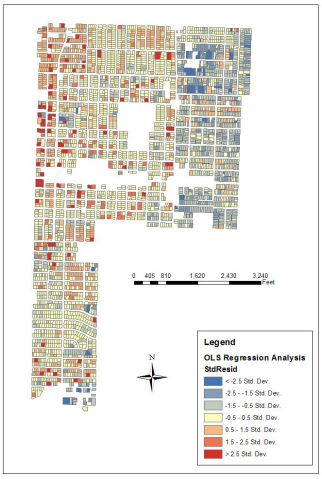


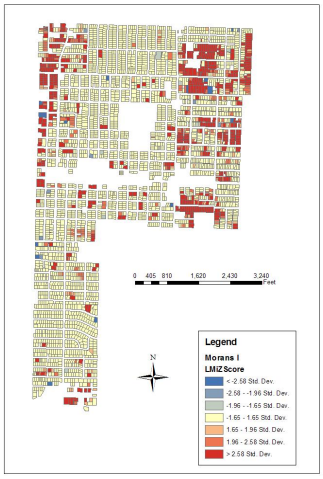
Sunnyside





Woodstock





Conclusions

- While our models were incomplete, we have established a foundation upon which we can expand and improve upon.
- Additional variables to include in a future model could be:
  - Crime, lot topography, lot frontage, schools, view, “hot” neighborhoods, etc.
  - Experiment with a nonlinear regression model

Questions?

References

- ArcGIS 9.3 Webhelp
- RLIS Metadata Viewer
- Dr. Geoffrey Duh
- *www.eecs.berkeley.edu/~asimma*