School of Geography and Earth Sciences
McMaster University

Advanced Topics in Spatial Statistics

Area Data
V & VI

This session

- Modeling area data
  - Non-spatial regression models
- Assumptions
  - Non-constant variance
    - Data transformations
  - Error autocorrelation
    - Moran's I
  - Normality

Example

- Voter turnout
  - What explains voter turnout in elections?
- What do we know about the data?
  - Variables follow a spatial pattern (Spatial autocorrelation)
  - Variables are probably correlated

Example

- Voter turnout: Variables
  - VOTE
  - Voting age population (POP)
  - Population with grade 12 or higher education (EDU)
  - Number of owner occupied houses (HOUSE)
  - Aggregate income (INCOME)
- Unit of analysis: county

Example

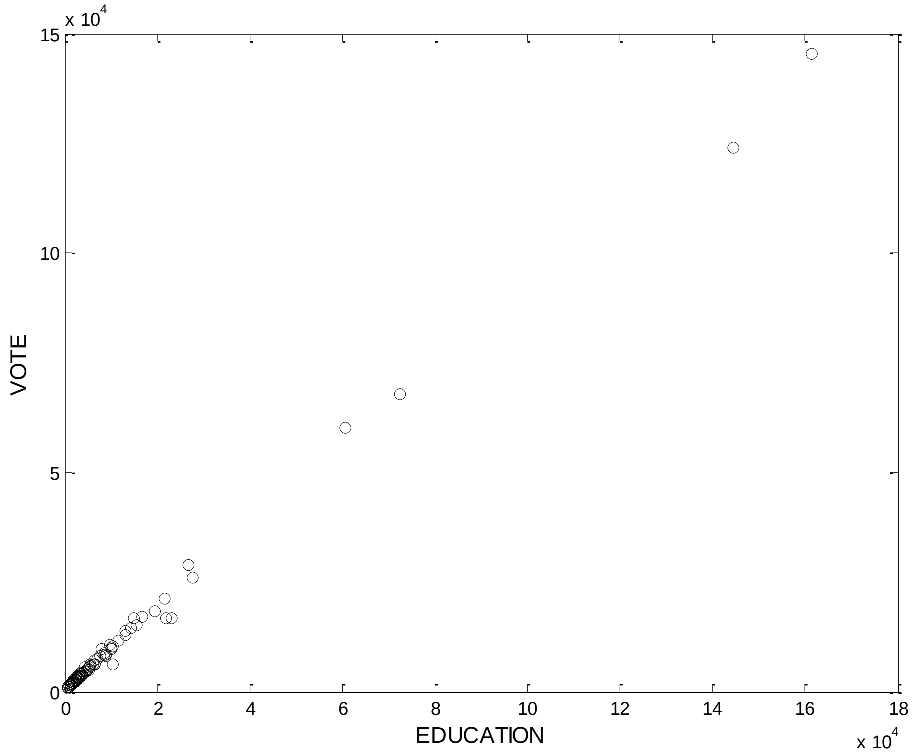
- Do we have prior expectations about the direction of the relationships?

Example

- Bivariate analysis
  - Scatterplots
  - Correlation coefficients

Example

- Scatterplot (EDU vs. VOTE)



Example

- Bivariate analysis
  - Correlation coefficients

	VOTE	POP	EDU	HOUSE	INCOME
VOTE	1.00				
POP	0.993	1.00			
EDU	0.998	0.990	1.00		
HOUSE	0.997	0.994	0.994	1.00	
INCOME	0.994	0.983	0.998	0.989	1.00

Example

- Modeling area data
  - Non-spatial regression
    - Variables can be spatial
    - Variables may be autocorrelated
    - Assumes that there is no spatial autocorrelation
  - Multiple regression

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{b} + \boldsymbol{e}$$

Modeling Area Data

- Try different models
- Model evaluation guidelines

Modeling Area Data

- Model 1

Model 1 (VOTE)

Variable	PARAMETER	t-value
CONST	695.21	5.23
EDU	0.89	163.61

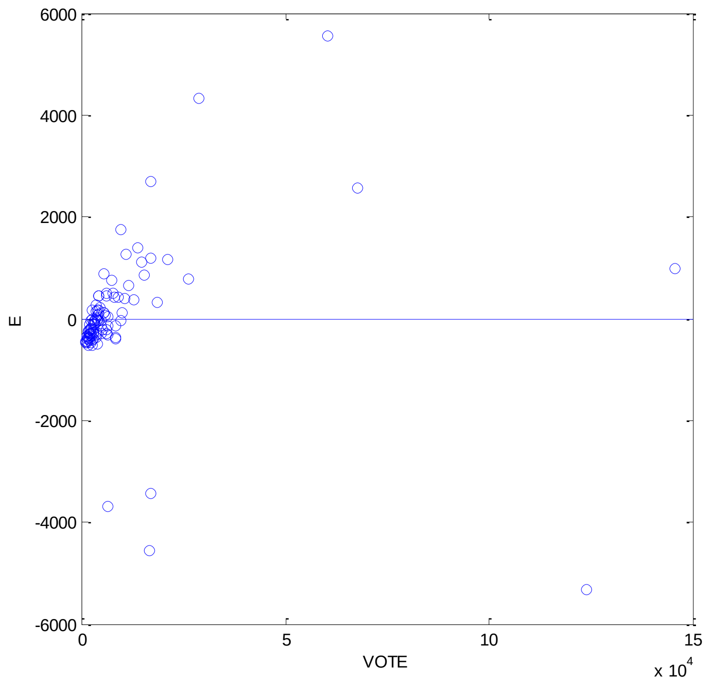
R^2=	0.996
R^2(adj)=	0.977
SIGMA^2=	1560013.858
SIGMA^2 (ML)=	1530299.308
SIGMA =	1249.005
n=	105

<< Normalized Moran's I >>

Z (I) = -2.821

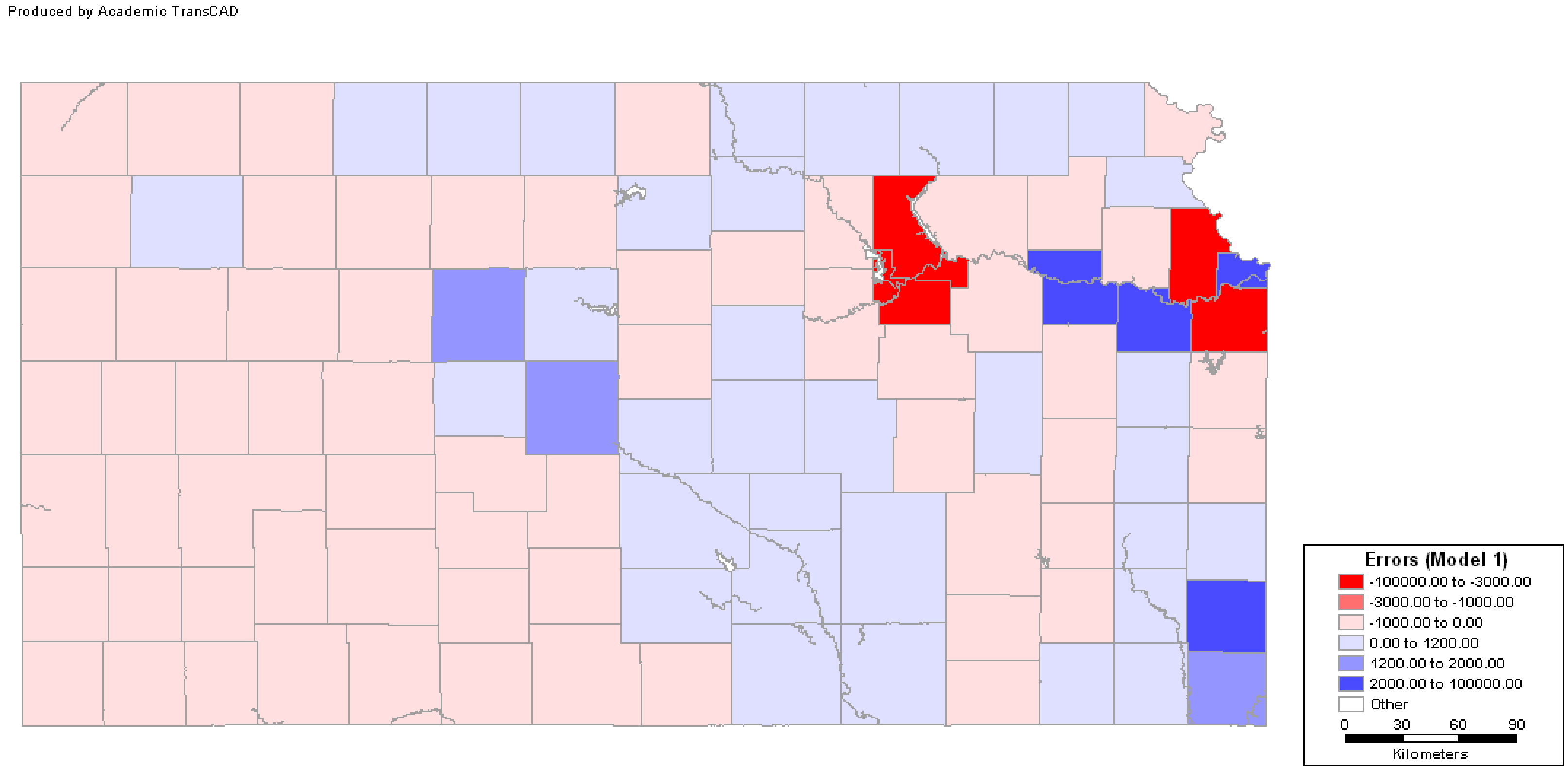
Analysis of Residuals

- Is variance constant?
  - Scatterplot of errors vs. variables



Analysis of Residuals

- Are errors independent?



Analysis of Residuals

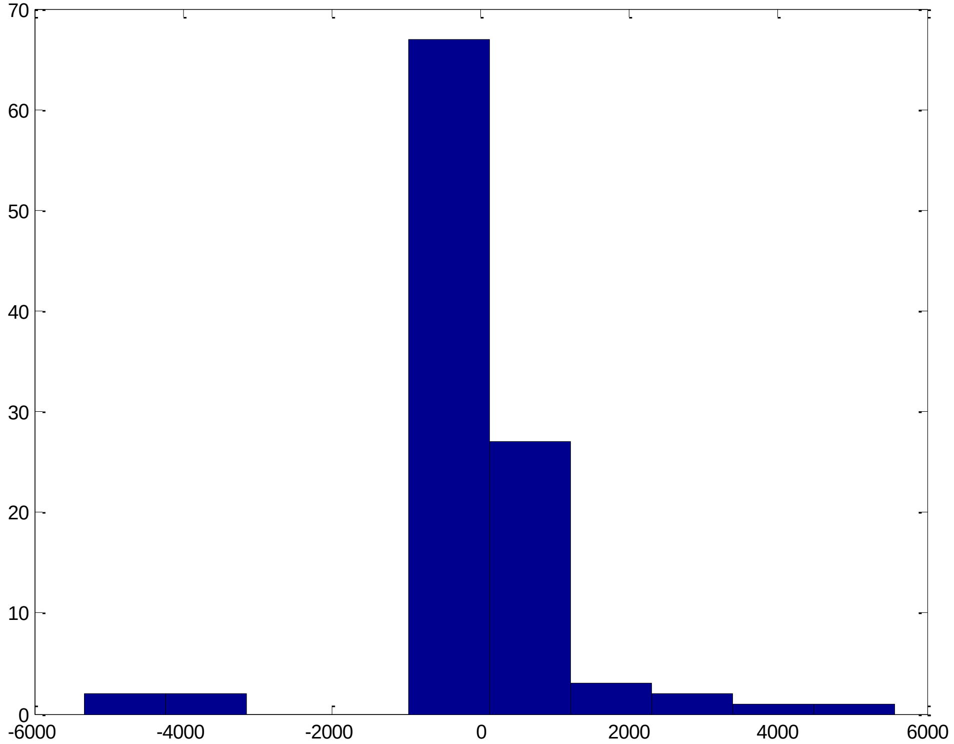
- Are errors independent?
  - Moran's I
  - Expected (mean) value of e is zero

Analysis of Residuals

- Are errors normally distributed?
  - Histogram
  - Probability plot
  - Other tests (Jarque-Bera, Kolmogorov-Smirnov,Lilliefors)

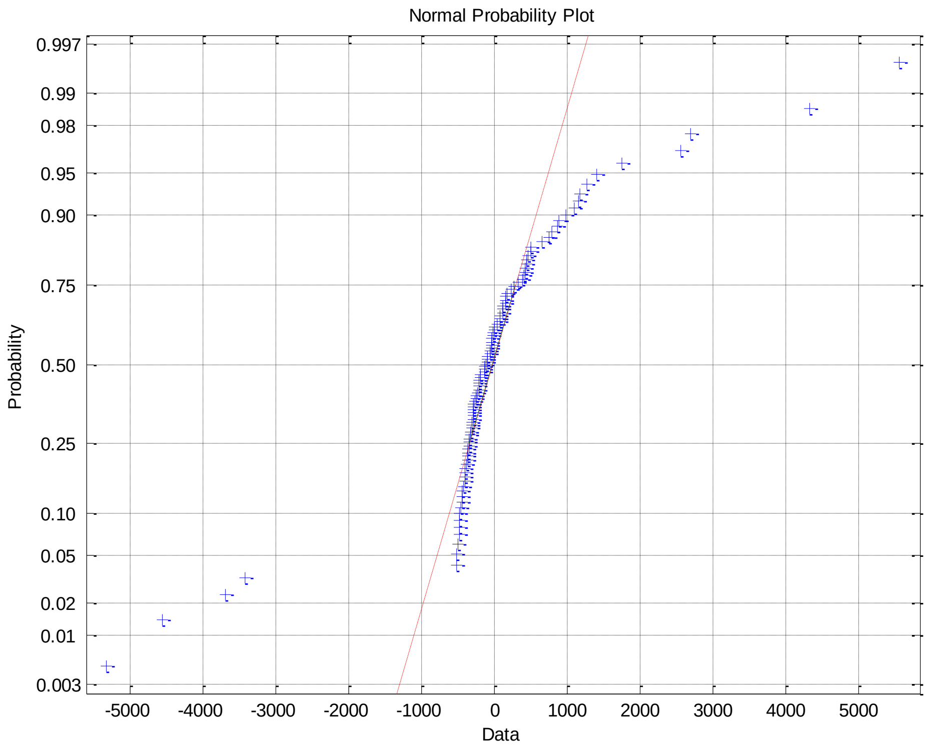
Analysis of Residuals

- Histogram



Analysis of Residuals

- Probability plot



Modeling Area Data

- What is the conclusion regarding this model?

Modeling Area Data

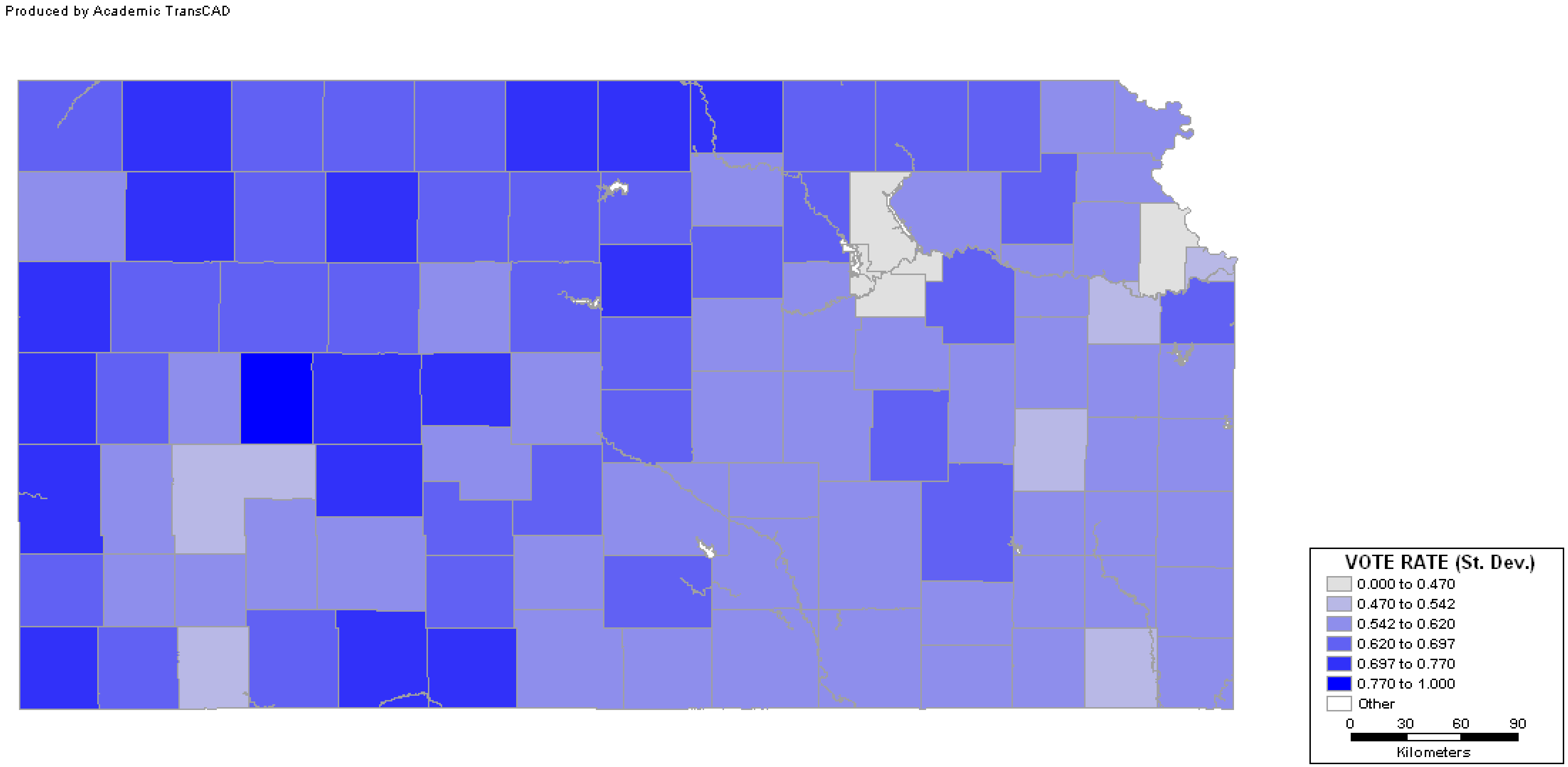
- Data transformations
  - May help to reduce size effects
  - May increase the interpretability of the models

Modeling Area Data

- Data transformations
  - Instead of modeling totals, model the proportions
  - Proportional voter turnout: VOTE/POP

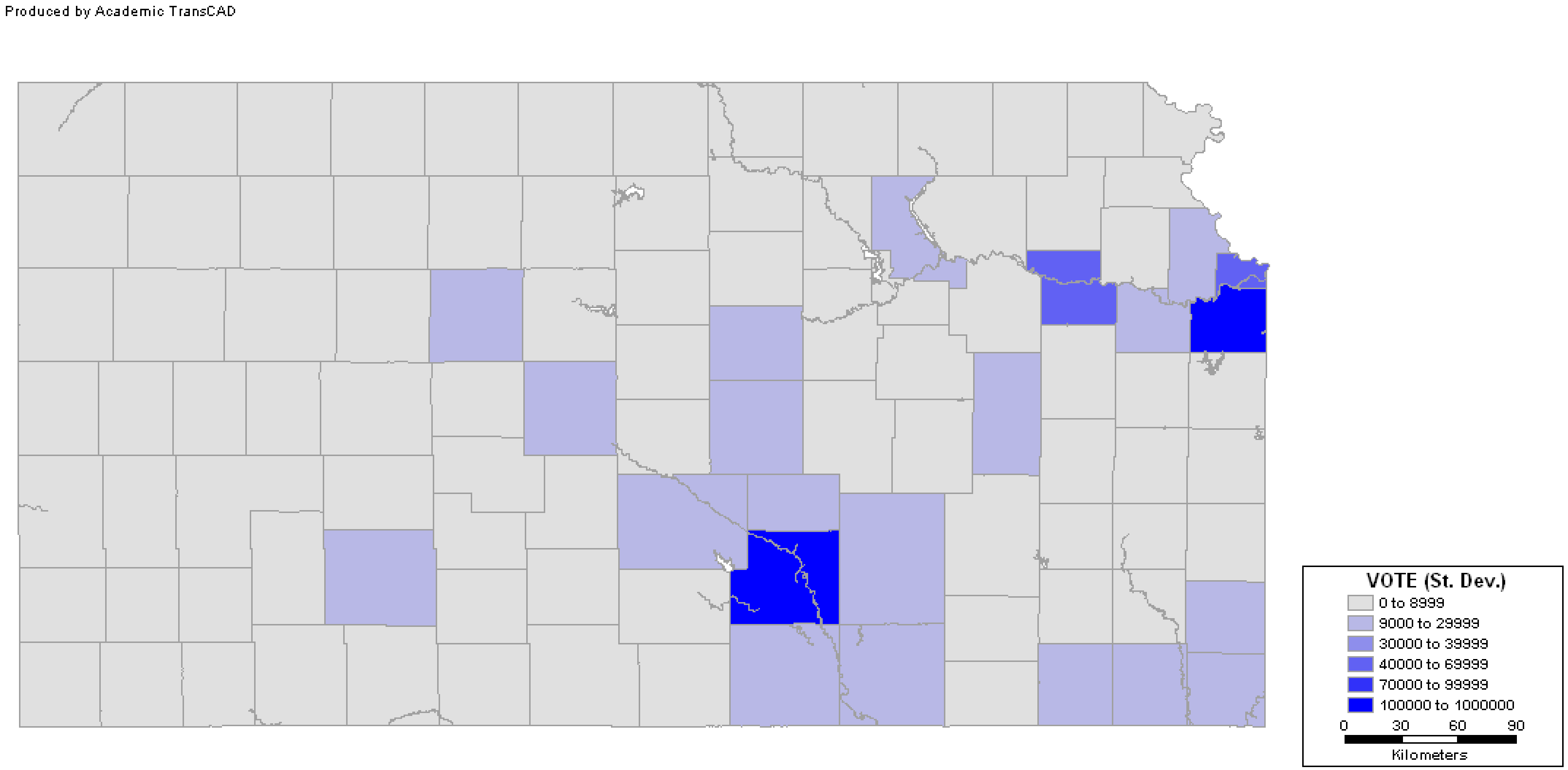
Exploring the Proportions

- Choropleth map: PrVotes (Std. Dev.)



Exploring the Proportions

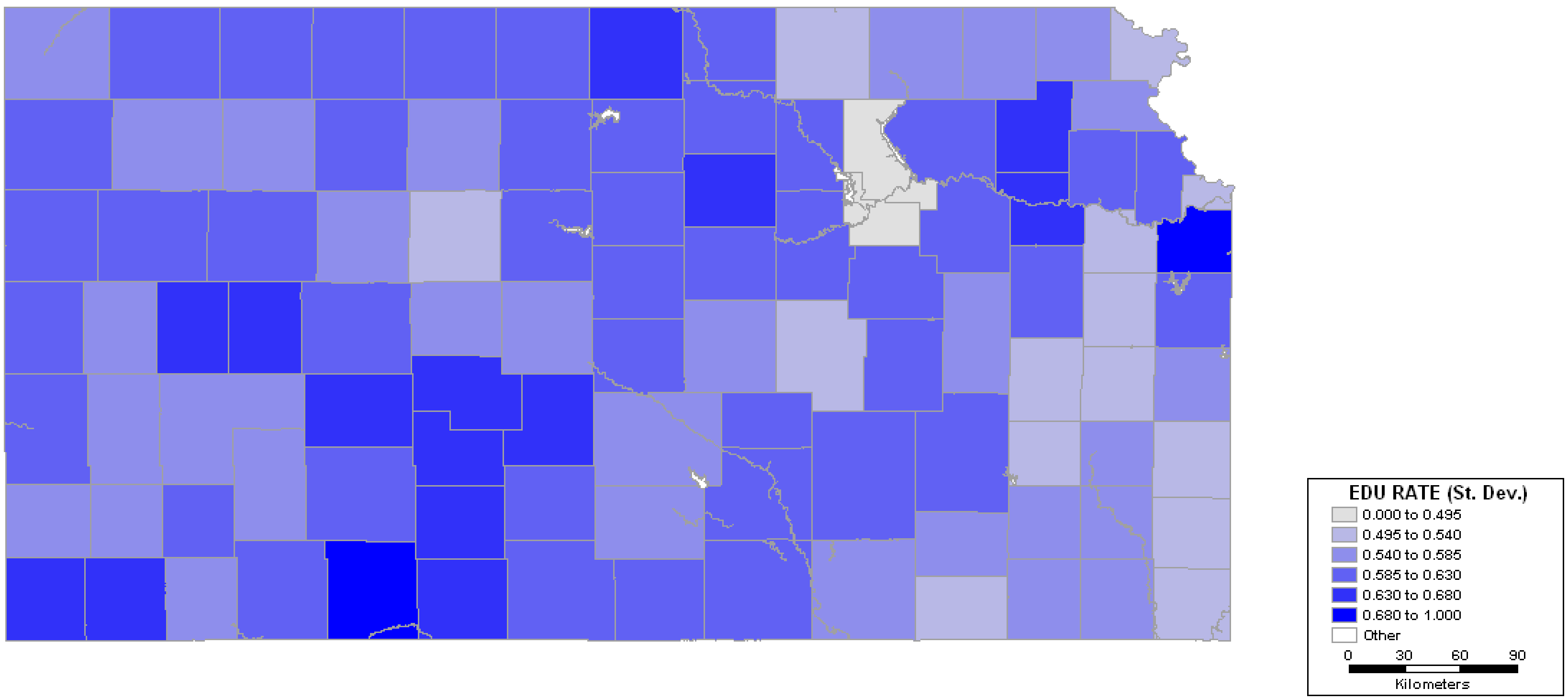
- Choropleth map: Votes (Std. Dev.)



Exploring the Proportions

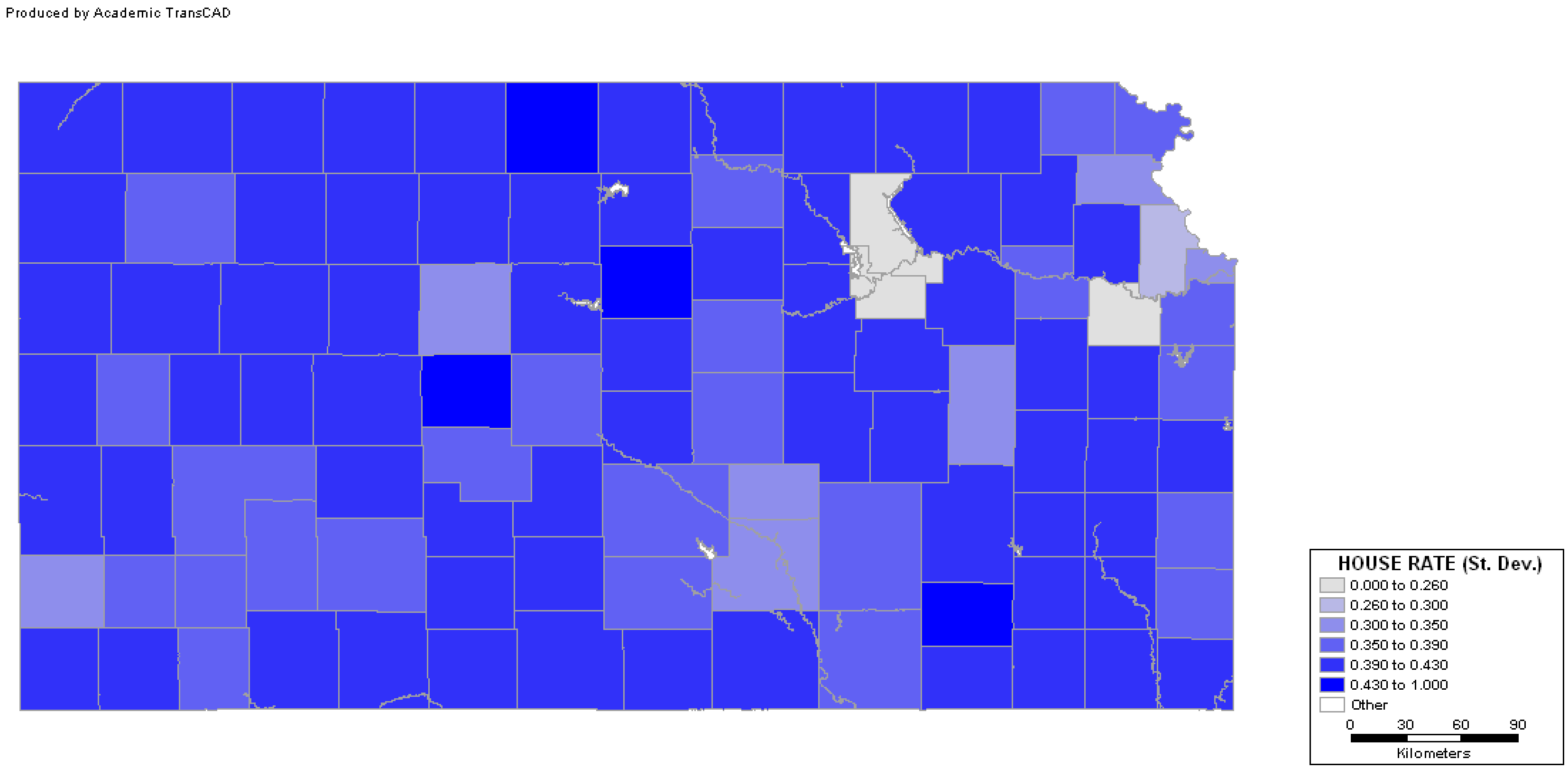
- PrEDU (Std. Dev.)

Produced by Academic TransCAD



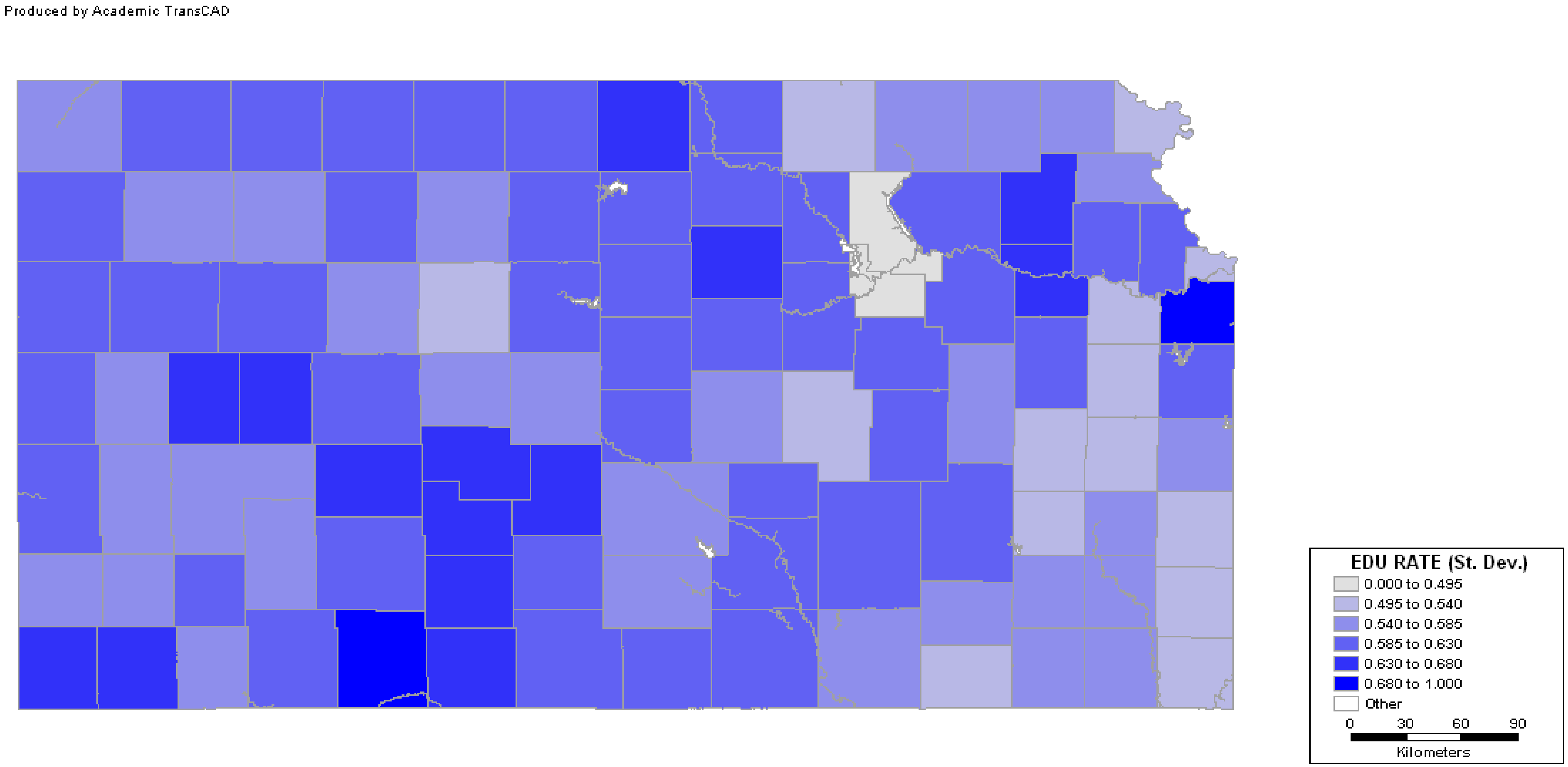
Exploring the Proportions

- PrHOUSE (Std. Dev.)



Exploring the Proportions

- PrINCOME (Std. Dev.)

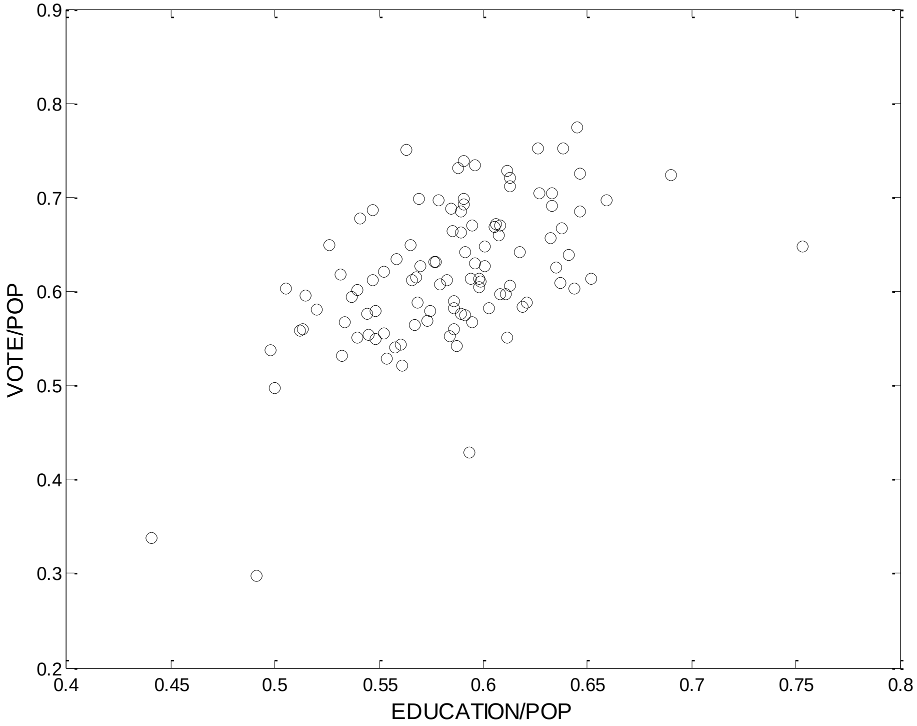


Example

- Bivariate analysis
  - Scatterplots
  - Correlation coefficients

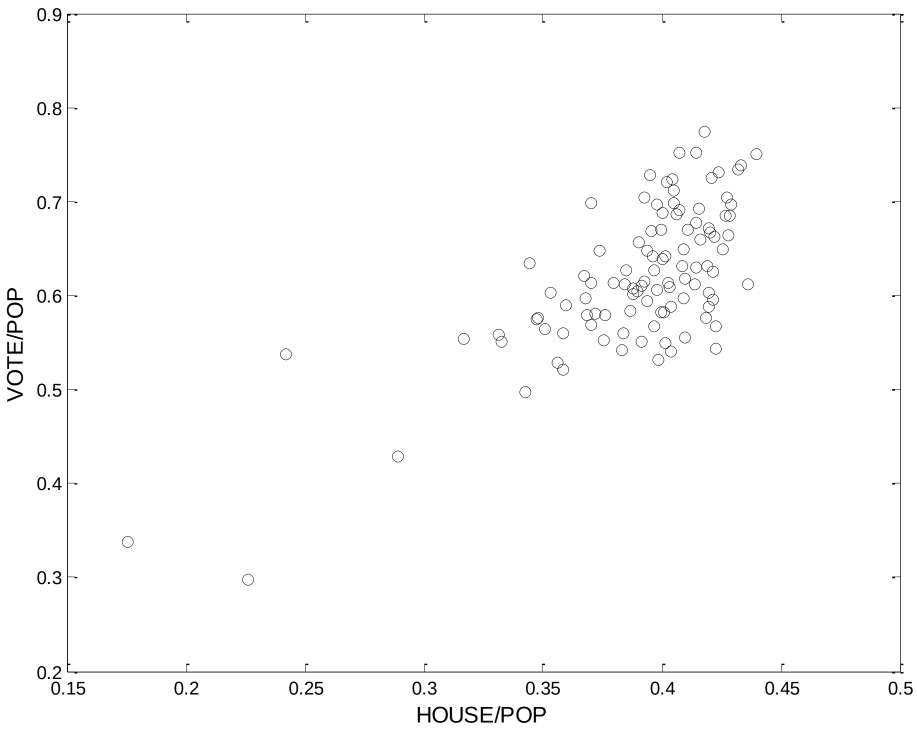
Example

- Scatterplot (EDU/POP vs. VOTE/POP)



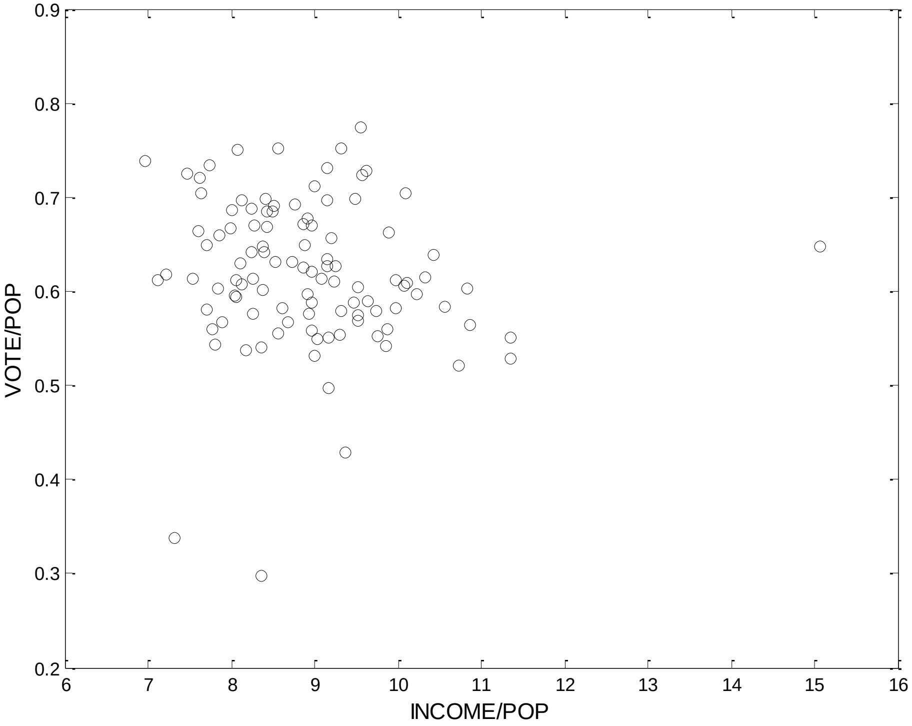
Example

- Scatterplot (HOUSE/POP vs. VOTE/POP)



Example

- Scatterplot (INCOME/POP vs. VOTE/POP)



Example

- Bivariate analysis
  - Correlation coefficients

	VOTE/POP	EDU/POP	HOUSE/POP	INCOME/POP
VOTE/POP	1.00			
EDU/ POP	0.563	1.00		
HOUSE/POP	0.718	0.392	1.00	
INCOME/POP	-0.111	0.426	-0.191	1.00

Example

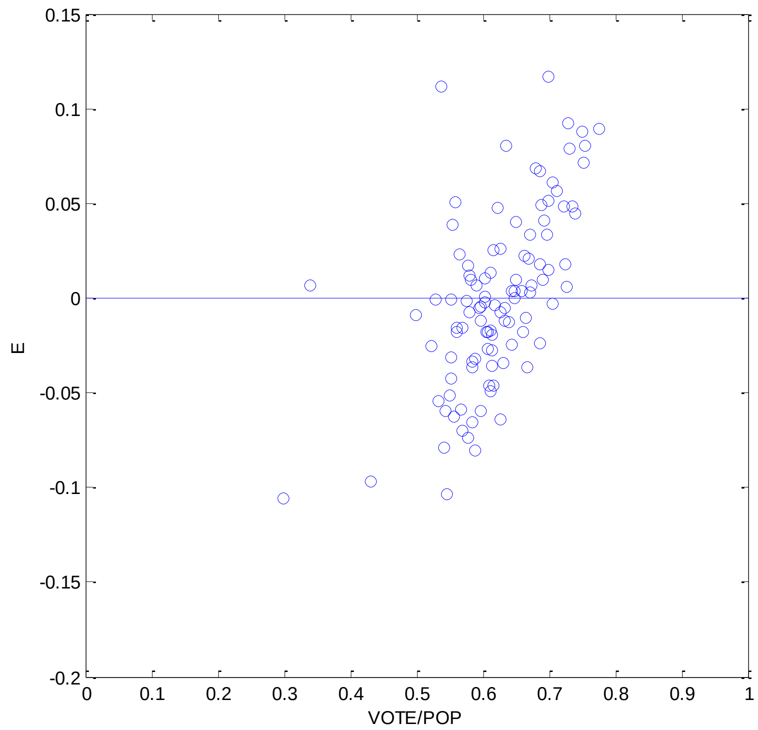
- Model 2

Model 2 (VOTE/POP)

Variable	PARAMETER	t-value
CONST	-0.071	-1.04
EDU/POP	0.782	5.76
HOUSE/POP	0.934	6.83
INCOME/POP	-0.015	-2.85
R^2=	0.636	
R^2(adj)=	0.611	
SIGMA =	0.047	
n=	105	
<< Normalized Moran's I >>		
Z (I) =	4.023	

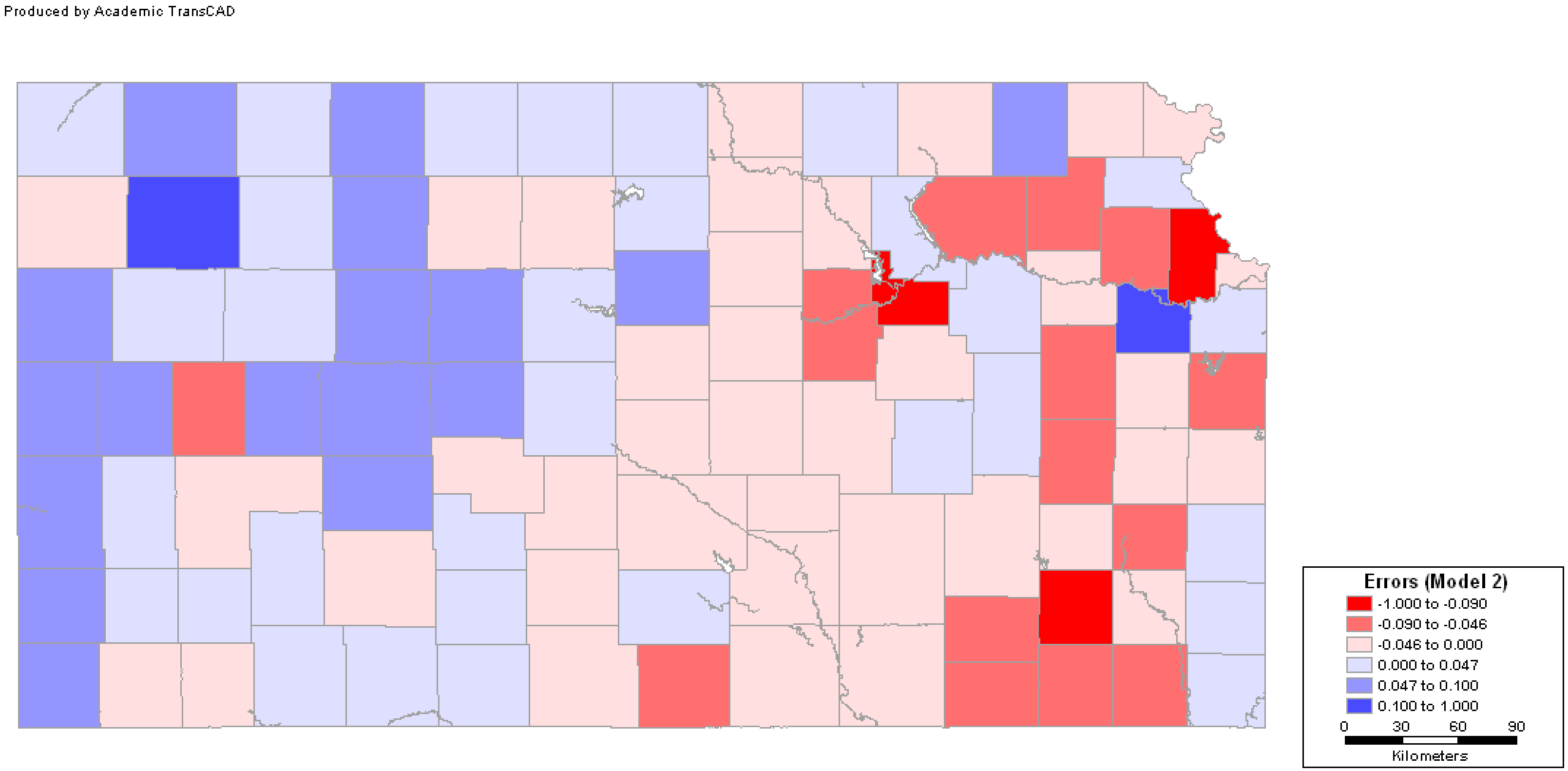
Analysis of Residuals

- Is variance constant?
  - Scatterplot of errors vs. variables



Analysis of Residuals

- Are errors independent?



Analysis of Residuals

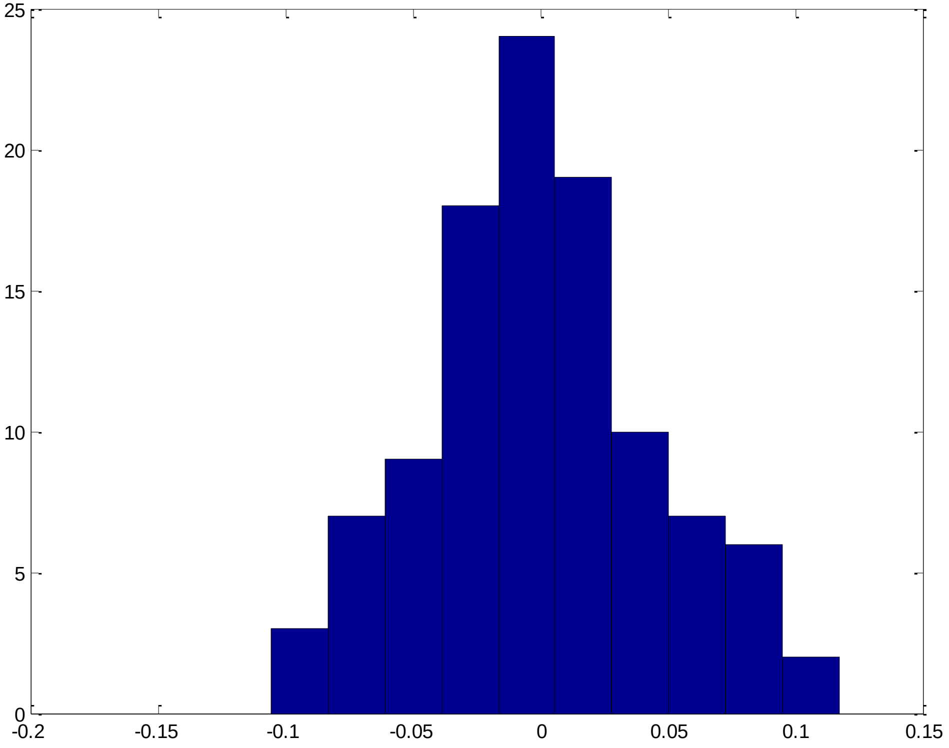
- Are errors independent?
  - Moran's I
  - Expected (mean) value of e is zero

Analysis of Residuals

- Are errors normally distributed?
  - Histogram
  - Probability plot
  - Other tests (Jarque-Bera, Kolmogorov-Smirnov,Lilliefors)

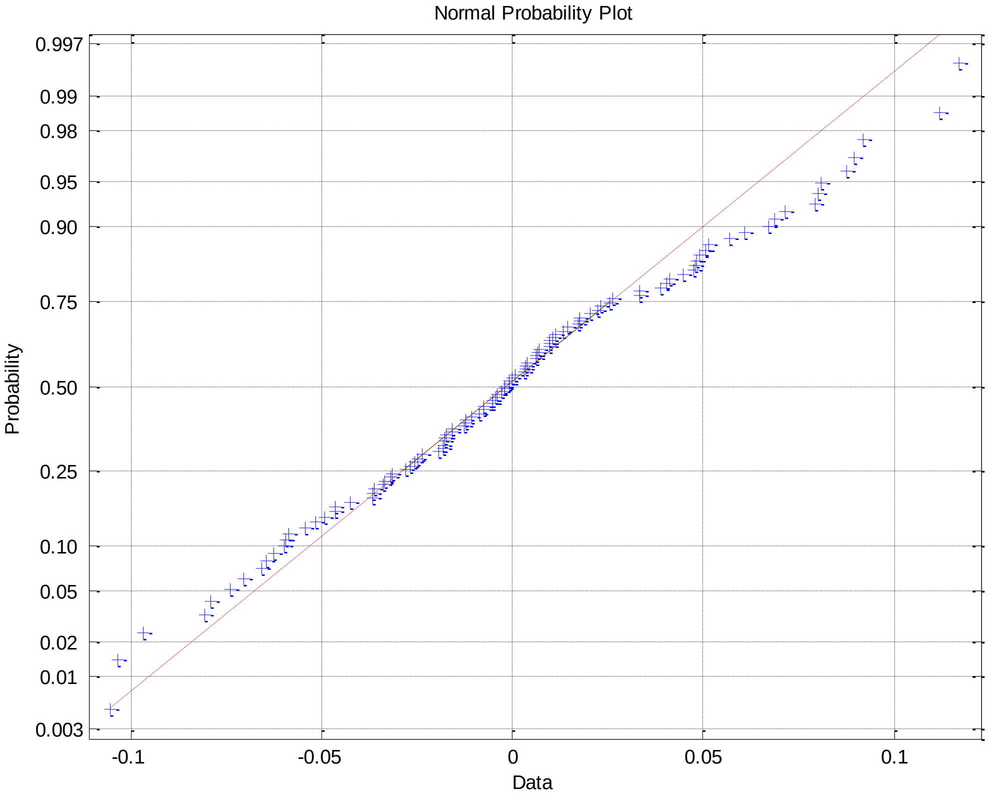
Analysis of Residuals

- Histogram



Analysis of Residuals

- Probability plot



Modeling Area Data

- Data transformations
  - Logarithm of proportional voter turnout:

$$\ln(PrVOTE)$$

  - Correlation coefficients
  - Model
  - Analysis of residuals

Variable Transformations

- Taking logarithms: $\ln(PrVOTE)$

Model 3 ln(VOTE/POP)

Variable	PARAMETER	t-value
CONST	0.926	4.70
ln(EDU/POP)	0.710	5.28
ln(HOUSE/POP)	0.614	8.21
ln(INCOME/POP)	-0.205	-2.57
R^2=	0.674	
R^2(adj)=	0.649	
SIGMA =	0.081	
n=	105	
<< Normalized Moran's I >>'		
Z (I) =	3.138	

Normality? No
Constant variance? Yes

Conclusion

- Model selection
  - Totals
  - Proportions
  - Logarithm of Proportions

Conclusion

- All three models still show residual error pattern (error autocorrelation)

Non-spatial Regression

- Data transformations may help to reduce problems with the residuals + increase interpretability
- In the example: All three models show residual error pattern (error autocorrelation)
- Spatial regression models

Spatial Regression Modeling

- Objective:
  - Testing for and estimating regression models that incorporate spatial effects
- Recall GLS

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{U}$$

$$E[\boldsymbol{U}] = \boldsymbol{0}$$

$$E[\boldsymbol{U}\boldsymbol{U}'] = \mathbf{C}$$

Spatial Regression Modeling

- Generalized Least Squares

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{U}$$

  - 2^{nd} order effects?
  - Questionable assumption of stationarity of 2^{nd} order
  - Distance measures?

Spatial Regression Modeling

- Autocorrelated errors model

$$\boldsymbol{Y} = \underbrace{\mathbf{X}\boldsymbol{\beta}}_{\text{Non-spatial part}} + \boldsymbol{U}$$

$$\underbrace{\boldsymbol{U} = \rho\mathbf{W}\boldsymbol{U} + \varepsilon}_{\text{Spatial part (residual pattern)}}$$

Autocorrelated Errors Modeling

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{U}$$

$$\boldsymbol{U} = \rho\mathbf{W}\boldsymbol{U} + \boldsymbol{\varepsilon}$$

$$E[\boldsymbol{\varepsilon}] = 0$$

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2\mathbf{I}$$

- Rewriting:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\boldsymbol{Y} - \rho\mathbf{W}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{C} = \sigma^2\left(\left(\mathbf{I} - \rho\mathbf{W}\right)^T\left(\mathbf{I} - \rho\mathbf{W}\right)\right)^{-1}$$

Estimation of the Spatial Model

- OLS estimation not useful
- Maximum likelihood estimation of β and ρ
- Assumptions:
  - y_i are observations on n random variables $\boldsymbol{Y}$, jointly normally distributed with Mean = $\mathbf{X}\boldsymbol{\beta}$ and Covariance matrix $\mathbf{C}$, where

$$\mathbf{C} = \sigma^2 \left((\mathbf{I} - \rho\mathbf{W})^T (\mathbf{I} - \rho\mathbf{W}) \right)^{-1}$$

Estimation of a Spatial Model

- Maximum likelihood estimation of $\boldsymbol{\beta}$ and ρ
- What is the maximum likelihood function?

Alternative Spatial Models

- Lag models

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{U}$$

$$\boldsymbol{U} = \rho_1 \mathbf{W}^{(1)}\boldsymbol{U} + \rho_1 \mathbf{W}^{(2)}\boldsymbol{U} + \ldots + \varepsilon$$

Alternative Spatial Models

- Mixed regressive autoregressive

$$\mathbf{Y} = \underbrace{\rho\mathbf{WY}}_{\text{Spatial part}} + \underbrace{\mathbf{X}\boldsymbol{\beta}}_{\text{Non-spatial part}} + \varepsilon$$

Alternative Spatial Models

- Pure autoregressive

$$\mathbf{Y} = \underbrace{\rho \mathbf{W}\mathbf{Y}}_{\text{Spatial part}} + \varepsilon$$

Alternative Spatial Models

- Moving average model (map pattern)

$$\mathbf{Y} = \underbrace{(\mathbf{I} - \rho \mathbf{W})}_{\text{Spatial part}} \varepsilon$$

Spatial Regression

- Fit autocorrelated errors model

Model 4 VOTE/POP

Variable	PARAMETER	t-value
CONST	0.032	0.52
EDU/POP	0.645	4.80
HOUSE/POP	0.898	6.92
INCOME/POP	-0.015	-3.09
RHO	0.468	4.14
SIGMA =	0.002	
n=	105	
<< Lagrange Multiplier >>		
Omitted Sp. Lag=	1.038	--> Chi2 (1 DF)

Example

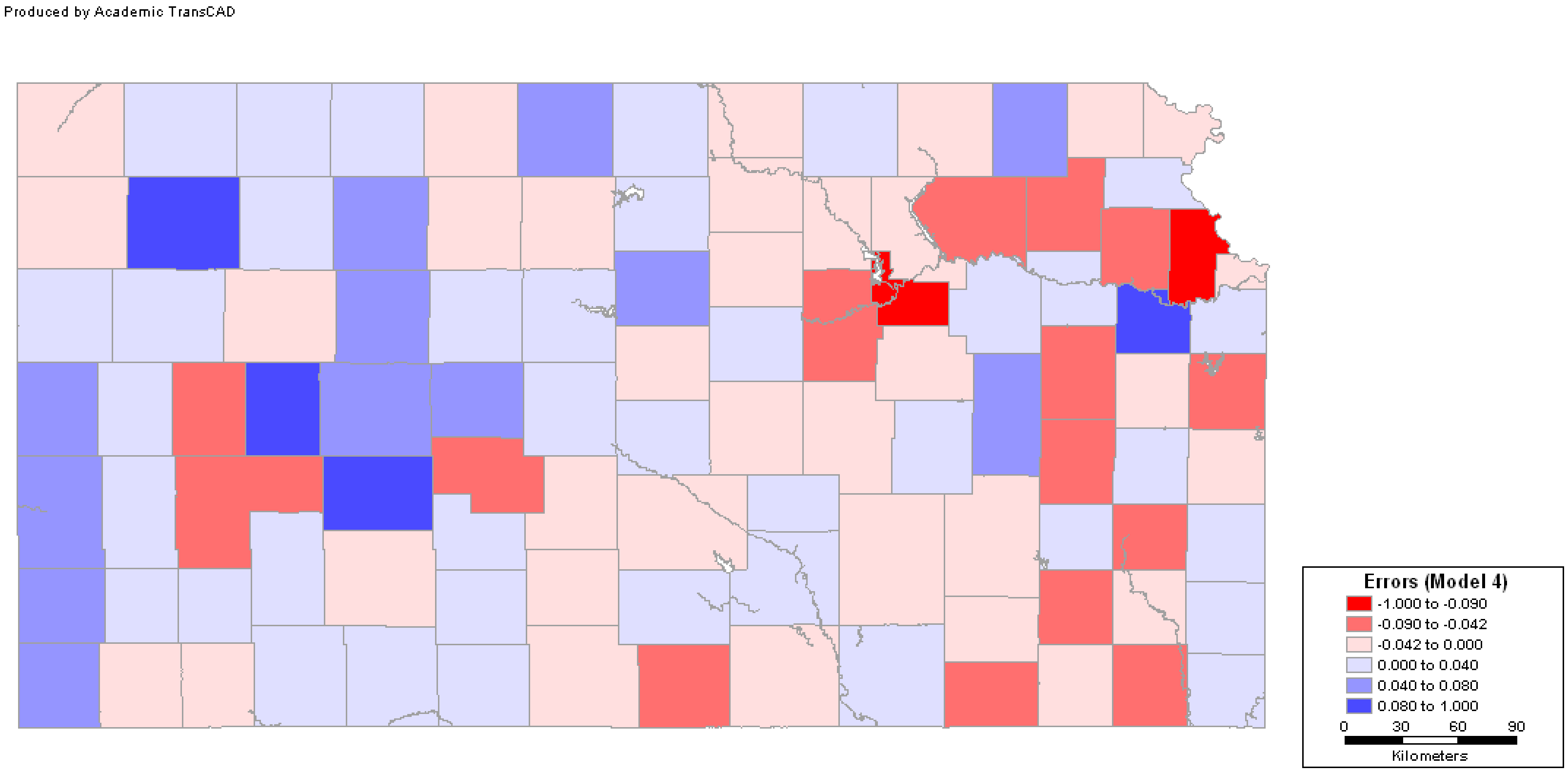
- Model 2

Model 2 (VOTE/POP)

Variable	PARAMETER	t-value
CONST	-0.071	-1.04
EDU/POP	0.782	5.76
HOUSE/POP	0.934	6.83
INCOME/POP	-0.015	-2.85
R^2=	0.636	
R^2(adj)=	0.611	
SIGMA =	0.047	
n=	105	
<< Normalized Moran's I >>		
Z (I) =	4.023	

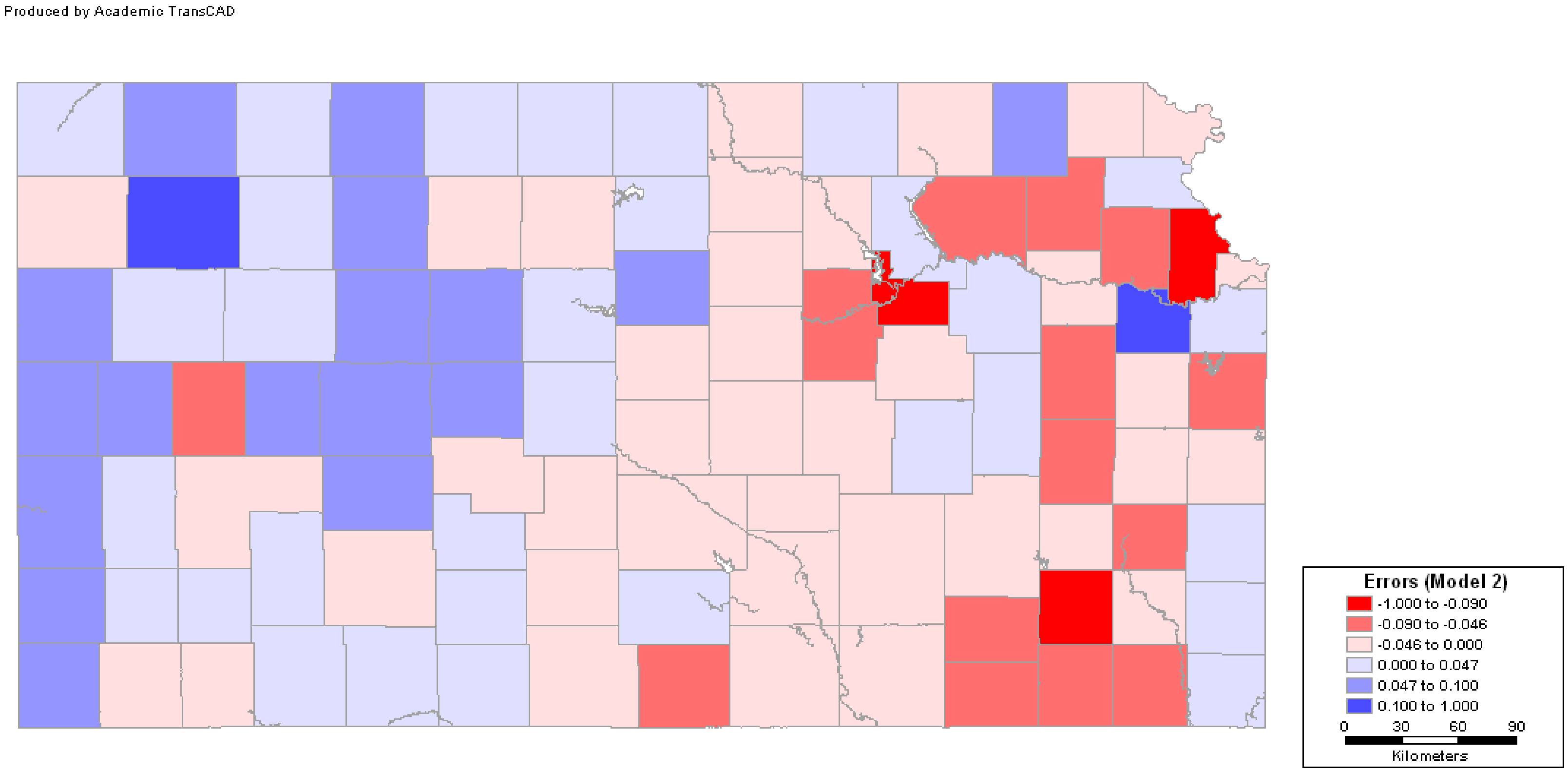
Analysis of Residuals

- Are errors independent? Model 4



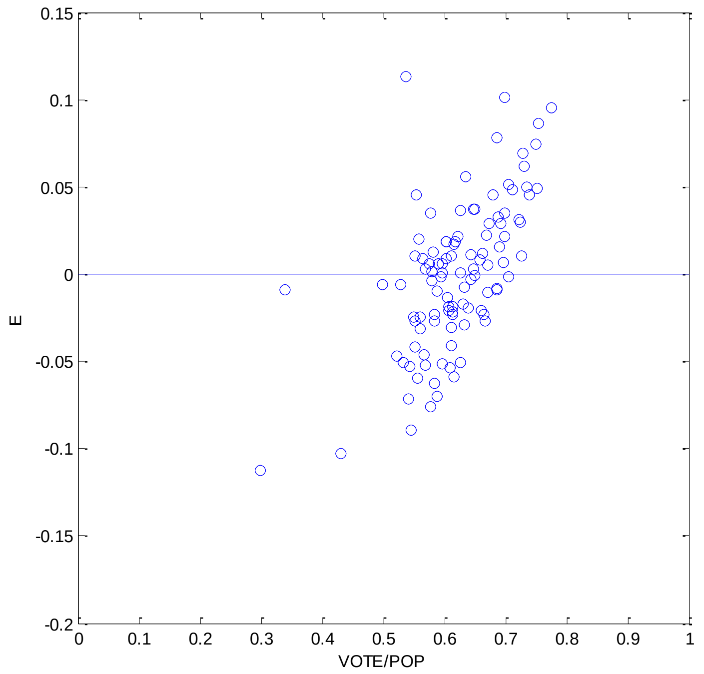
Analysis of Residuals

- Are errors independent? Model 2



Analysis of Residuals

- Is variance constant?
  - Scatterplot of errors vs. variables

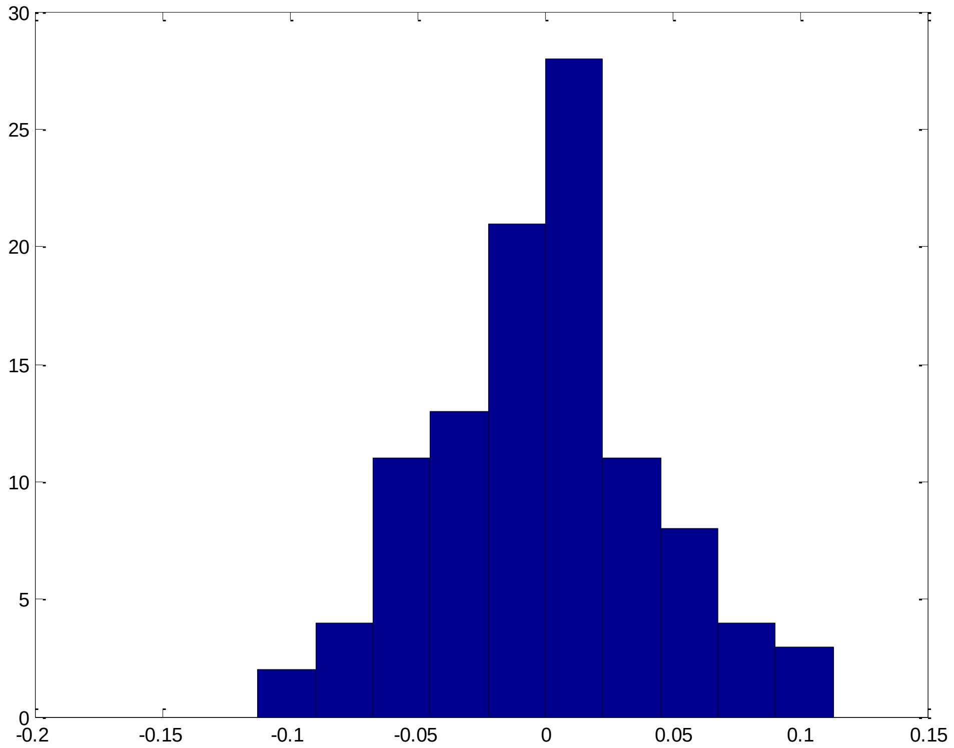


Analysis of Residuals

- Are errors normally distributed?
  - Histogram
  - Probability plot
  - Other tests (Jaques-Bera, Kolmogorov-Smirnov,Lilliefors)

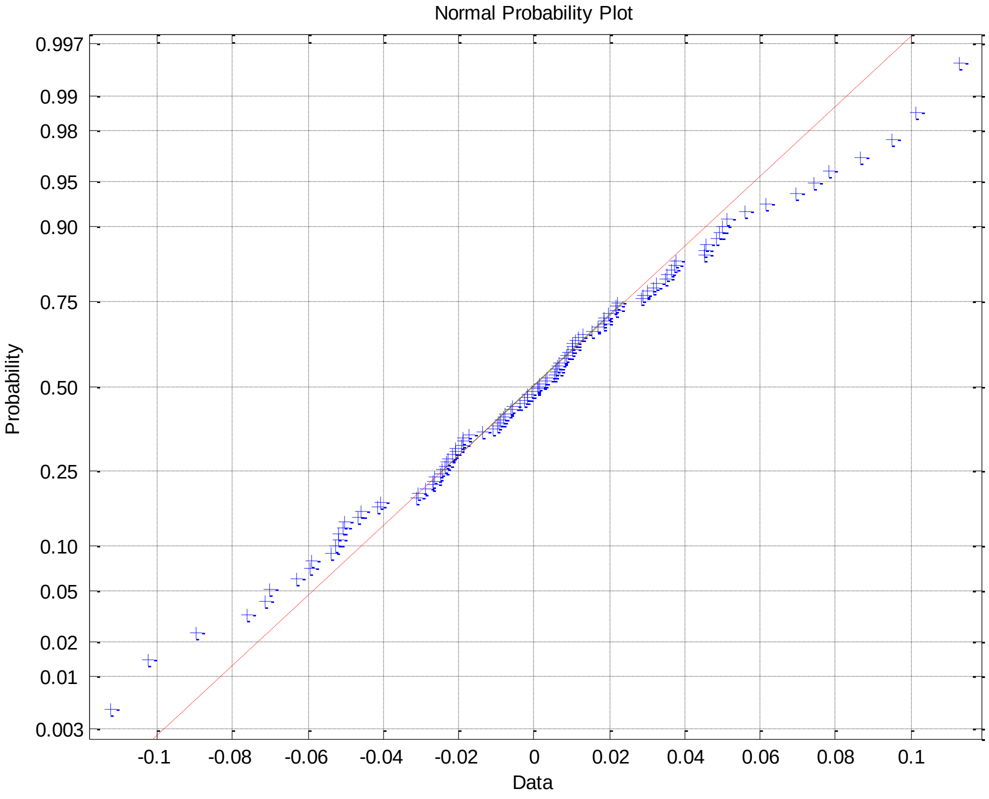
Analysis of Residuals

- Histogram



Analysis of Residuals

- Probability plot



Analysis of Residuals

- Other tests

Test	Normal?
Jarque-Bera	YES
Kolmogorov-Smirnov	YES
Lilliefors	YES

Conclusion

- The autocorrelated errors model is the best alternative
  - Satisfies all major assumptions
  - Accounts for all systematic spatial variation

Summary of modeling approach

- Exploration
- Models – non-spatial models, check assumptions
- Models – spatial models, check assumptions and statistical fit
- Think Theory!