An Exploration of Geographically Weighted Regression on Spatial Non-Stationarity and Principal Component Extraction of Determinative Information from Robust Datasets

A Case Study in Socio-Economic Patterns of Coronary Disease

James Norris, B.Sc, M.GIS

Table of Contents

1. Case Study Introduction
2. Background on PCA, GWR, and Getis-Ord Gi*
3. Methodology, Analyses and Results
4. Discussion and Interpretation
5. Summary and Conclusions

Golder Associates

Case Study Introduction

- An investigation of Spatial-Analytical processes using a Heart Disease Dataset

- Study Area: Calgary, Alberta, Canada
  - Population of Approximately 1 million
  - 181 Census Tracts
  - Structure of City: Monocentric > Concentric
  - Oil and Gas Based Economy
  - Divided into 4 quadrants (NW, NE, SW, SE)

- Study Time Frame: 1998-2002

Golder Associates

Cardiac Catheterization

- Proxy Dependent Variable for Risk of Coronary Disease (APPROACH, 2008)
- Procedure commonly performed post-incidence of myocardial infarction
- Also performed on patients presenting symptoms of cardiovascular disease (AHA, 2008)
- Beneficial to study in Canada, due to Universal Health Care

Golder Associates

Study Motivation

1. A method is needed to *reduce information loss* during Multiple Regression Analyses on Robust Datasets, and provide more *explanatory* predictors.

2. A method must be utilized to observe the influence of *Spatial Non-Stationarity* in relationships and examine its removal from datasets

The methods proposed here are transferable across many research fields, both medical and non medical

Golder Associates

Methods of Addressing Problems

Information Retention

- Principal Component Analysis (PCA)
  - Reduces dimensionality within large datasets containing inter-correlated variables
  - Large number of Variables reduced into small group of Components
  - Components are rotated orthogonally, so minimal inter-correlation exists between variables

Golder Associates

Methods of Addressing Problems

Spatial Non-Stationarity

- Geographically Weighted Regression (GWR)
  - Localized Method of Regression
  - Disaggregates global statistics by generating a model at each data point

Golder Associates

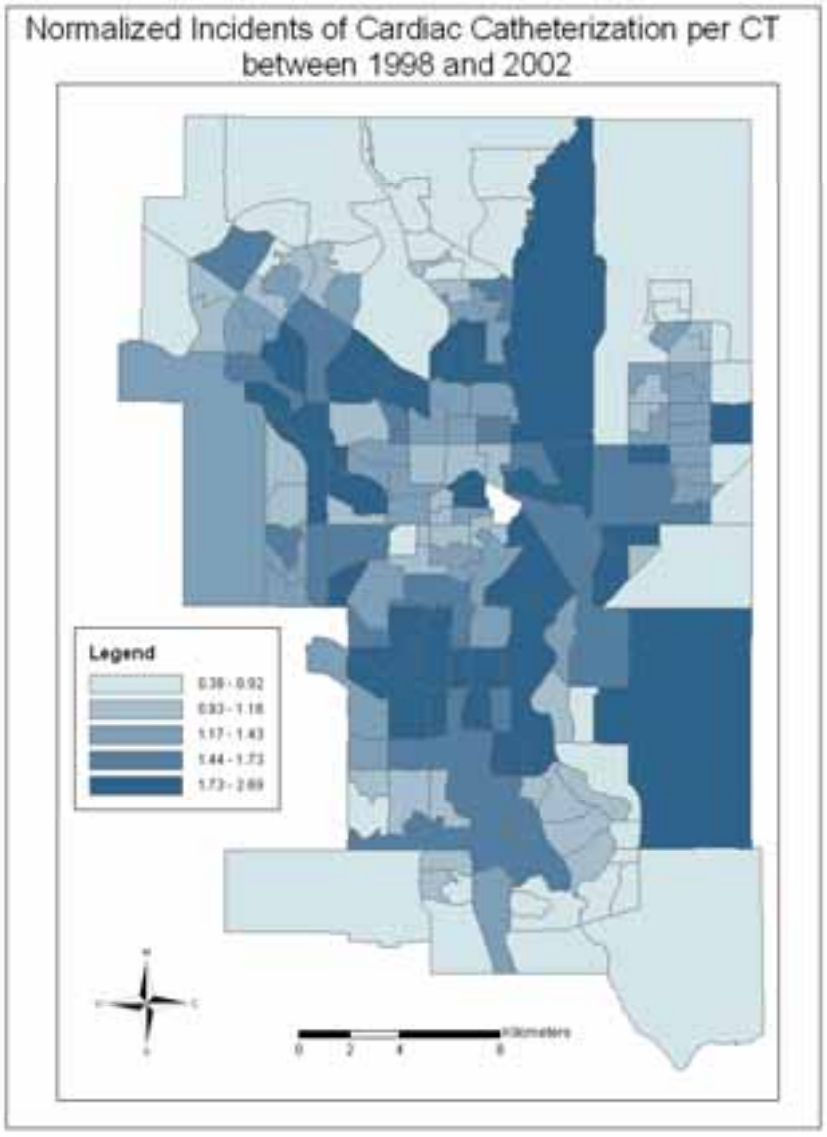
Revised Research Goals

1. To analyze and compare the results between Multiple Linear Regression (MLR) on Predictor Variables versus MLR on Extracted Components (OLS Method)

2. To implement the locally focused GWR as a tool for investigating the effects of spatial non-stationarity on regression modeling

Golder Associates

Dependent Variable and Study Area



Golder Associates

Available Variables

Predictor Variables used in Analysis (Per CT)				
Age	**Housing**	**Family**	**Income**	**Education**
% Pop. Ages 45-54 % Pop. Ages 55-64	Owned Dwellings Rented Dwellings	Number of Males Male to Female Ratio	Annual Household Income Unemployment Rate	Adults with Grade 13 or less education In University
% Pop. Ages 65-74	Separate Detached Homes	Widowed, Divorced and Separated		Non University Post Secondary Degrees
% Pop. Ages 75-84	Duplexes	2 Parent Families		University With Degrees
% Pop. Ages 84 +	Apartments	2 Parents w/Kids at Home		
		2 Parents No Kids at Home		
		Single Parent Households		
		Number of Children		
		Number of Single Adults		
		Married and Common Law		
		People per Family		

Golder Associates

Bivariate Pearson's Correlation

	n.cases	n.nu.chld	n.single	n.mar.claw	n.w.d.s	n.2par	n.2p.wchld	n.2p.nchld	n.sing.p	n.owned	n.rented	n.sdetach
n.cases	1.000											
n.males	-.144											
mf.ratio	-.132											
n.a45.54	.041											
n.a55.64	.569											
n.a65.74	.838											
n.a75.84	.720											
n.a85pl	.397											
ppl.fam	-.127											
n.nu.chld	-.361	1.000										
n.single	.105	-.735	1.000									
n.mar.claw	-.377	.757	-.911	1.000								
n.w.d.s	.577	-.664	.550	-.790	1.000							
n.2par	-.175	.092	-.163	.244	-.280	1.000						
n.2p.wchld	-.495	.931	-.737	.819	-.798	.156	1.000					
n.2p.nchld	.389	-.954	.586	-.600	.530	-.024	-.888	1.000				
n.sing.p	.317	-.173	.469	-.620	.702	-.280	-.450	-.008	1.000			
n.owned	-.285	.797	-.879	.909	-.711	.186	.818	-.662	-.499	1.000		
n.rented	.300	-.768	.879	-.914	.703	-.195	-.802	.633	.522	-.993	1.000	
n.sdetach	-.273	.751	-.800	.871	-.728	.148	.803	-.636	-.516	.912	-.901	1.000
n.duplex	.177	-.075	.341	-.431	.465	-.205	-.287	-.047	.723	-.348	.368	-.502
n.apts	.200	-.808	.710	-.736	.529	-.048	-.735	.754	.135	-.835	.822	-.827
n.gr13ls	.181	.134	.163	-.224	.378	-.215	-.087	-.250	.661	-.133	.151	-.120
n.non.uni	-.235	.111	-.046	.127	.035	.035	.044	-.152	.197	.099	-.094	.093
n.uni	-.051	-.154	-.115	.132	-.322	.163	.053	.267	-.620	.069	-.086	.060
n.uni.wd	-.036	-.155	-.130	.154	-.328	.166	.048	.275	-.629	.086	-.105	.078
unemp	.171	-.060	.328	-.367	.259	-.063	-.153	-.044	.407	-.347	.353	-.322
hh.m.inc.k	-.344	.648	-.763	.823	-.783	.254	.781	-.499	-.735	.816	-.821	.765

Golder Associates

Variables Selected for Regression

- Households with 2 Parents
- Households with 2 Parents without children living at home
- Male to Female Ratio
- Percentage Ages 45-54
- Percentage Ages 65-74
- Percentage Ages 75-84
- Percentage Ages 85 +
- Adults with University Degree
- Adults with Post Secondary Non-University Degree
- Married and Common Law Couples
- Unemployment Rate
- People Per Family
- Family Income
- Number of Owned Houses
- 2 Parent Households

Golder Associates

Standard Multiple Linear Regression

- Backwards Selection Regression Method
  - All variables are initially entered into the regression, and are subsequently removed according to their significance values
- The final model includes the remaining set of significant variables that describe the dependent variable

Golder Associates

Multiple Linear Regression Results

- Predictor Variables

Adjusted R^2: 0.790

  - Ages 65-74
  - University Degree
  - Ages 45-54
  - 2 Parents No Kids
  - Ages 85 Plus
  - Post Secondary Non-University Degree

- Components

Adjusted R^2: 0.775

  - 1
  - 2
  - 3
  - 4
  - 5

Golder Associates

Principal Component Analysis

- Following Thompson, the five components were named in as broad and descriptive a manner as possible. (Thompson, 2004)

- Factor scores were saved as Regression scores for further comparative analyses

Principal Component Extraction

Rotated Component Matrix[a]

	Component				
	1	2	3	4	5
n.nu.chld	.958	-.027	.113	.094	-.096
n.2p.wchld	.917	.151	-.080	.118	-.220
n.owned	.903	.029	-.348	.015	.081
n.apts	-.895	.180	.012	.015	-.020
n.2p.nchld	-.888	.102	-.279	-.063	.218
n.rented	-.883	-.035	.380	-.007	-.061
n.sdetach	.844	.006	-.400	.091	.051
n.mar.claw	.838	.056	-.473	.046	-.034
n.single	-.833	-.022	.364	.219	-.208
hh.m.inc.k	.744	.469	-.372	.038	-.003
n.w.d.s	-.671	-.361	.357	-.383	.235
ppl.fam	.660	-.071	.132	-.294	-.104
n.uni	-.076	.911	-.315	-.195	.014
n.uni.wd	-.068	.897	-.335	-.176	.033
n.gr13ls	.048	-.753	.489	.212	.134
n.non.uni	.089	-.729	-.210	.053	-.302
n.a45.54	.474	.592	.188	.173	.312
n.duplex	-.116	-.250	.733	-.196	-.062
n.sing.p	-.274	-.521	.712	-.143	.047
unemp	-.138	.046	.690	.197	.081
n.2par	.099	.141	-.305	.193	-.033
n.males	-.193	-.137	-.017	.920	-.021
mf.ratio	-.226	-.131	-.012	.909	-.015
n.a85pl	-.388	.129	.102	-.730	.146
n.a75.84	-.504	.085	.019	-.622	.482
n.a55.64	.120	.189	.090	.053	.883
n.a65.74	-.279	.002	-.044	-.353	.816

Extraction Method: PCA with Varimax Rotation

1
2
3
4
5

Components

1. Young Families and Housing Type
2. Level of Education
3. Lower Societal Position
4. Male Population and *negative* relationship to eldest population
5. Middle Age and Retiree Age Group

Golder Associates

Total Variance Explained

Total Variance Explained

Comp#	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	10.469	38.774	38.774	10.469	38.774	38.774	9.441	34.965	34.965
2	5.015	18.573	57.347	5.015	18.573	57.347	3.962	14.676	49.641
3	3.272	12.120	69.467	3.272	12.120	69.467	3.327	12.321	61.962
4	1.887	6.988	76.455	1.887	6.988	76.455	3.322	12.305	74.267
5	1.559	5.773	82.227	1.559	5.773	82.227	2.149	7.961	82.227

Golder Associates

Multiple Linear Regression Results

- Predictor Variables

Adjusted R^2: 0.790

  - Ages 65-74
  - University Degree
  - Ages 45-54
  - 2 Parents No Kids
  - Ages 85 Plus
  - Post Secondary Non-University Degree

- Components

Adjusted R^2: 0.775

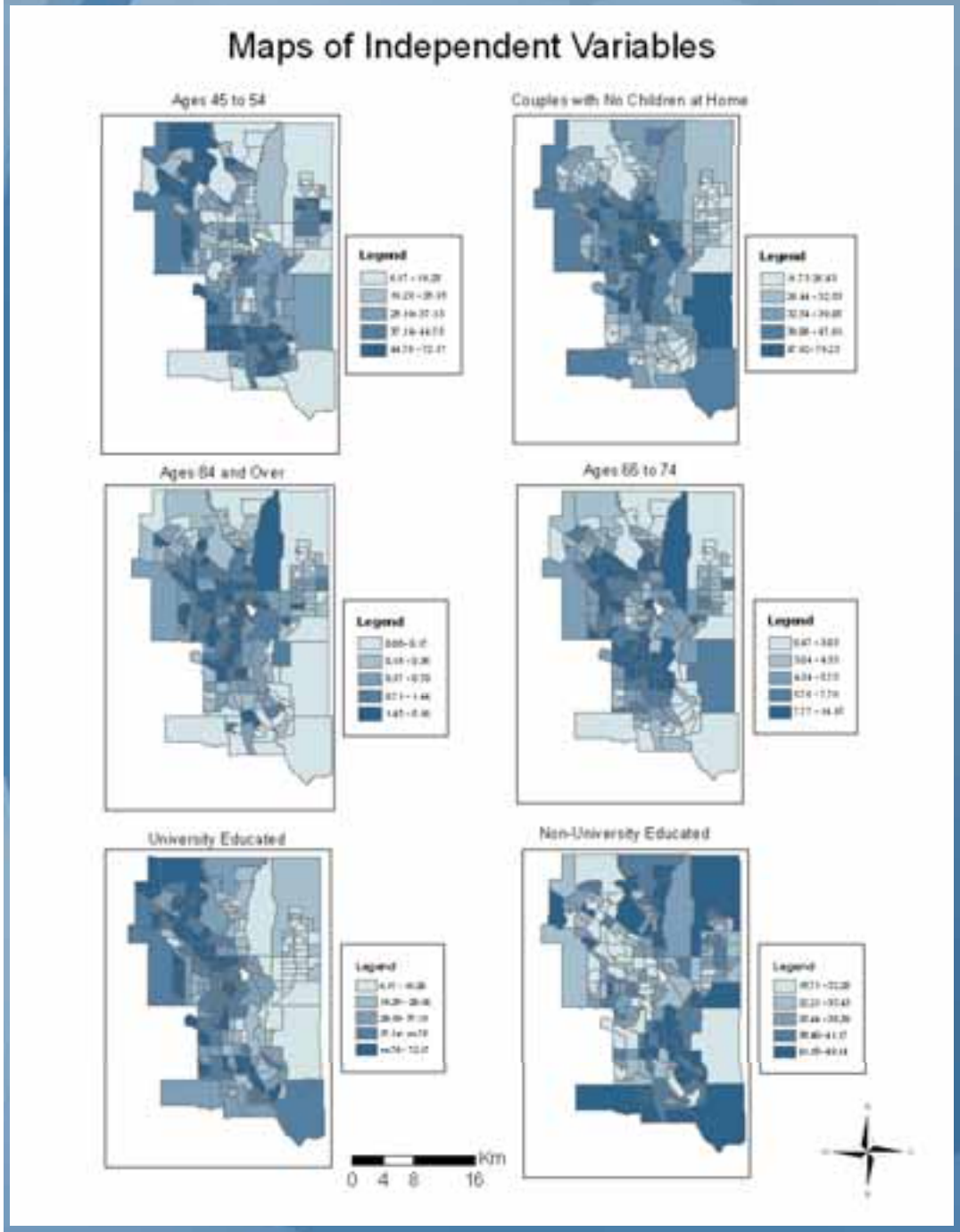
  - Family Structure and Housing Type
  - Level of Education
  - Lower Societal Position
  - Male Population and *negative* Population above the Age of 75
  - Middle Age and Retiree Age Group

Golder Associates

Standardized Beta Coefficients

Variable	Beta		Component	Beta
Ages 65 to 74	**0.722**		Family and Housing	**0.771**
University Degree	**-0.408**		Education Level	**-0.299**
Ages 45 to 54	**0.253**		Low Societal Status	**-0.248**
2 parents with no children at home	**0.204**		Males and negative age groups 75 and above	**0.165**
Ages 85 and over	**0.133**		Middle Age and Retirees	**-0.095**
Post Secondary Degree	**-0.121**			

Golder Associates

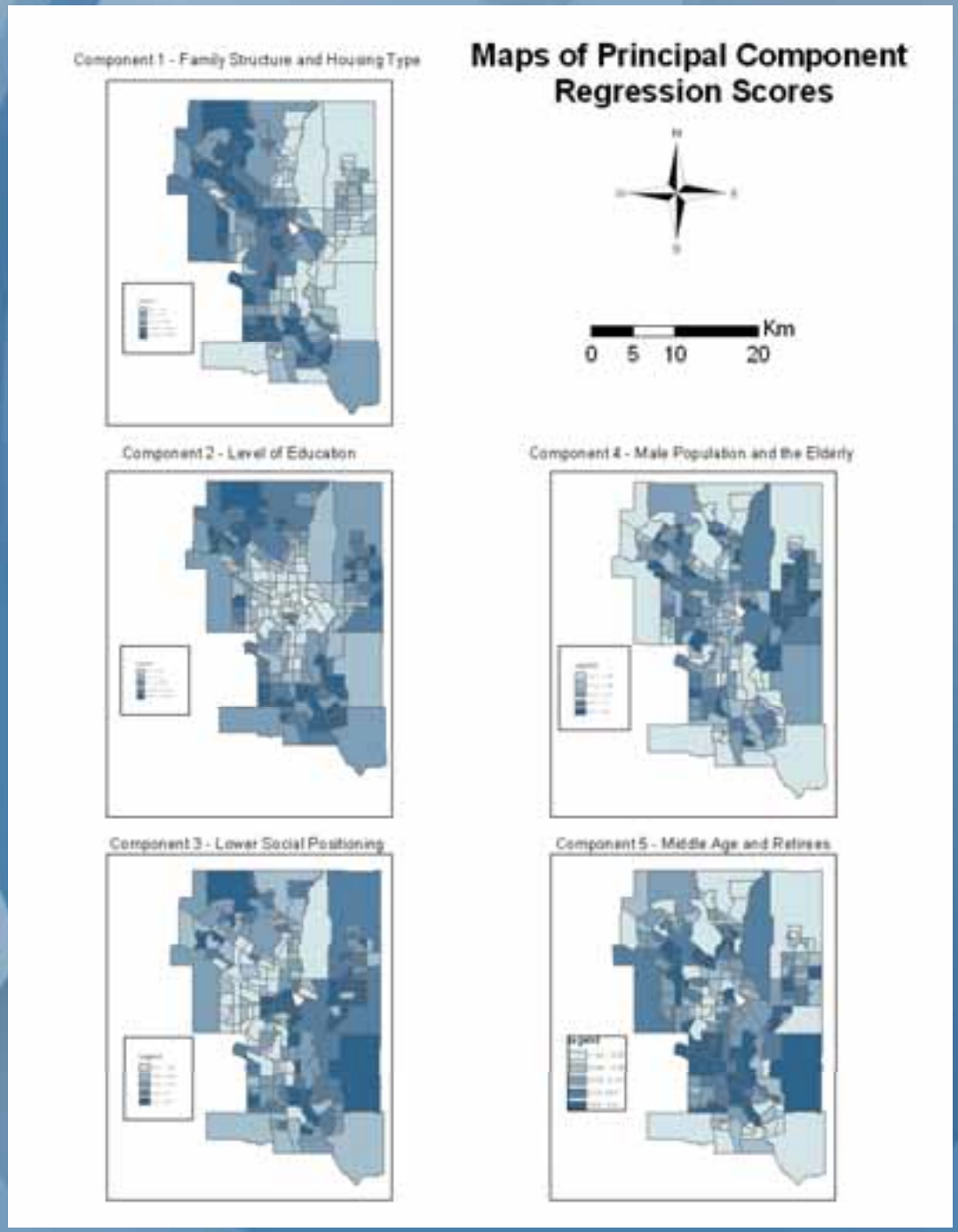


- Dominance of University Degree holders in the West. Non-University post secondary degrees more prominently observed in the east.

- Residents aged 65 to 74 are concentrated in the older communities nearer downtown*.

- Groups aged 45 to 54 dominantly live in newer suburbs on the outer periphery of the city.

*Bear this fact in mind for later



•Family Structure and Housing Type shows a NW-SE trend.

• The highest levels of post secondary education are seen in the periphery of the city, with lowest levels found in the inner city and older suburbs.

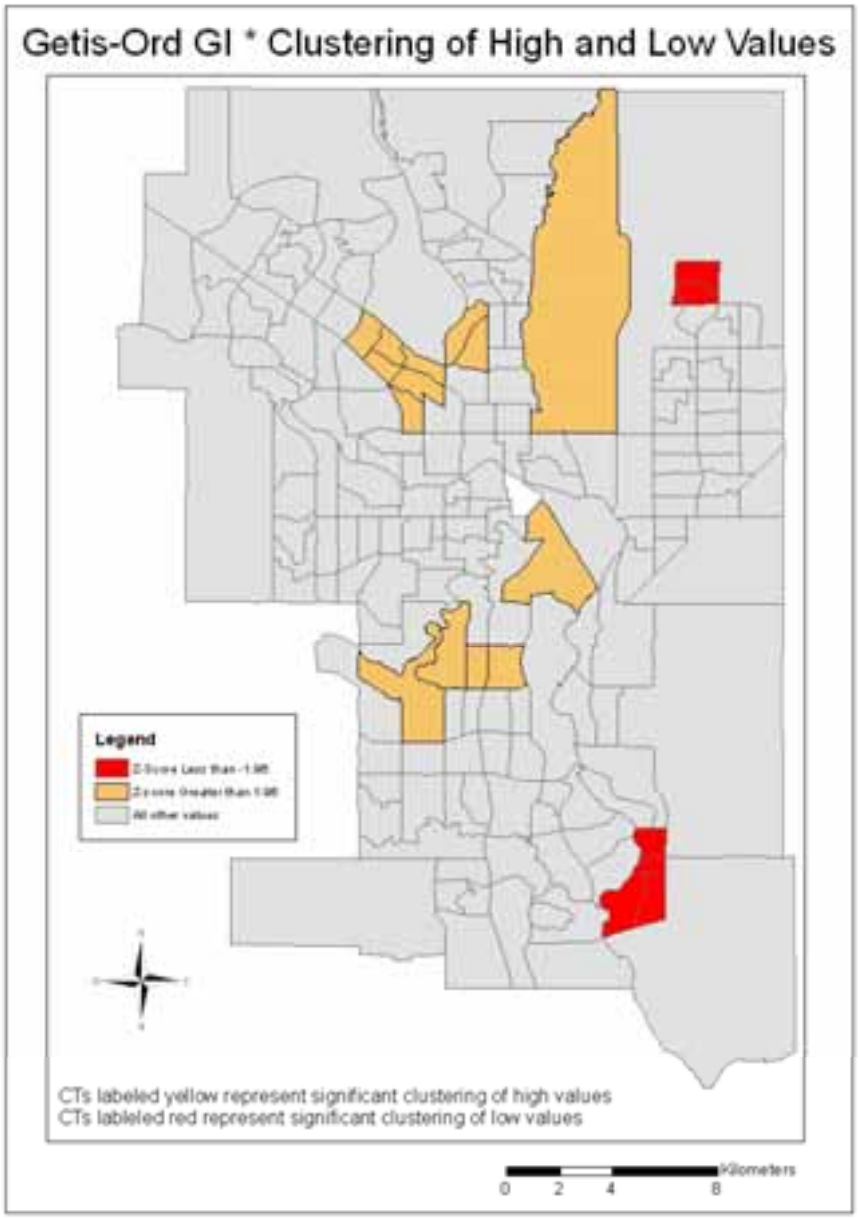
•High proportion of young males and lower social positioning in the NE of the city, with anomalies in the NW visible

Spatial Non-Stationarity

- Social processes are often non-stationary across space
- Observed spatial non-stationarity *within* a dependent variable may lead to examination of spatial non-stationarity *between* variables.

Golder Associates

Getis-Ord Gi* Statistic



The Gi* statistic is a method of examining local relationships.

Census tracts highlighted in orange have been found to display significant clustering of high values in the dependent variable

Census tracts marked in red represent significant clustering of low values in the dependent variable

Geographically Weighted Regression

- Founded by Fotheringham, Brunsdon and Charlton
- Promoted as a means of removing spatial non-stationarity through local analysis
- **KEY POINT**: A researcher cannot know whether spatial non-stationarity exists in a modeled relationship unless GWR is run

Golder Associates

Standard Regression Vs. GWR

Standard Linear Regression Equation

$$Y = a_0 + \sum_{k=1}^{p} a_k x_k + \varepsilon$$

GWR Equation

$$Y_i = a_{i0} + \sum_{k=1}^{P} a_{ik} x_{ik} + \varepsilon_i$$

$i = 1,2,\ldots,n$

Where a_{ik} is the value of the *k*th parameter at location *i*

k represent the variables or components, and x represents the parameters

Golder Associates

Geographically Weighted Regression

- Bandwidth Selection
  - AICc Method
- Adaptive Kernel Selection
  - Attempts to correct for boundary effects
  - Used here also due to community size differences

Golder Associates

GWR Results

- All diagnostic indicators of GWR effectiveness were *slightly* better than their global model counterparts for both variables and components *
- Very small difference in Adjusted R^2 value, with improvements of 0.011 and 0.013 respectively.
- Spatial Non-Stationarity was found to be present in 4 of the 6 variables
- Spatial Non-Stationarity was found to be present in 4 of the 5 Components
- Problems with apparent oversmoothing, assumed to be due to huge bandwith and adaptive kerneling.

*However, since these diagnostics represent *averages* of the GWR points, the overall "locality" of the model is removed.

Golder Associates

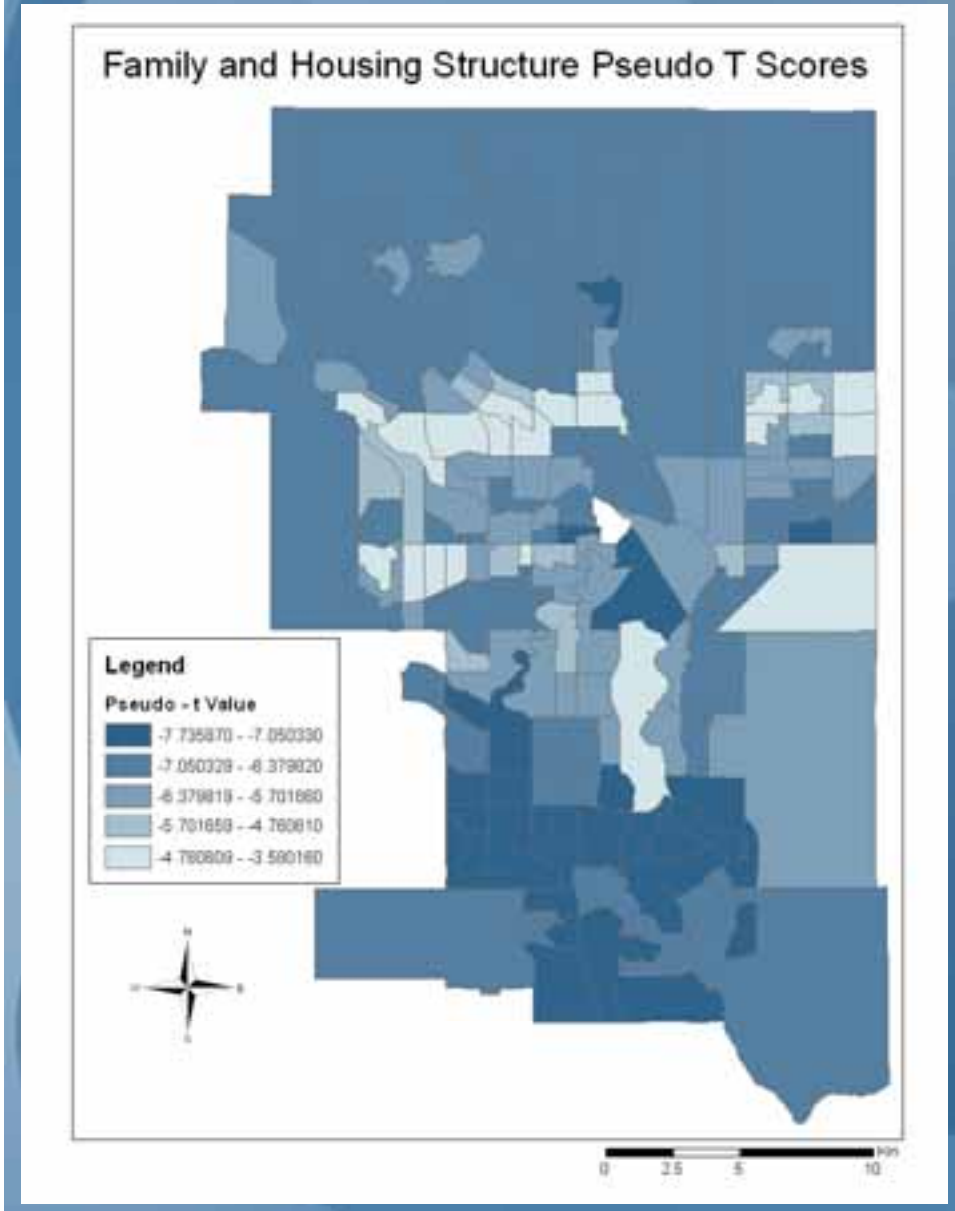
Pseudo T Score Significance

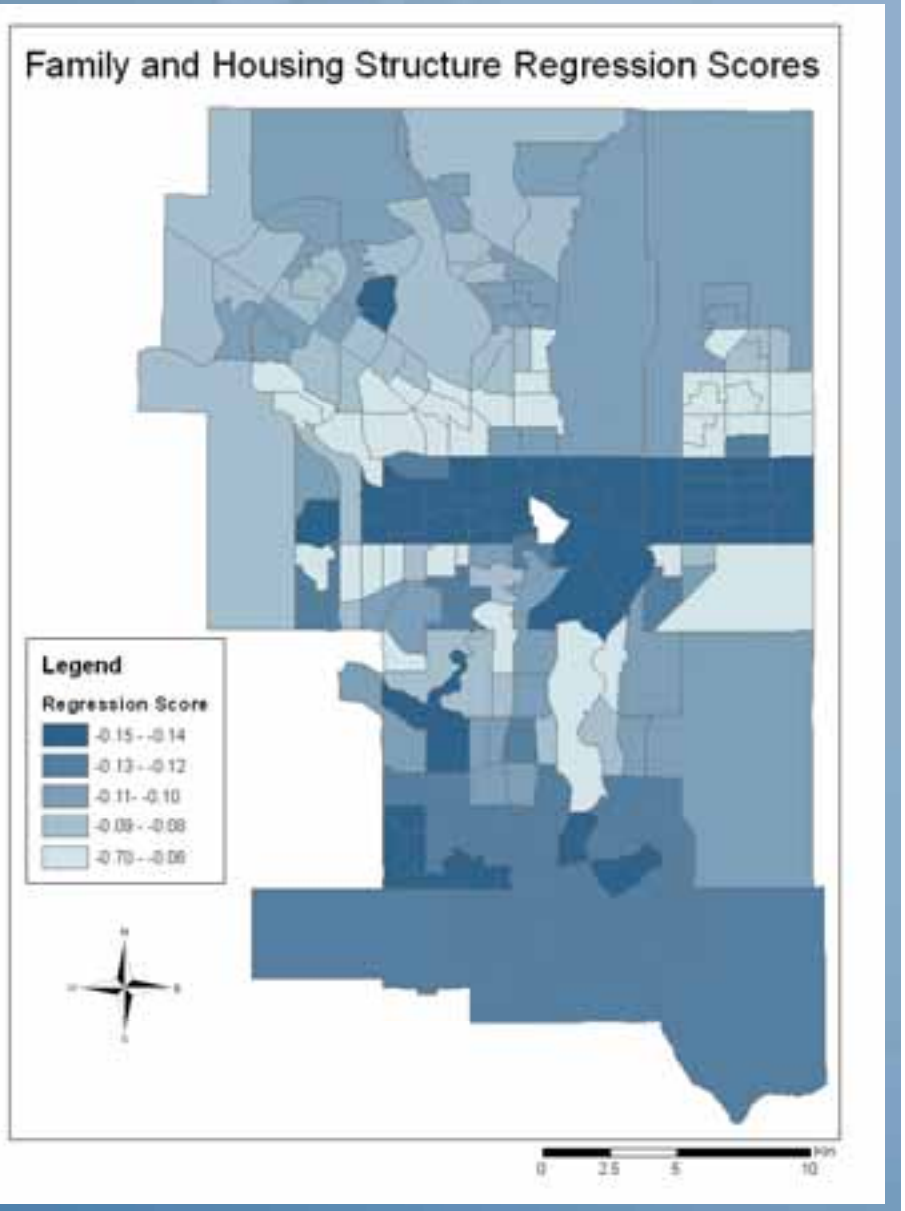
Pseudo T Scores from GWR

Variable	T Score Range	Component	T Score Range
Ages 45-54	2.72 to 5.56	Family/Housing	-3.58 to -7.73
Ages 65-74	12.66 to 17.26	Education Level	-1.53 to 3.13
Ages 85+	0.015 to 0.018	Lower Societal Position	1.61 to 4.48
2 Parents, No kids at home	0.001 to 0.002	Males and *negative* 74+	-4.23 to -7.44
Post-Secondary Degree	0.004 to 0.006	Middle Age and Retirees	12.29 to 18.27
University Degree Holders	0,002 to 0.003		

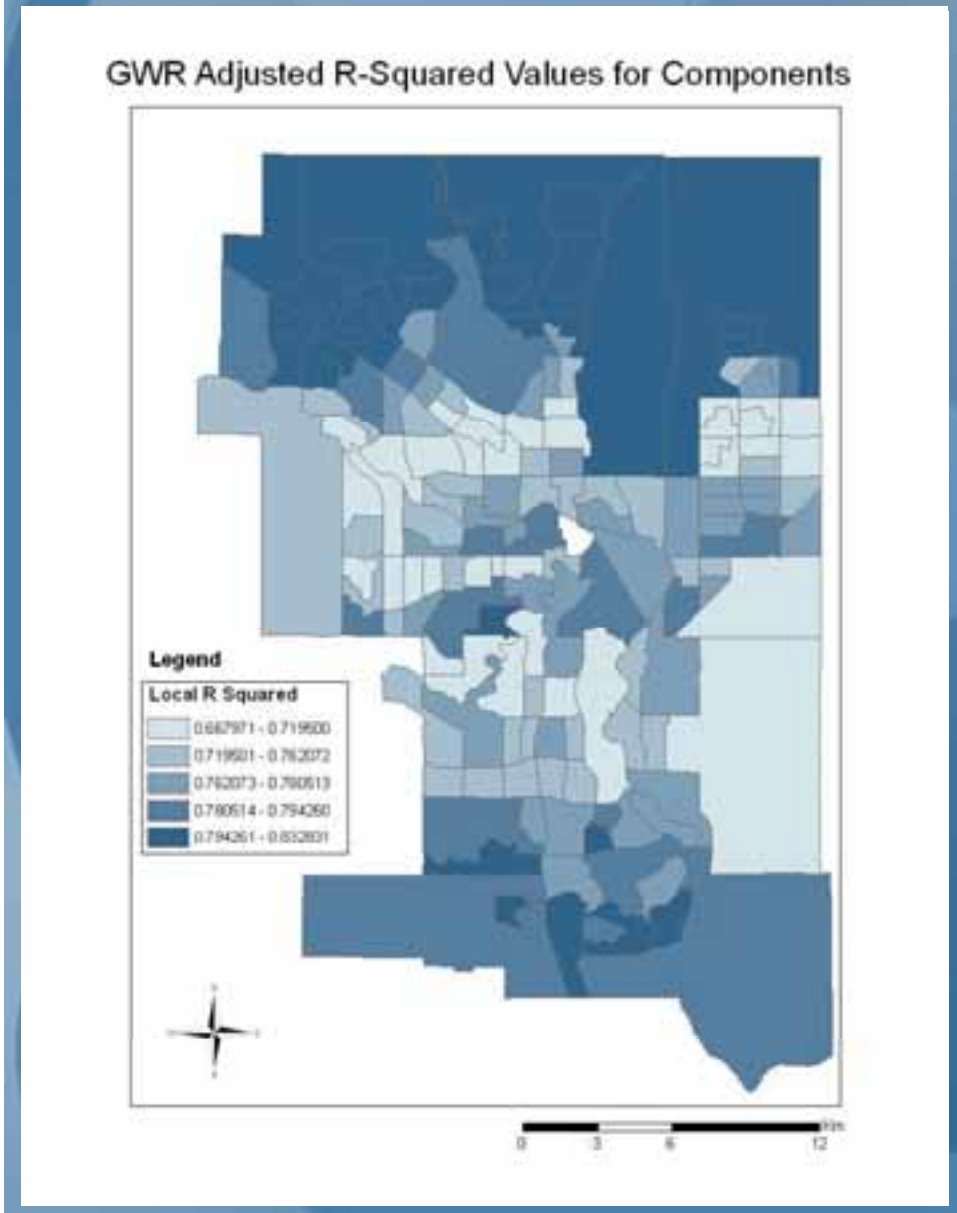
Golder Associates

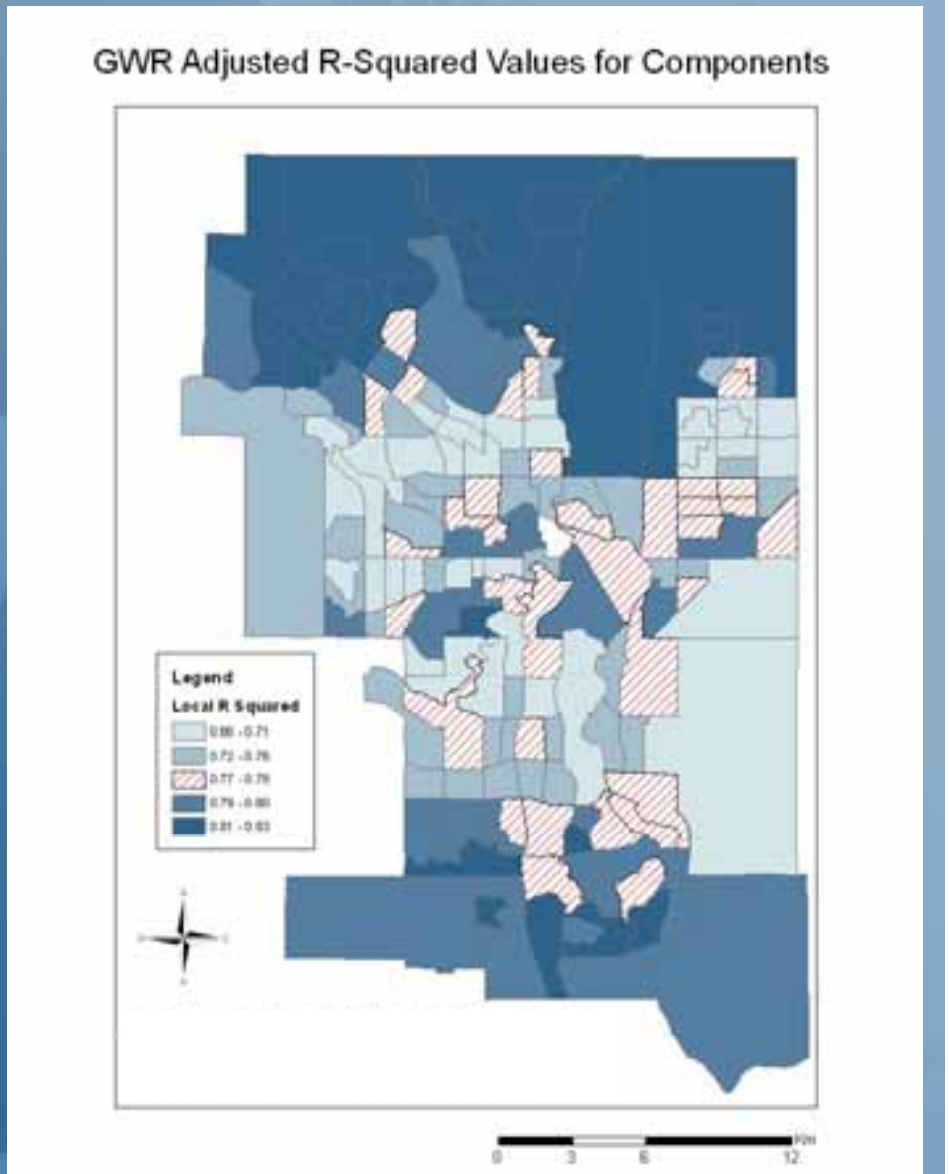
Pseudo T-Score Vs. Regression Score



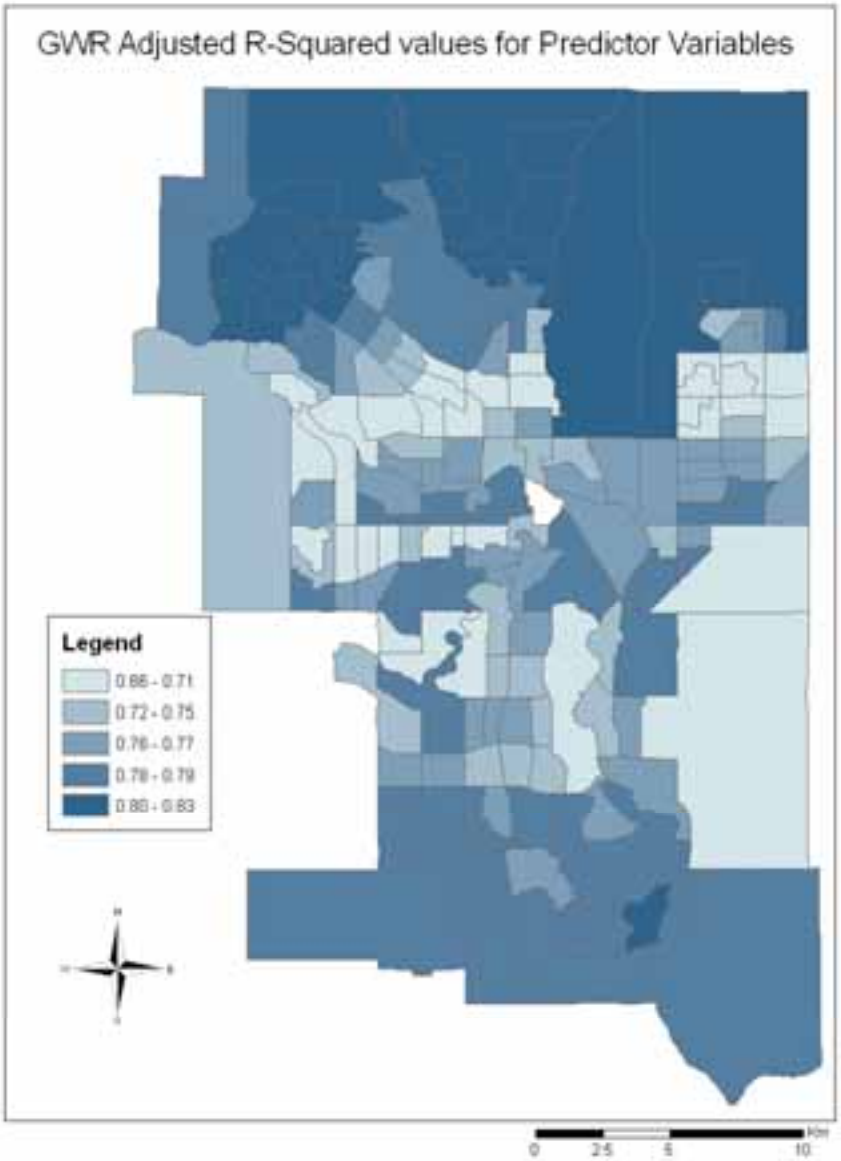


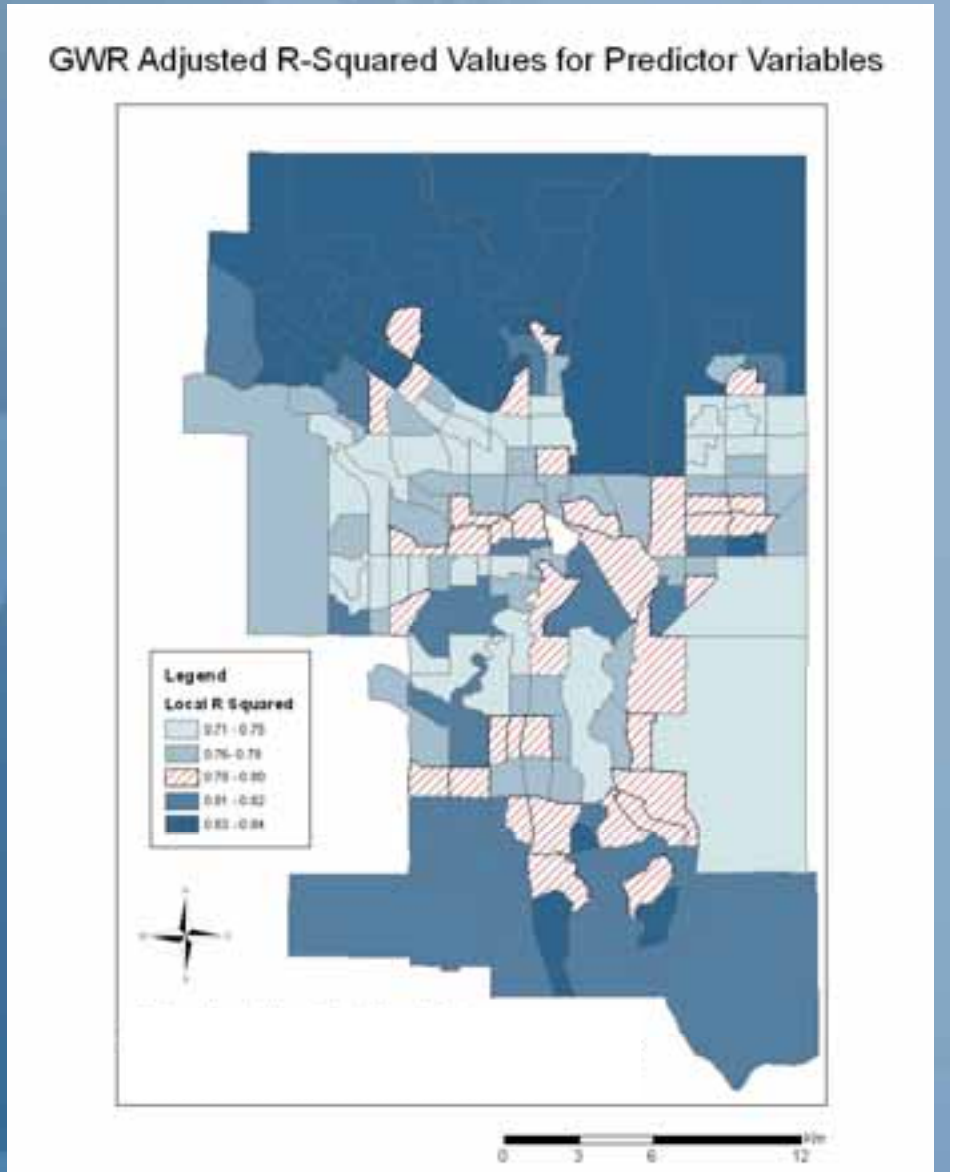
Comparison of Adjusted R-Squared between GWR and Standard Regressions for Components



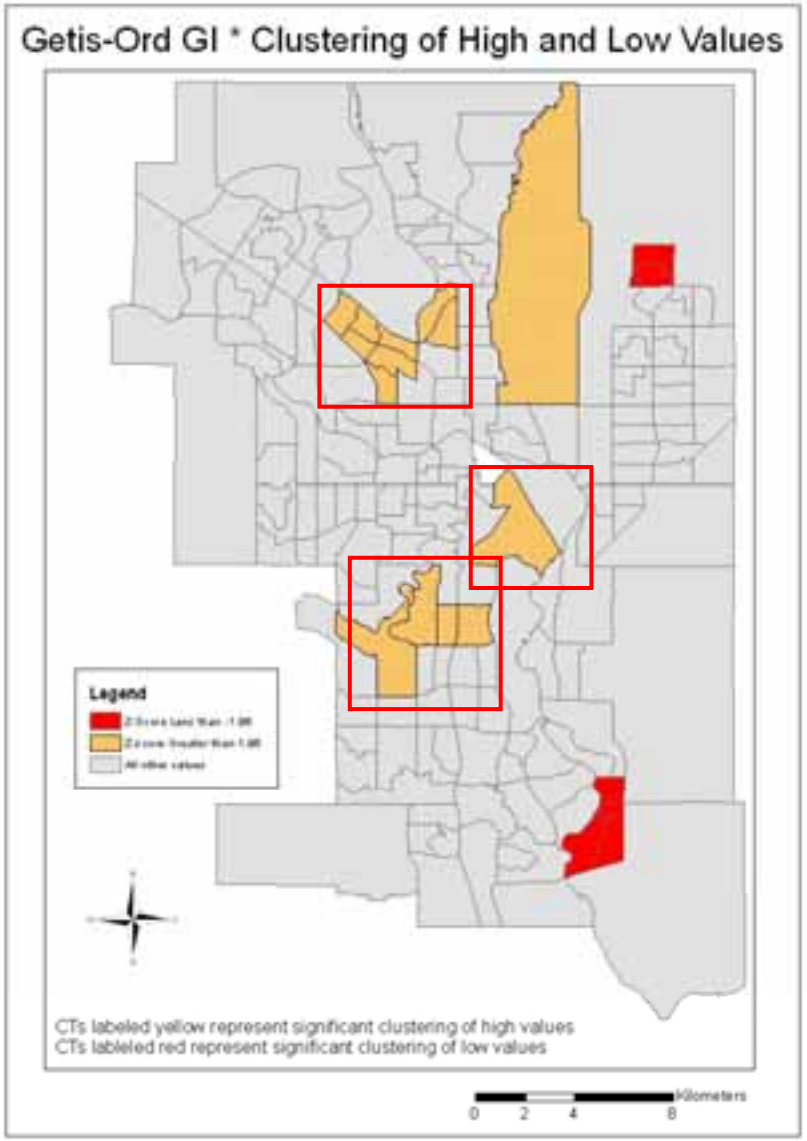


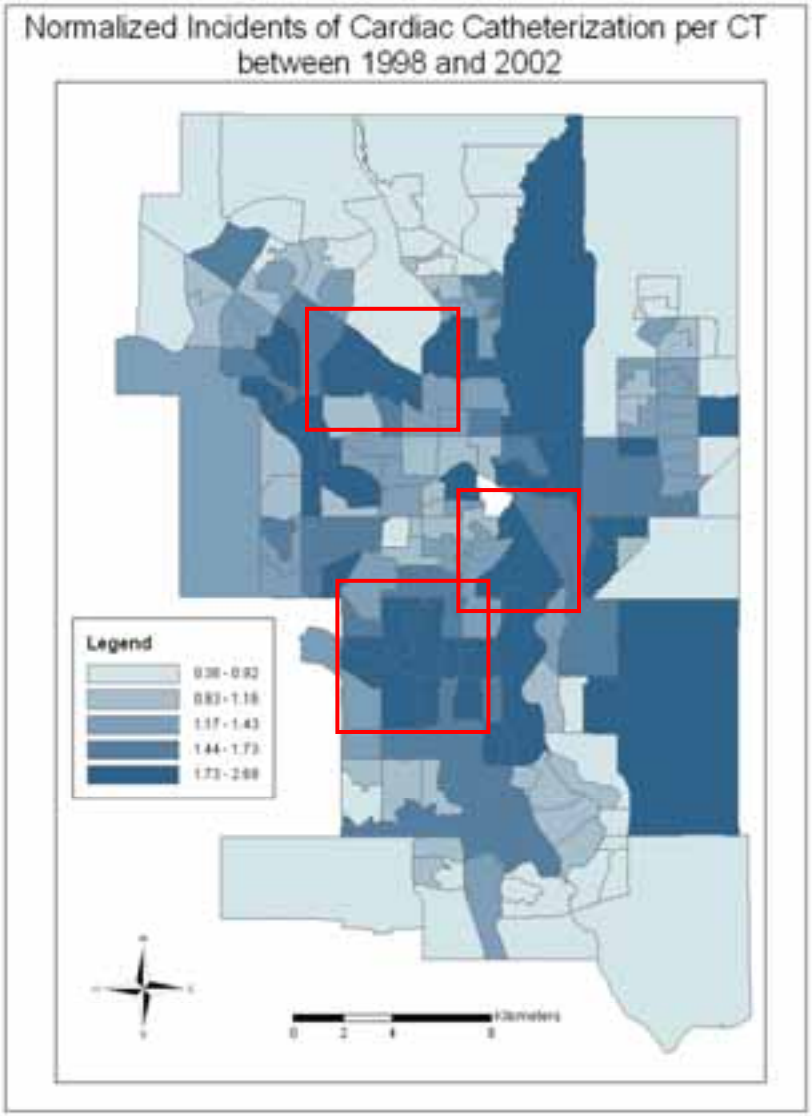
Comparison of Adjusted R^2 between GWR and Standard Regressions for Predictor Variables





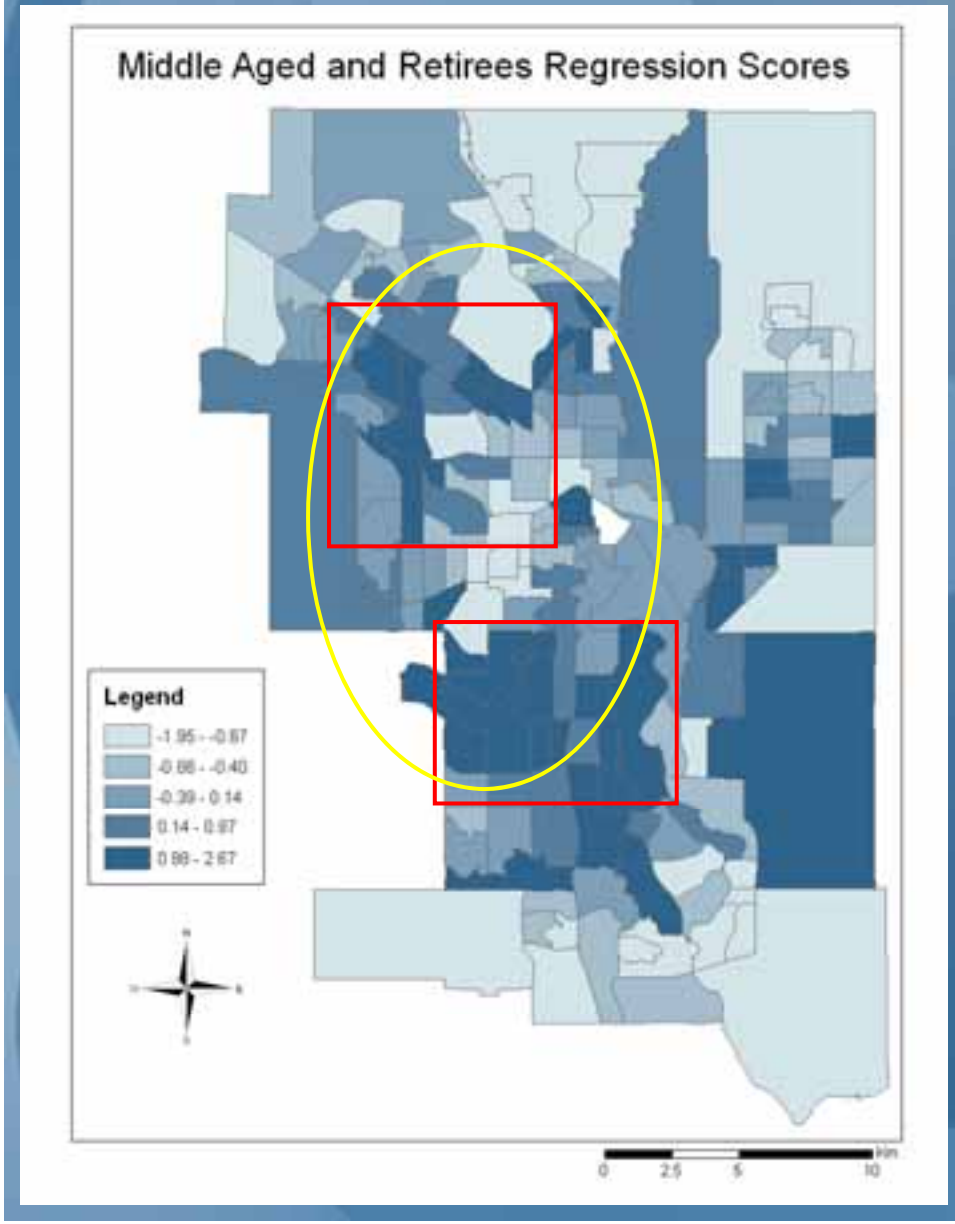
Focus Areas for Further Studies



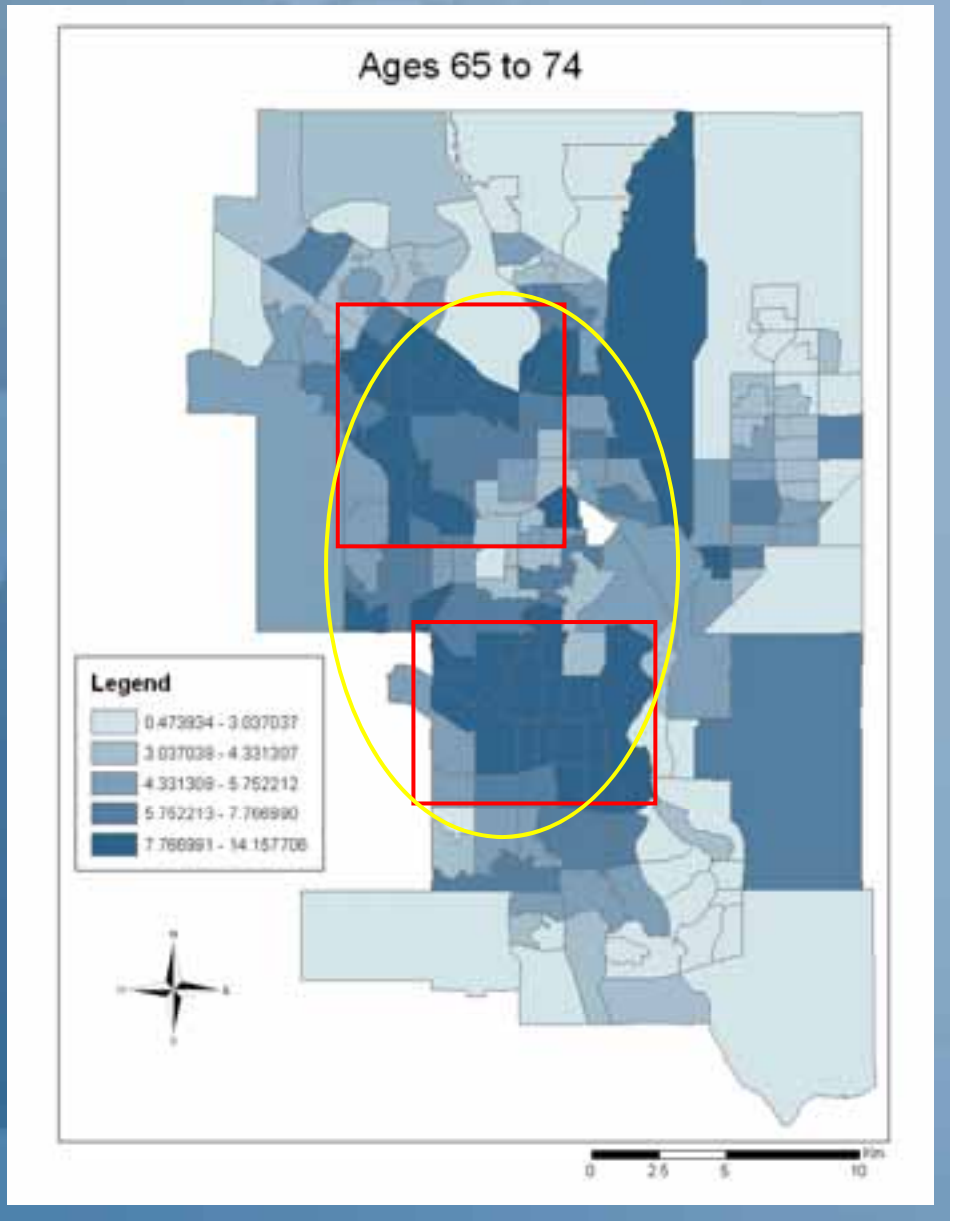


Focus Areas for Further Studies

Component



Variable



Discussion

- GWR not convincingly “localized” in this case. Most effective bandwidths were measured at 148 nearest neighbors for components, and 169 nearest neighbors of variables out of 181. Arc 9.3*
- Some Census Tracts should be eliminated to reduce skewing of data, and for cartographic purposes.
- A new census is coming out, making the study repeatable over another 5 years
- Dataset limited by the census. Larger datasets including ethnicity, diet, obesity, number of smokers are likely to follow.
- As more data becomes available, it could lead to further studies in Binary Logistic Regression on myocardial infarction likelihoods.

Golder Associates

Conclusions

- According to diagnostics, PCA proved to be just as effective a predictive modeling tool as MLR without sacrificing variance. When the two methods were compared, *PCA* was determined to be more useful analytically due to the broader explanation provided by components. Components were also more significant in GWR.

- GWR made slight improvements on both standard linear regressions, but improvements were minimal.
  - GWR highlighted non-spatial variables and components individually, suggesting non-stationarity in the process.
  - However, the overall improvement of R^2 scores observed in GWR models were *based on an average of the local scores*, which makes the model no less global than MLR regressions.

- A directional trend was observed through the use of these methods, highlighting spots worthy of detailed research

Golder Associates

Thank You

Questions?

Please come and say Hello after the session

jdnorris@ucalgary.ca

jnorris@golder.com

Golder Associates

American Heart Association: Link as of June 2

http://www.americanheart.org/presenter.jhtml?identifier=4491

APPROACH., Alberta Provincial Project for Outcome Assessment in Coronary Heart Disease: Link as of June 2