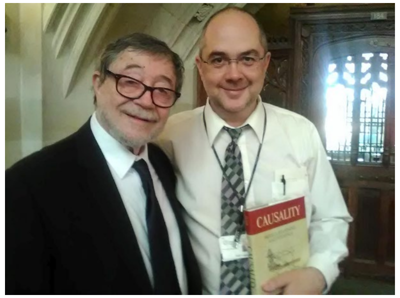
Modern Spatial Path Analytic Tools to Investigate the Geography of Medical Debt across a US State

Modern Modeling Methods – 2024, Storrs CT,
June 25-26, 2024 **Session 1B: Modeling Spatial Data**

Slides at https://tinyurl.com/mmmdebtct



Emil Coman, PStat, SEMNET ‘moderator’; Samuel Bruder

comanus@gmail.com

tinyurl.com/agecause

knowledge as a human right

Ⓒ Humanity, Earth, Milky Way, Universe

General plan

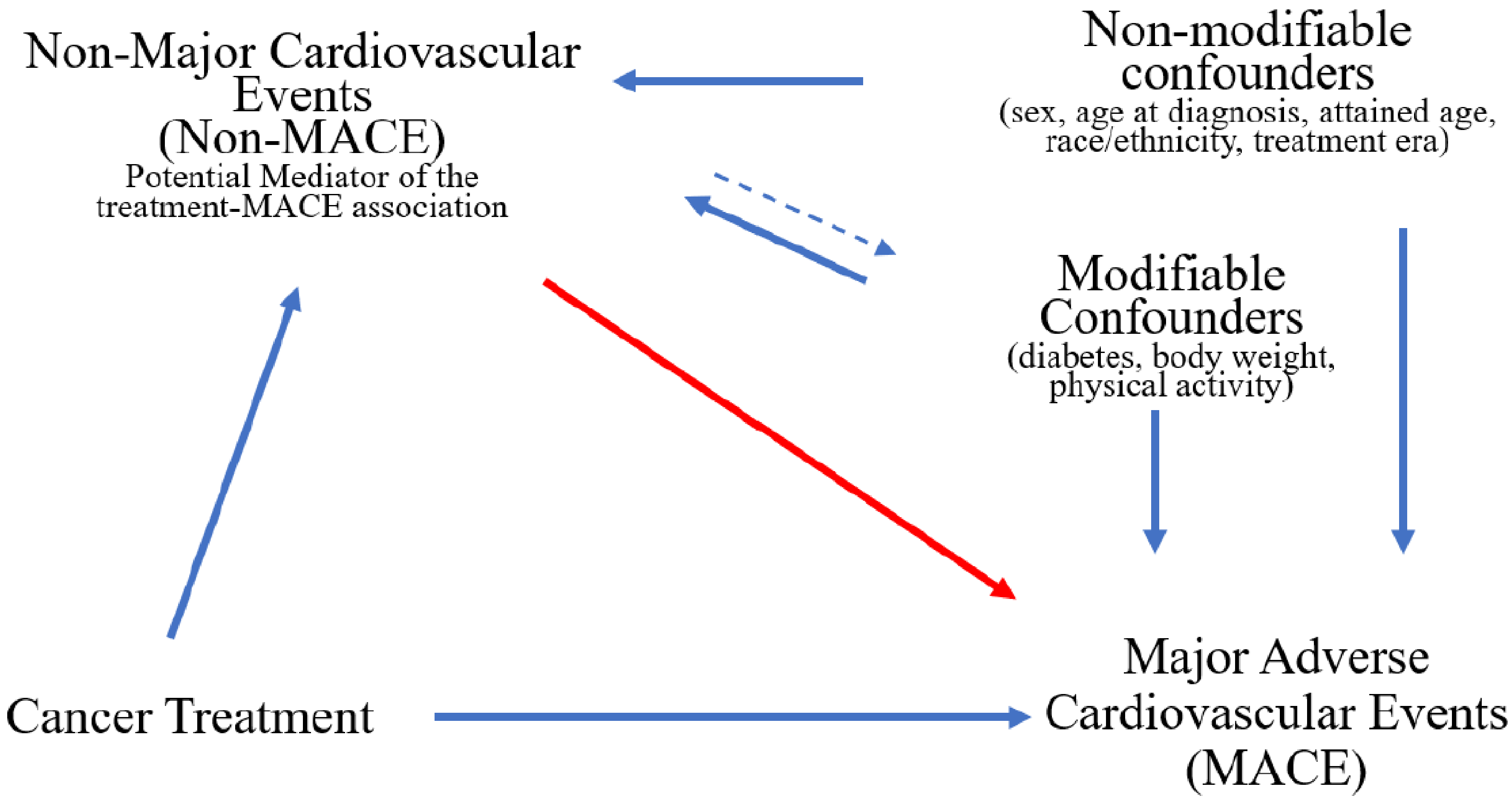
1. Causal logic with spatial data

2. Spatial non-independence intuition
   - Modeling solutions

3. Example with N=8 regions

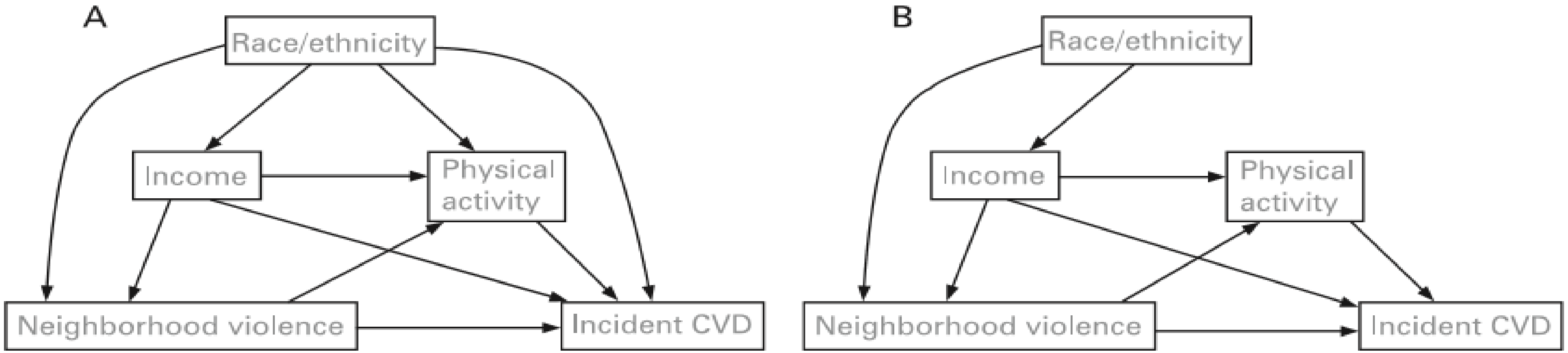
'Visual' causal reasoning

Figure S2· Directed acyclic graph showing selected factors involved in the lifetime risk of major adverse cardiovascular events (MACE) after childhood cancer survivorship



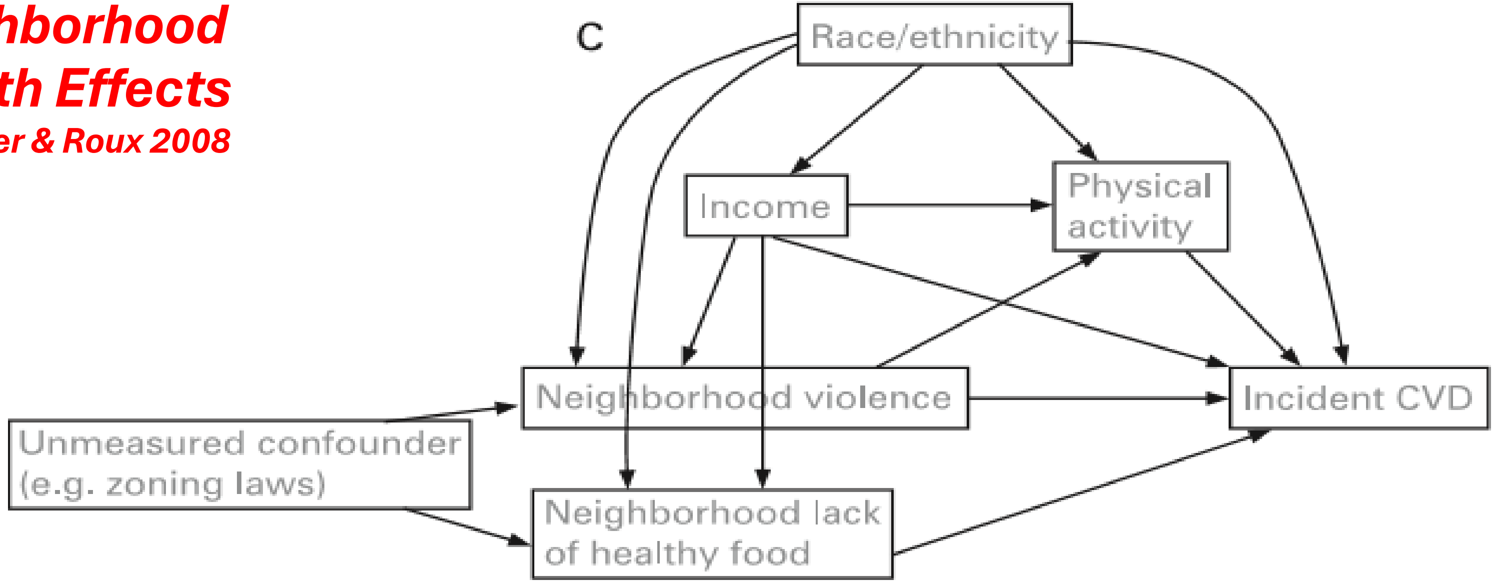
Rueegg, C. S. (2024). Directed acyclic graphs to foster transparency and scientific dialogue. The Lancet Oncology, 25(6), 693-694. doi:10.1016/S1470-2045(24)00271-7
Hammoud, R. A., Liu, Q., Dixon, S. B., Onerup, A., Mulrooney, D. A., Huang, I. C., . . . Armstrong, G. T. (2024). The burden of cardiovascular disease and risk for subsequent major adverse cardiovascular events in survivors of childhood cancer: a prospective, longitudinal analysis from the St Jude Lifetime Cohort Study. The Lancet Oncology, 25(6), 811-822. 6. Westreich, D., & Greenland, S. (2013). The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. American Journal of

Figure 1 Using directed acyclic graphs to identify variables that need to be controlled for in estimating neighbourhood health effects.



Neighborhood Health Effects

Fleischer & Roux 2008



Fleischer, N., & Roux, A. D. (2008). Using directed acyclic graphs to guide analyses of neighbourhood health effects: an introduction. *Journal of Epidemiology & Community Health, 62(9), 842-846.*

Causal reasoning

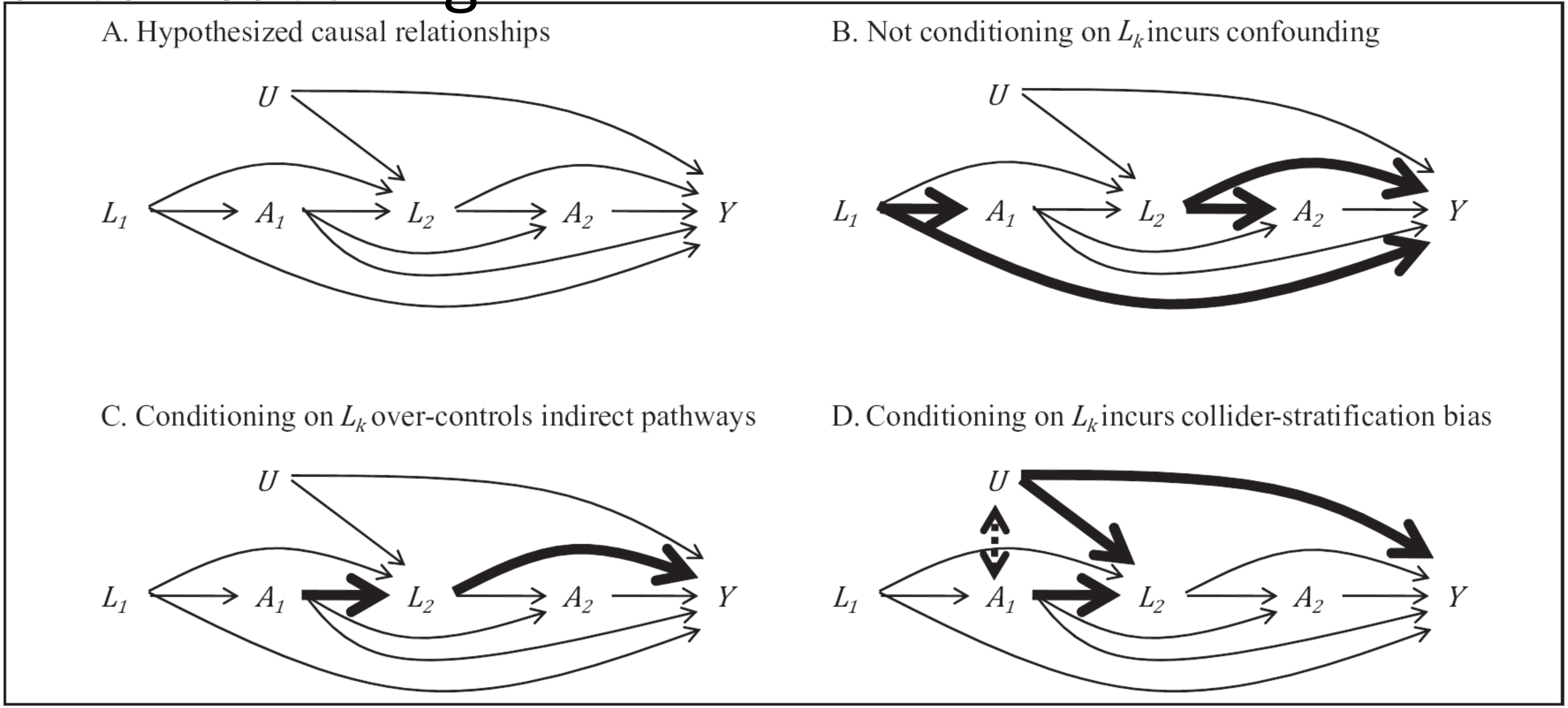
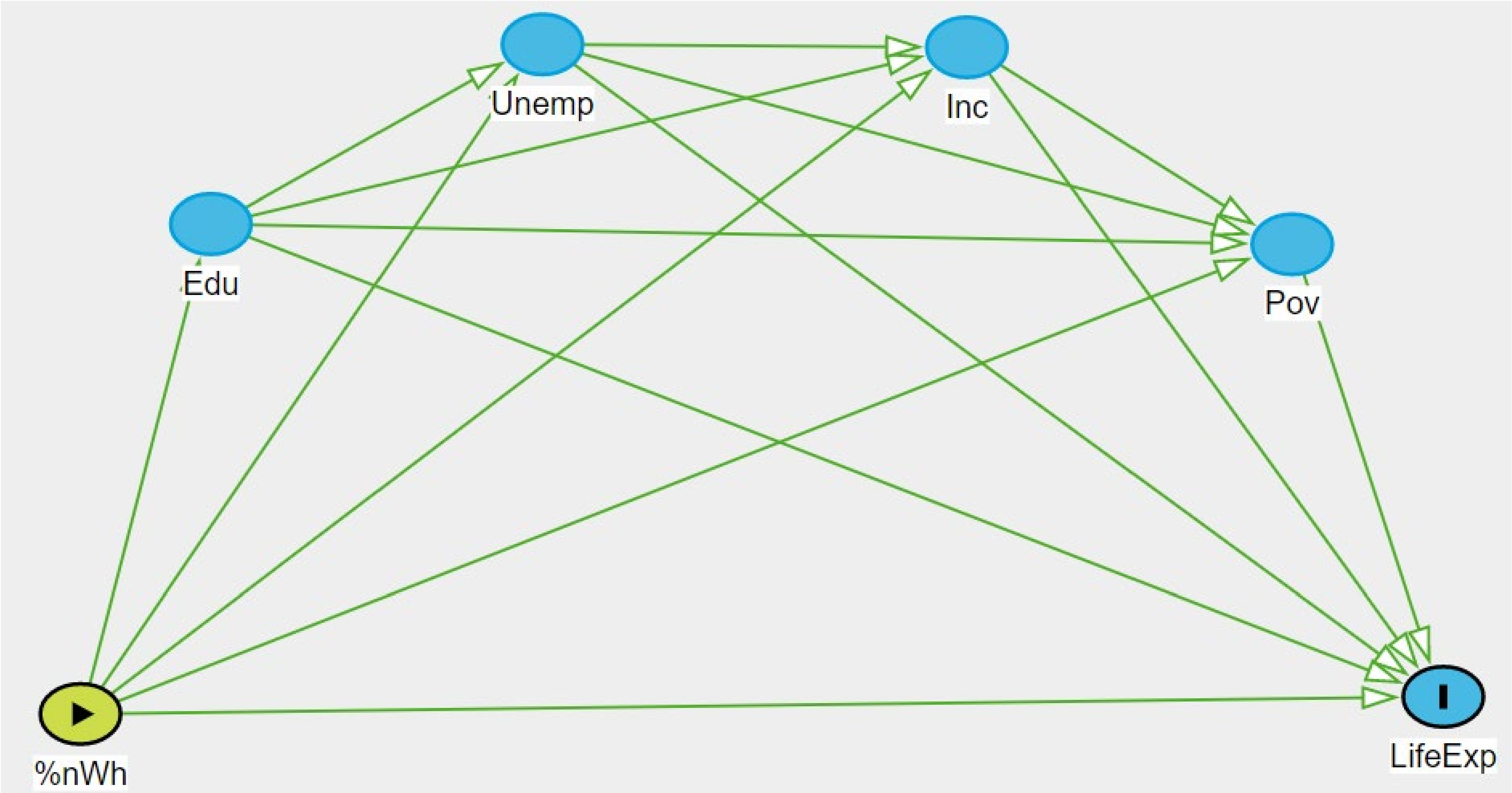


Figure 1. Causal Graphs for Exposure to Disadvantaged Neighborhoods with Two Waves of Follow-up

Note: A_k = neighborhood context, L_k = observed time-varying confounders, U = unobserved factors, Y = outcome.

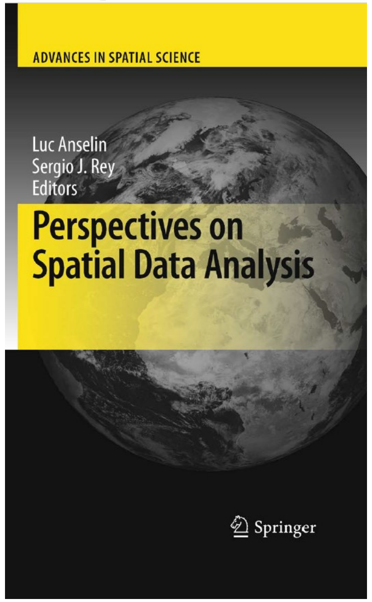
Wodtke, G. T., Harding, D. J., & Elwert, F. (2011). Neighborhood effects in temporal perspective: The impact of long-term exposure to concentrated disadvantage on high school graduation. *American Sociological Review, 76(5), 713-736.*

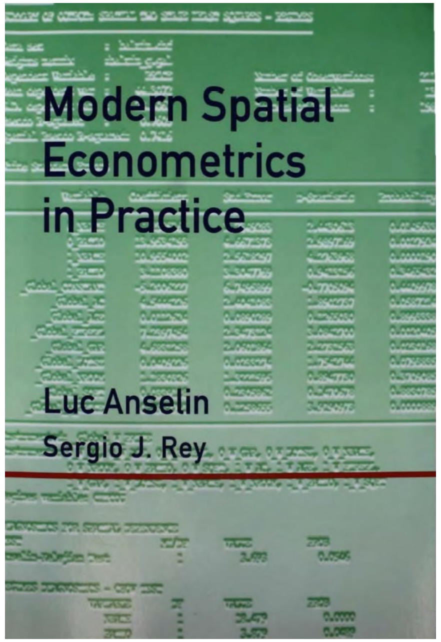
Life Expectancy and its determinants
http://dagitty.net/dags.html?id=GtvICQ

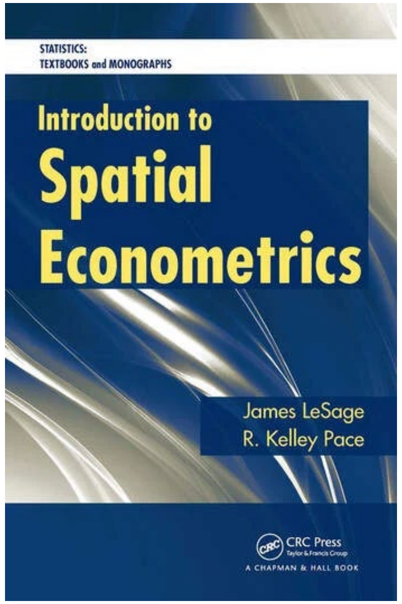


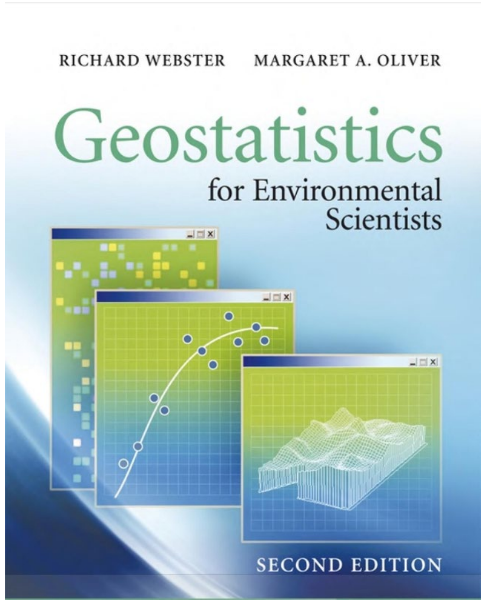
Simplest model' is: everything relates to everything: the 'saturated' model ('reference' in path analytic/SEM lingo)

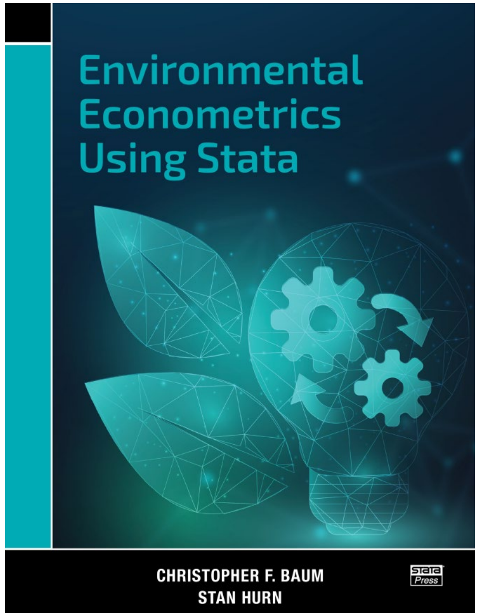
Quick look:

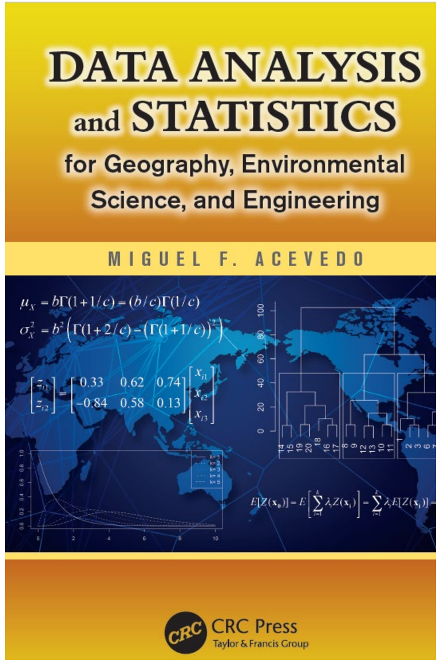

ADVANCES IN SPATIAL SCIENCE
Luc Anselin
Sergio J. Rey
Editors
Perspectives on Spatial Data Analysis
Springer

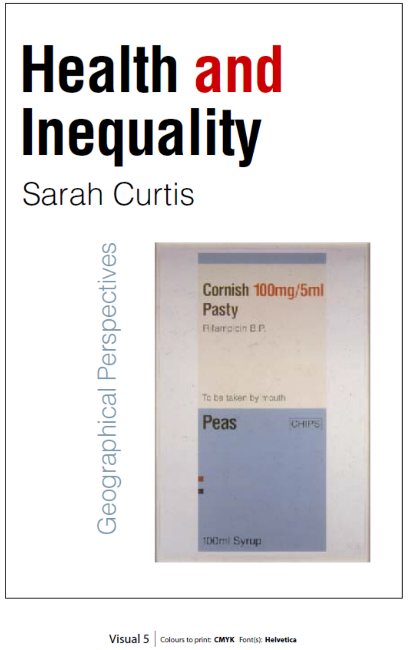

Modern Spatial Econometrics in Practice
Luc Anselin
Sergio J. Rey

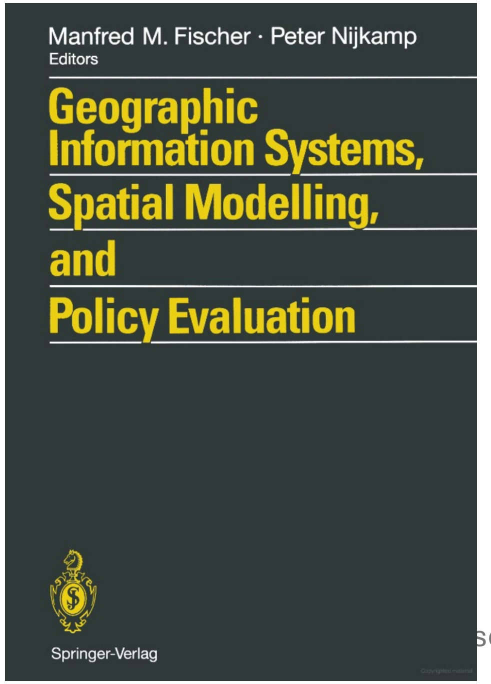

STATISTICS: TEXTBOOKS and MONOGRAPHS
Introduction to Spatial Econometrics
James LeSage
R. Kelley Pace
CRC Press
A CHAPMAN & HALL BOOK

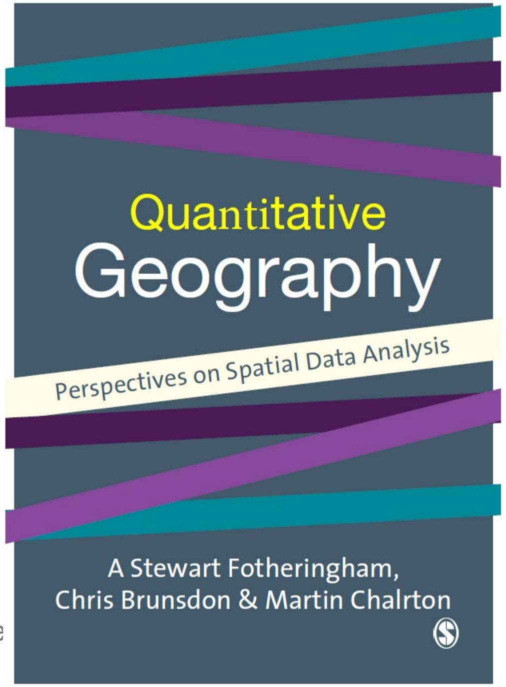

RICHARD WEBSTER
MARGARET A. OLIVER
Geostatistics
for Environmental Scientists
SECOND EDITION

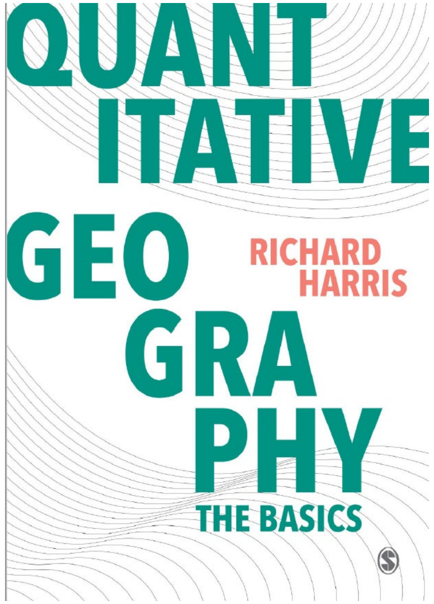

Environmental Econometrics Using Stata
CHRISTOPHER F. BAUM
STAN HURN


DATA ANALYSIS and STATISTICS
for Geography, Environmental Science, and Engineering
MIGUEL F. ACEVEDO
CRC Press
Taylor & Francis Group

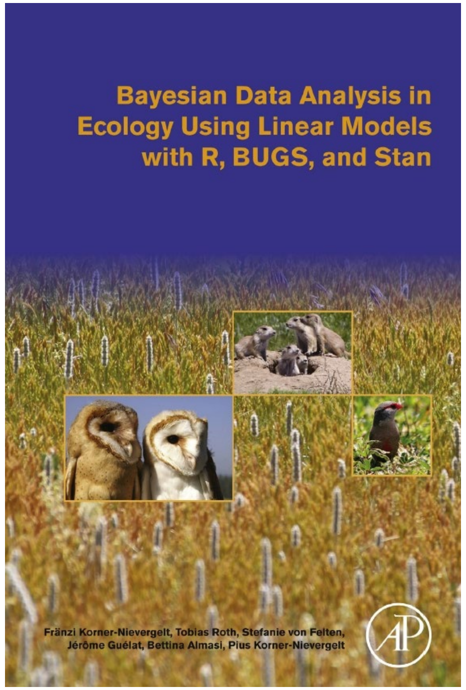

Health and Inequality
Sarah Curtis
Geographical Perspectives
Visual 5

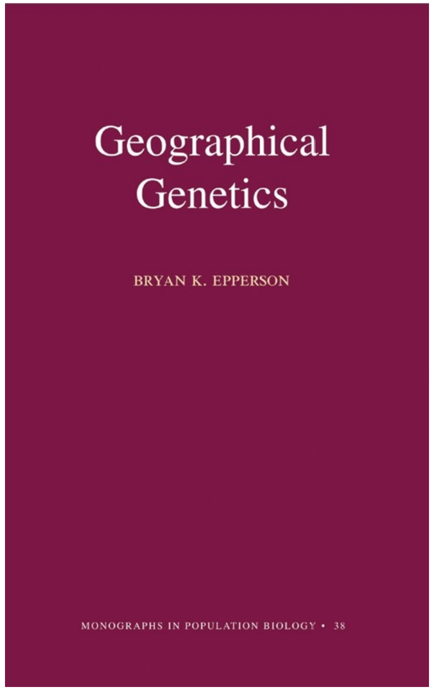

Manfred M. Fischer · Peter Nijkamp
Editors
Geographic Information Systems, Spatial Modelling, and Policy Evaluation
Springer-Verlag

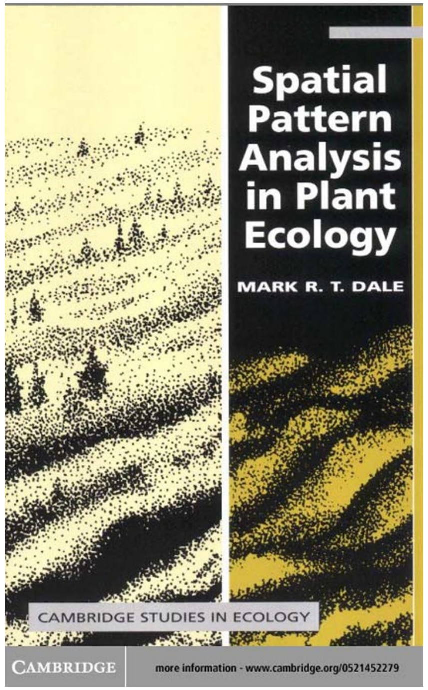

Quantitative Geography
Perspectives on Spatial Data Analysis
A Stewart Fotheringham,
Chris Brunsdon & Martin Charlton

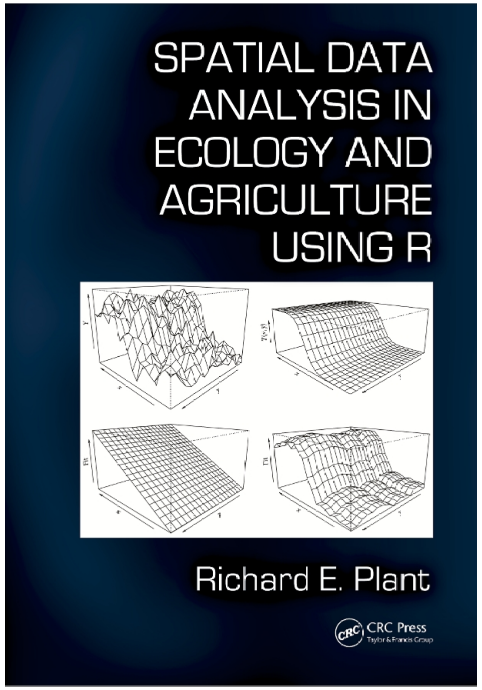

QUANTITATIVE GEOGRAPHY
THE BASICS
RICHARD HARRIS

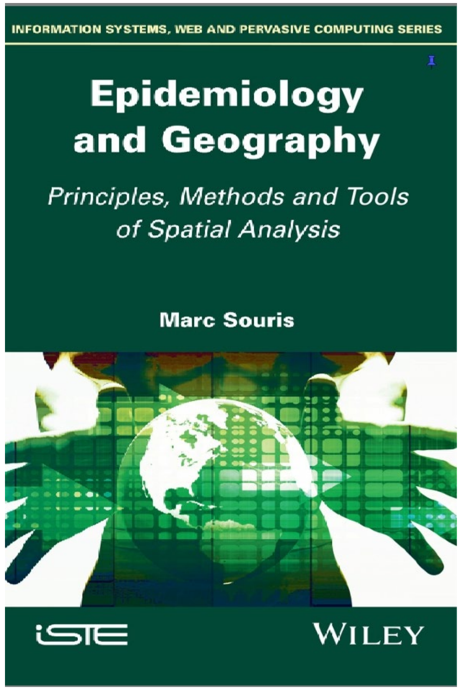
Quick look 2:

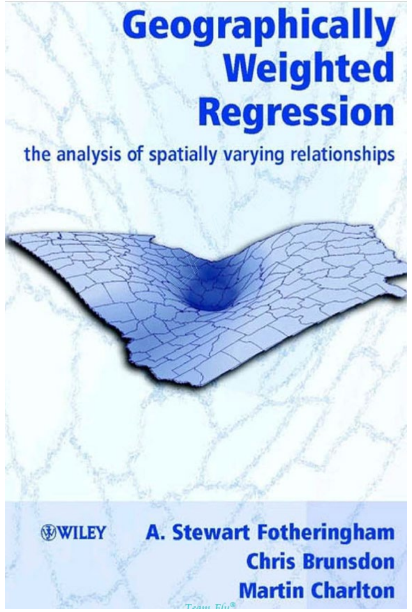

Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS, and Stan
Fränzi Korner-Nievergelt, Tobias Roth, Stefanie von Felten, Jérôme Guélat, Bettina Almasi, Pius Korner-Nievergelt
AP

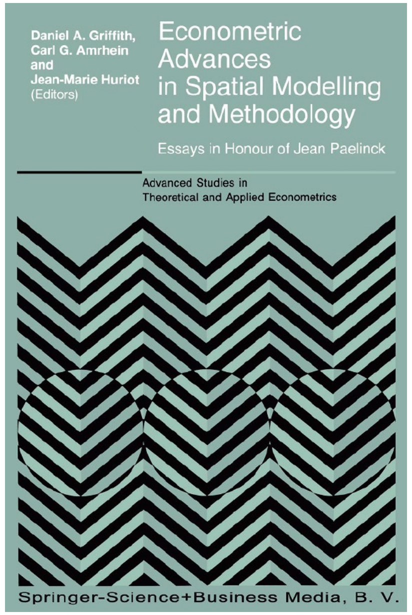

Geographical Genetics
BRYAN K. EPPERSON
MONOGRAPHS IN POPULATION BIOLOGY • 38

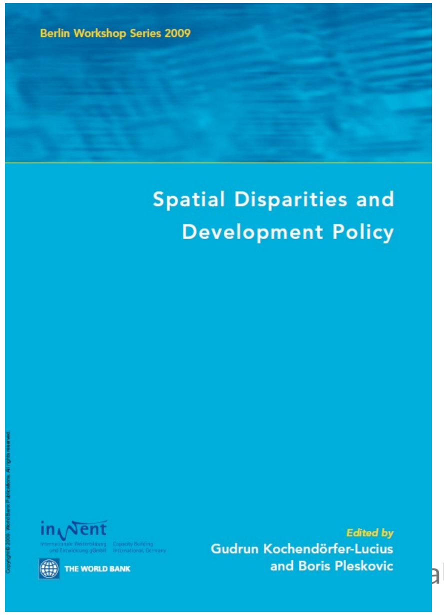

Spatial Pattern Analysis in Plant Ecology
MARK R. T. DALE
CAMBRIDGE STUDIES IN ECOLOGY
CAMBRIDGE
more information - www.cambridge.org/0521452279

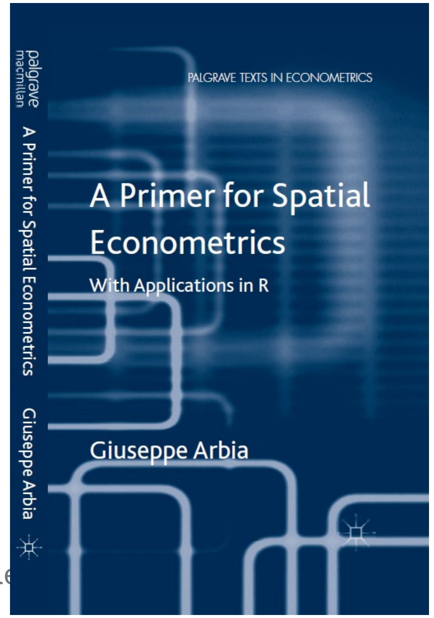

SPATIAL DATA ANALYSIS IN ECOLOGY AND AGRICULTURE USING R
Richard E. Plant
CRC Press

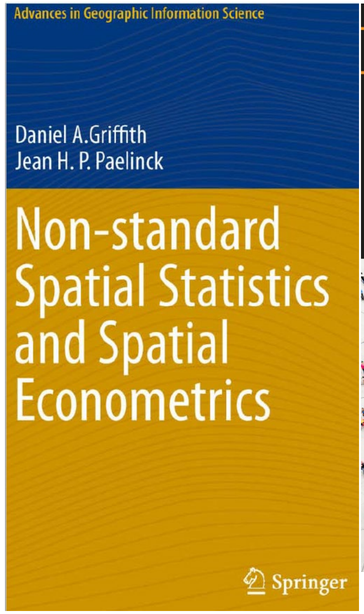

INFORMATION SYSTEMS, WEB AND PERVASIVE COMPUTING SERIES
Epidemiology and Geography
Principles, Methods and Tools of Spatial Analysis
Marc Souris
ISTE
WILEY

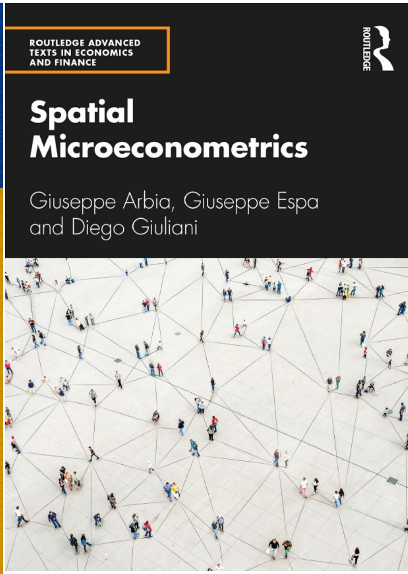

Geographically Weighted Regression
the analysis of spatially varying relationships
WILEY
A. Stewart Fotheringham
Chris Brunsdon
Martin Charlton


Daniel A. Griffith, Carl G. Amrhein and Jean-Marie Huriot (Editors)
Econometric Advances in Spatial Modelling and Methodology
Essays in Honour of Jean Paelinck
Advanced Studies in Theoretical and Applied Econometrics
Springer-Science+Business Media, B. V.


Berlin Workshop Series 2009
Spatial Disparities and Development Policy
InWEnt
THE WORLD BANK
Edited by
Gudrun Kochendörfer-Lucius and Boris Pleskovic

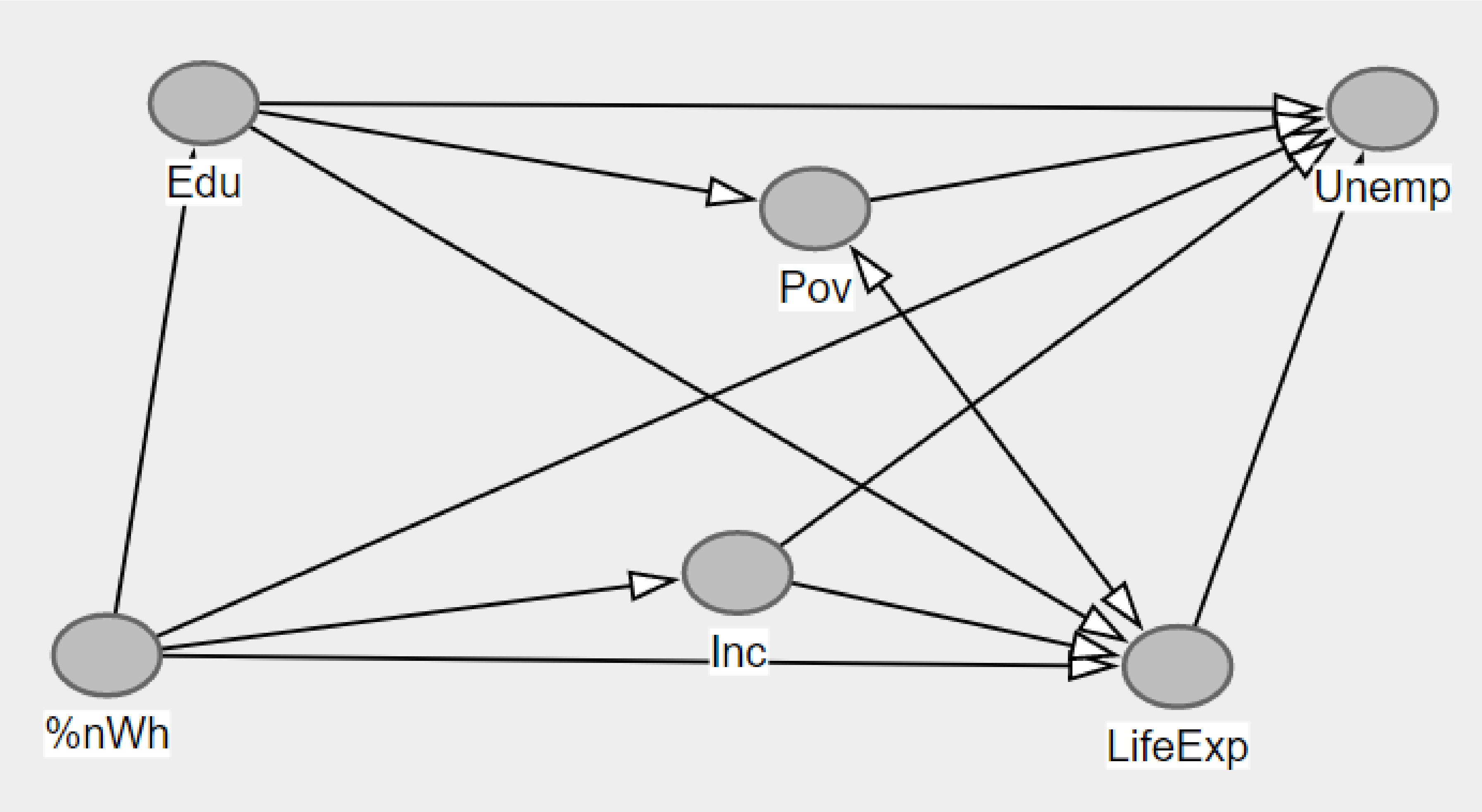

palgrave macmillan
A Primer for Spatial Econometrics
Giuseppe Arbia
PALGRAVE TEXTS IN ECONOMETRICS
A Primer for Spatial Econometrics
With Applications in R
Giuseppe Arbia


Advances in Geographic Information Science
Daniel A.Griffith
Jean H. P. Paelinck
Non-standard Spatial Statistics and Spatial Econometrics
Springer


ROUTLEDGE ADVANCED TEXTS IN ECONOMICS AND FINANCE
ROUTLEDGE
Spatial Microeconometrics
Giuseppe Arbia, Giuseppe Espa and Diego Giuliani

Life Expectancy data informed model

http://dagitty.net/dags.html?id=4TETpl



There are many factors to consider, or course http://bit.ly/HD_causal_model , including molecular: “Scientists Discover a Molecular Switch That Controls Life Expectancy”

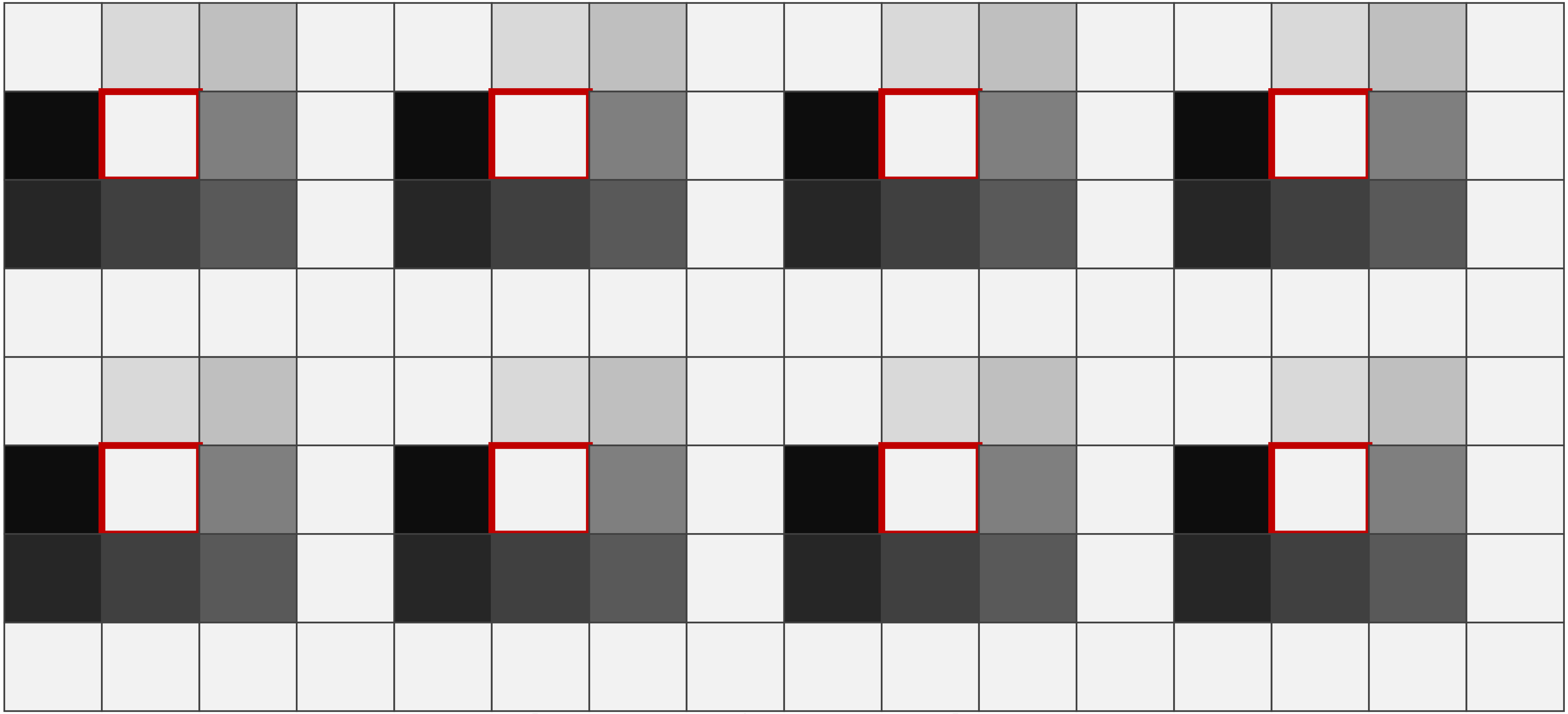
Some troubles with spatial/regional/geographic data

A. Averaging to talk about 'typical region' does not work :
i. A region with 1 resident with 100y LfEx and another with 100 residents with 80ybLfEx do not yield a 101 aggregate with 90y LfEx.
ii. If a region's LfEx value is *identical* to its neighbors', then this is too much similarity: much like spousal data, or family data.
B. Clustering within higher level regions due to all-belong-to-higher structure is distinct from clustering due to each-to-its-neighbors spatial structure: there are as many clusters as regions!

***** Multilevel modeling does not address the spatial structure, much like it can't address e.g. friendship relational structure in student-in-classrooms settings.

Intuition for minimum Moran's I

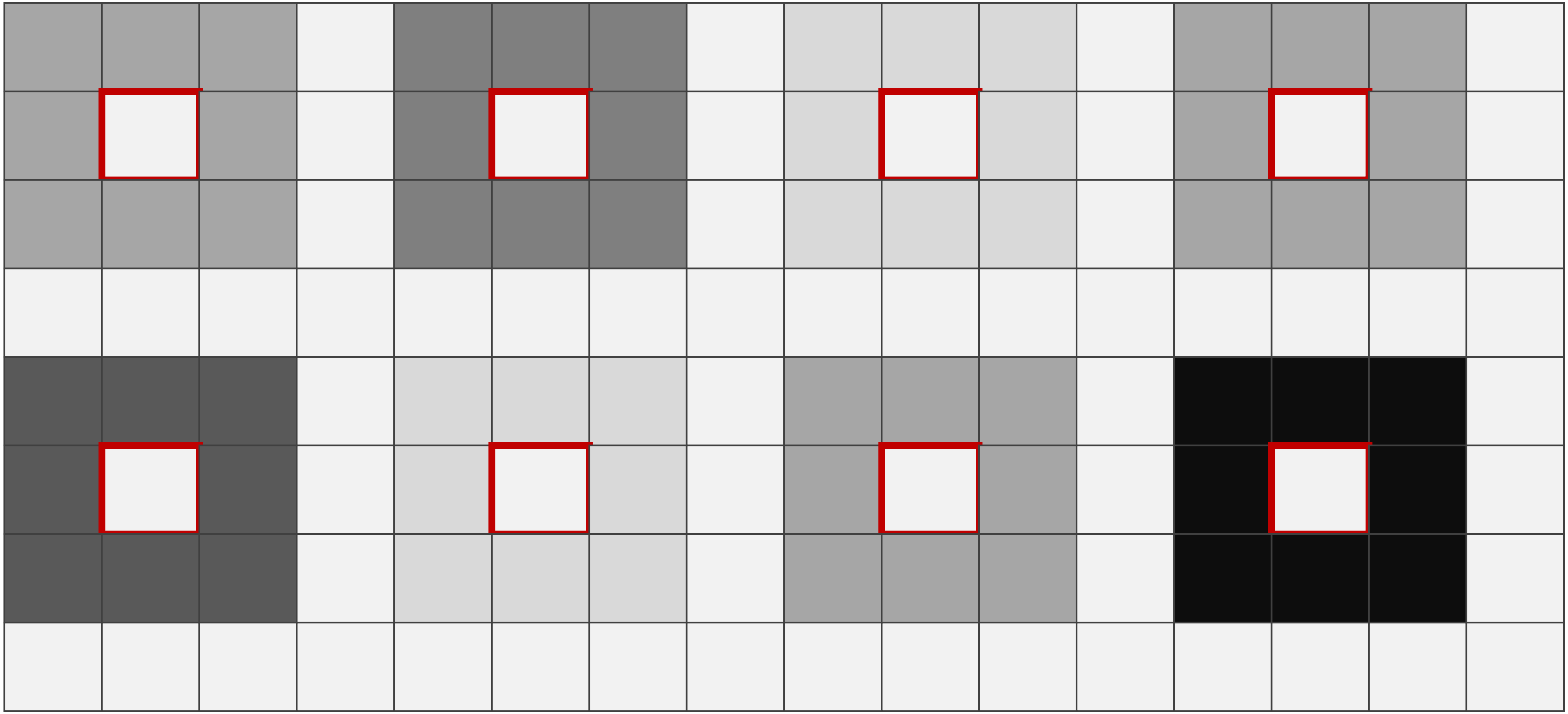
Haggard, E. A. (1958). Intraclass correlation and the analysis of variance



"all the variation is within classes [neighbors of red squares], with the result that there is no variation between class (i.e., each class sum equals [the same #])."

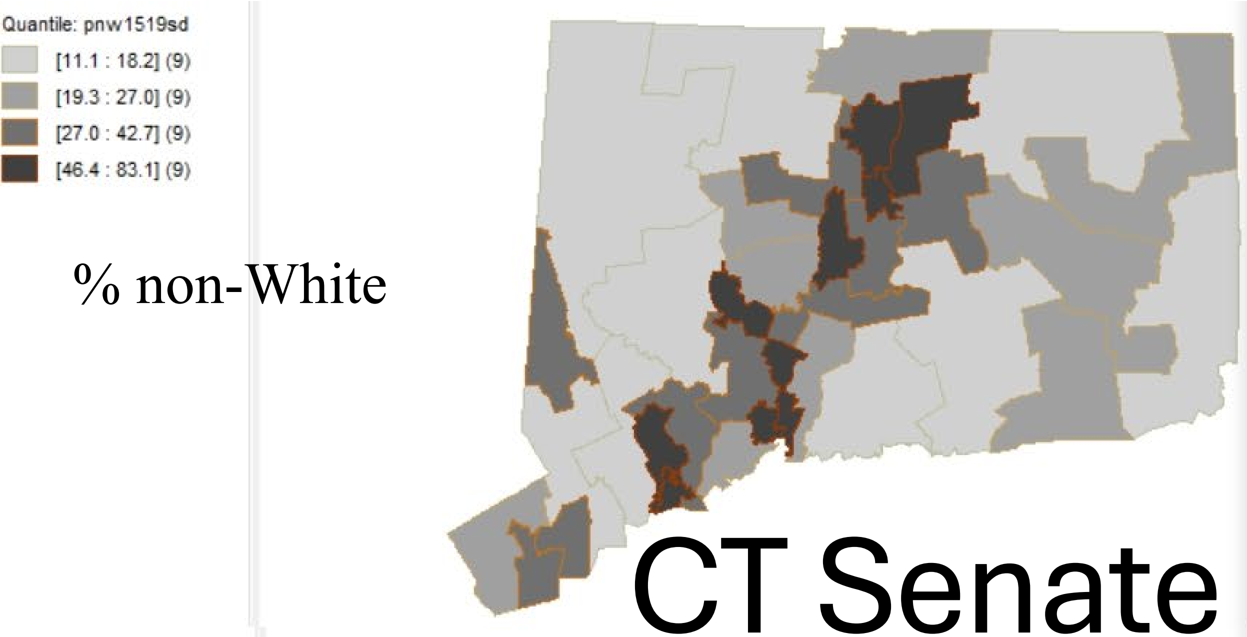
Intuition for Maximum Moran's I

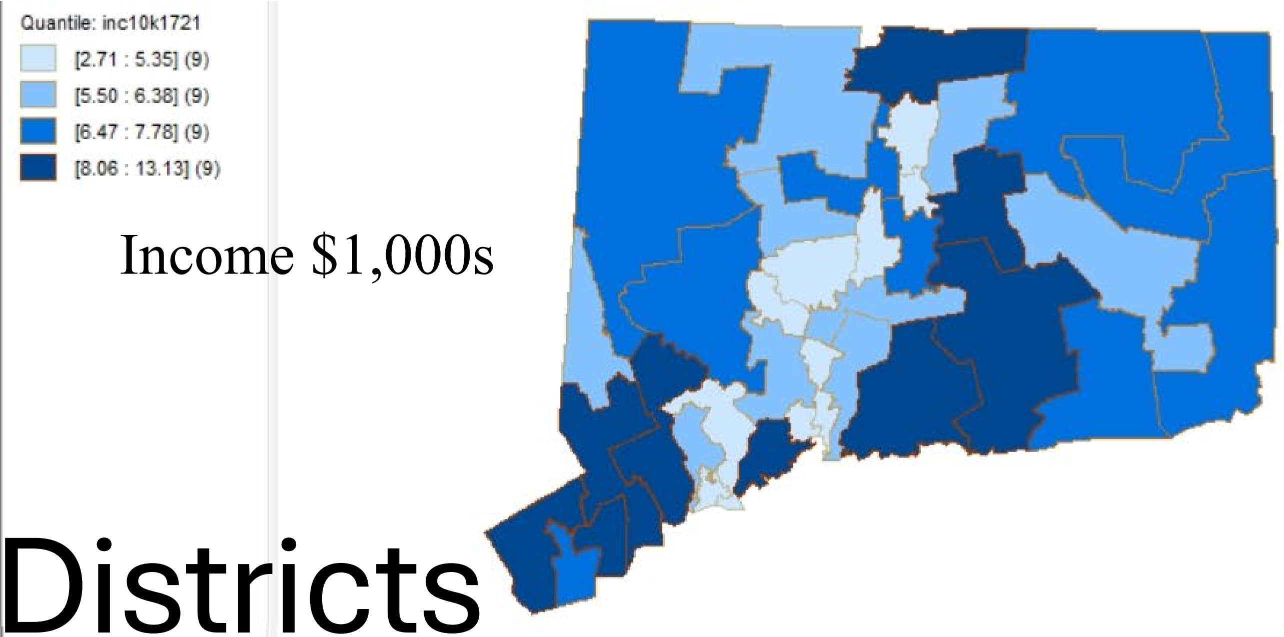
Haggard, E. A. (1958). [Intraclass correlation and the analysis of variance]

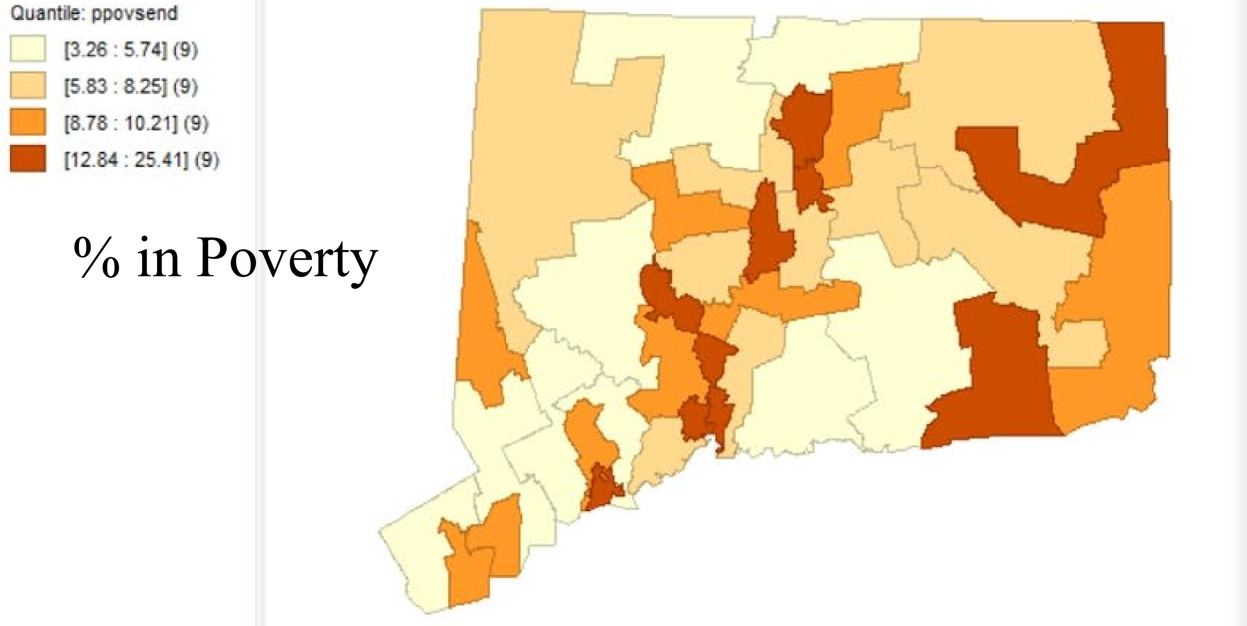


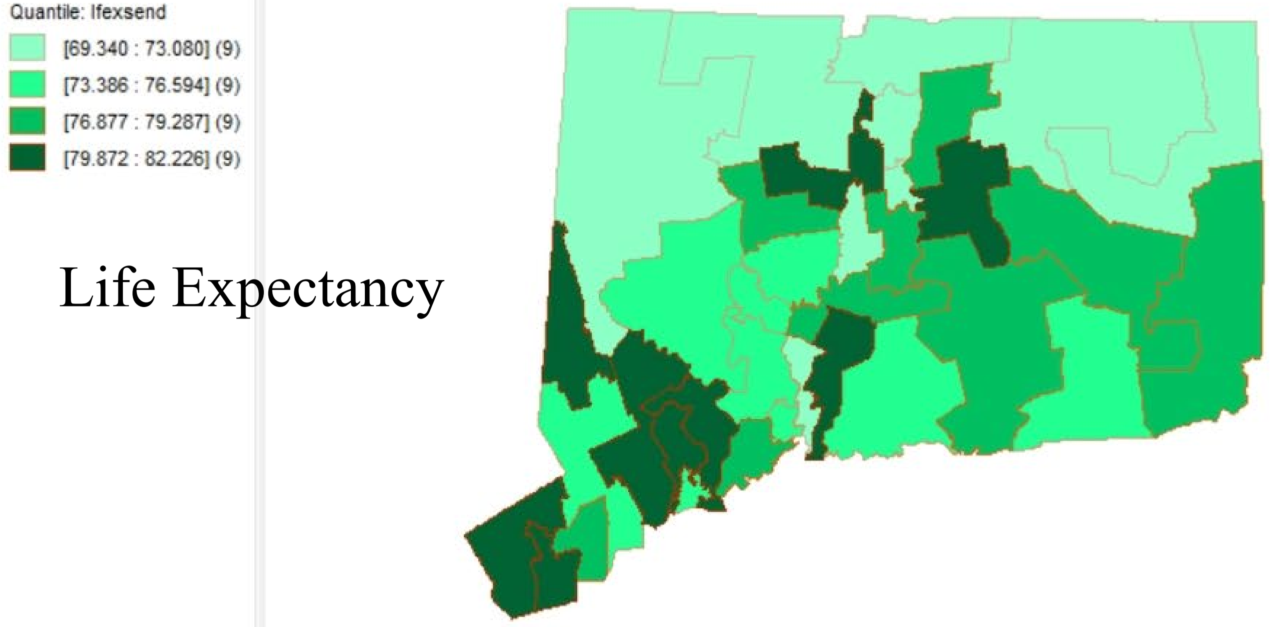
"there is no variation between the scores in any of the [classes [neighbors of red squares]; rather all the variation is between the [classes [the same #])."

CT Senate Districts





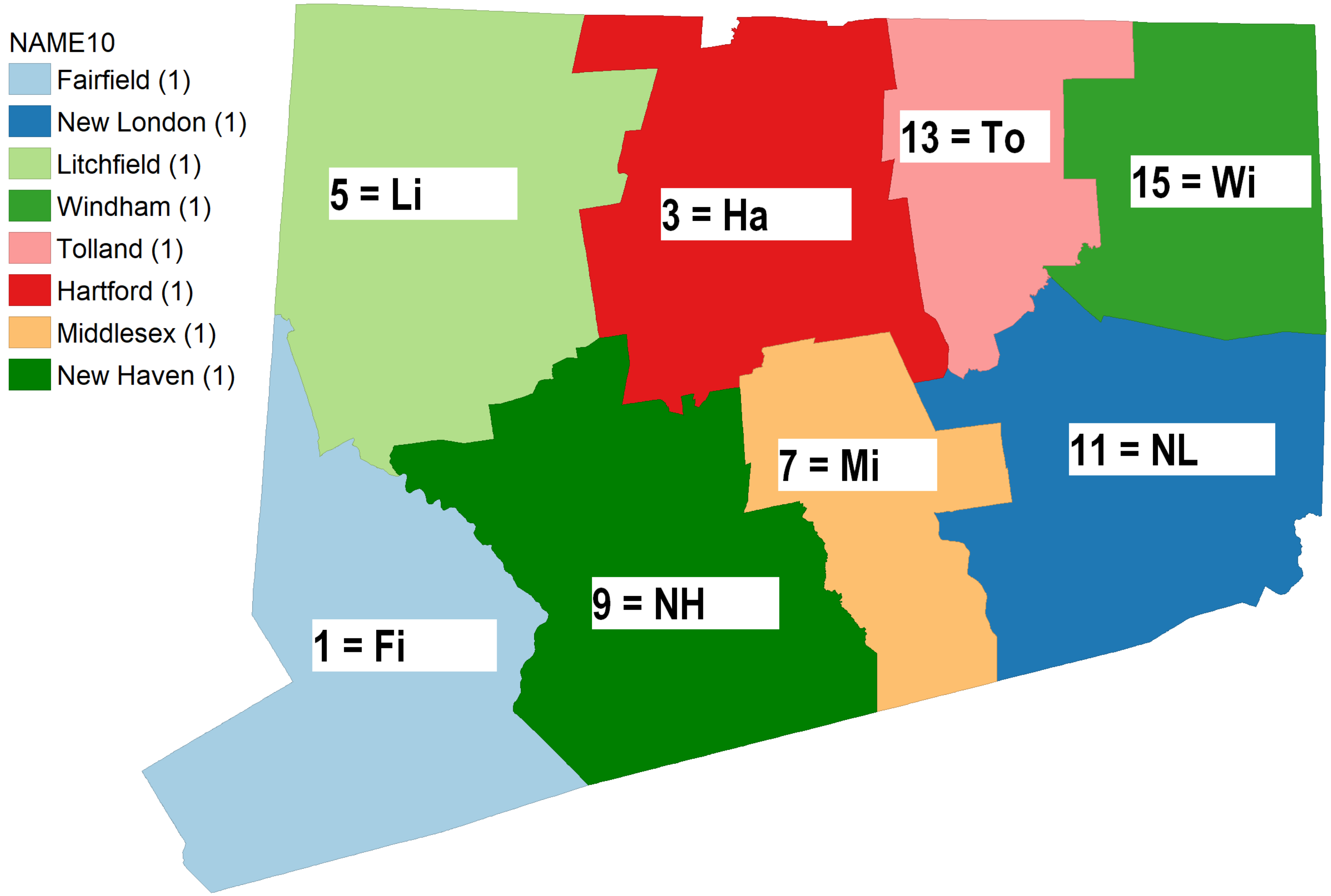




Queen Contiguity Weight Matrix – CT 8 counties

GEOID10

1 2
5 9
3 5
11 5 13 7
5 3
1 3 9
7 3
11 3 9
9 4
3 5 1 7
11 4
15 13 3 7
13 3
11 15 3
15 2
11 13



Queen Contiguity Weight Matrix – CT 8 counties

```
GEOID10
  1 2
  5    9
  3 5
  11    5    13    7    9
  5 3
  1    3    9
  7 3
  11    3    9
  9 4
  3    5    1    7
  11 4
  15    13    3    7
  13 3
  11    15    3
  15 2
  11    13
```

STANDARDIZE weights:

UCONN HEALTH
HEALTH DISPARITIES INSTITUTE

W_{ij}	9001	9003	9005	9007	9009	9011	9013	9015	
	Fairfield	Hartford	Litchfield	Middlesex	New Haven	New London	Tolland	Windham	Neighbors
Fairfield			0.50		0.50				2
Hartford			0.20	0.20	0.20	0.20	0.20		5
Litchfield	0.33	0.33			0.33				3
Middlesex		0.33			0.33	0.33			3
New Haven	0.25	0.25	0.25	0.25					4
New London		0.25		0.25			0.25	0.25	4
Tolland		0.33				0.33		0.33	3
Windham						0.50	0.50		2

15

Spelling out the ‘auto’-correlation – CT counties

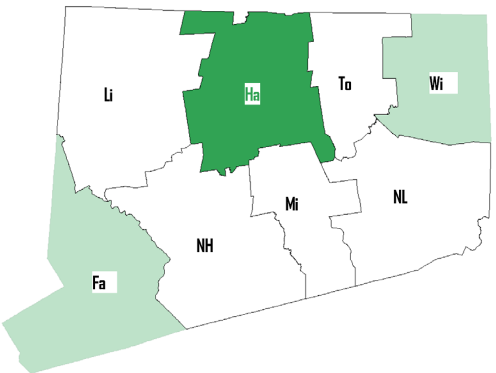
“In essence, it is a cross-product statistic between a variable and its spatial lag, with the variable expressed in deviations from its mean.” GeoDa

$$I_Y = \sum_i \sum_j [(W_{ij} \cdot (y_i - \bar{Y}) \cdot (y_j - \bar{Y})] / \ S_0 \ ] / \ [\sum_i (y_i - \bar{Y})^2 / n]$$

with w_{ij} as the elements of the spatial weights matrix, $S_0 = \sum_i \sum_j w_{ij}$ as the sum of all the weights, and n as the number of observations. For the 8 CT counties, one then would get

$({}^{Ha}[1/5 \cdot (y_{Ha} - \bar{Y}) \cdot (y_{Li} - \bar{Y})] + 1/5 \cdot (y_{Ha} - \bar{Y}) \cdot (y_{Mi} - \bar{Y})] + 1/5 \cdot (y_{Ha} - \bar{Y}) \cdot (y_{NH} - \bar{Y})] + 1/5 \cdot (y_{Ha} - \bar{Y}) \cdot (y_{NL} - \bar{Y})] +$
$1/5 \cdot (y_{Ha} - \bar{Y}) \cdot (y_{To} - \bar{Y})] +$
$[{}^{Li}\ldots] + [{}^{Fa}\ldots] + [{}^{NH}\ldots] + [{}^{Mi}\ldots] + [{}^{To}\ldots] + [{}^{NL}\ldots] + [{}^{Wi}\ldots] +) / 8)$ /
$([(y_{Fa} - \bar{Y})^2 + (y_{Ha} - \bar{Y})^2 + \ldots + (y_{Wi} - \bar{Y})^2] / 8)$

(if we use the standardized weights, to sum up to 1 per case)



The ‘clustering’/spatial structure is contained in the Weight Matrix: how the ‘clusters’ are built:

- Each case/region has its own ‘cluster’!
- ‘Clusters’ overlap: same regions can belong to > 1 ‘cluster’!
- There is cyclical influences between ‘members’:

Spelling out the spatial regression – CT counties

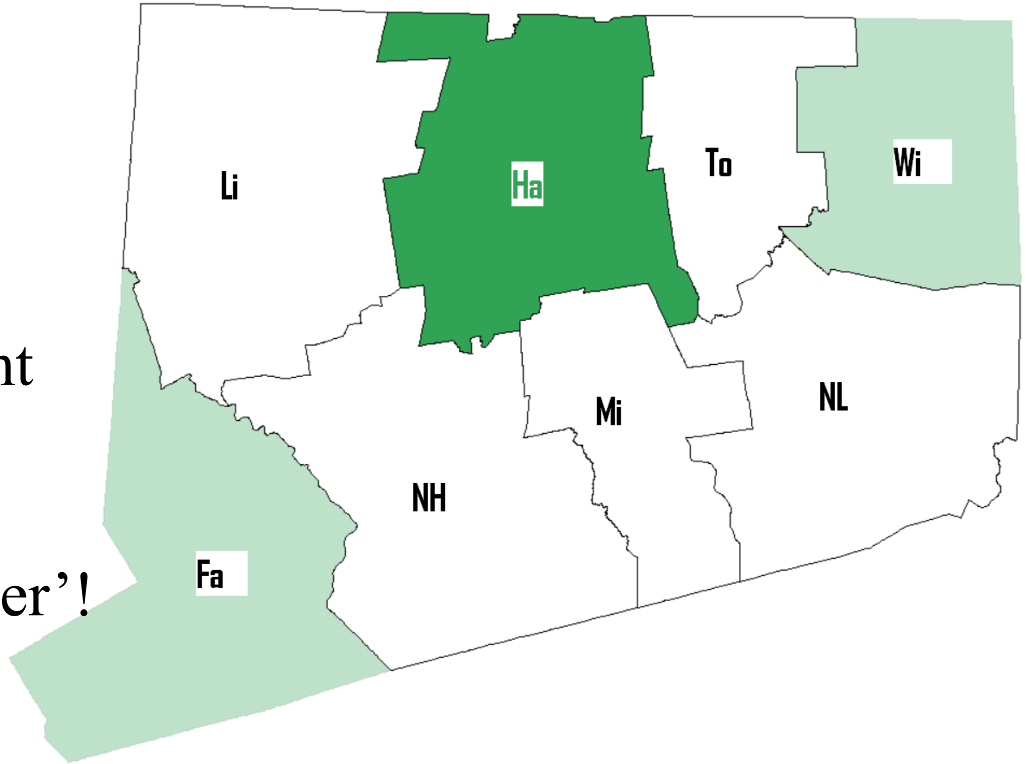
A classic regression $Y_i = \alpha. + \beta.\cdot X_i + \varepsilon_i$ would become for spatially connected/nonindependent data e.g., from

$Y_{Ha} = \alpha. + \beta.\cdot X_{Ha} + \varepsilon_{Ha}$, etc. to:

$Y_{Ha} = \rho\cdot(1/5\cdot Y_{Li} + 1/5\cdot Y_{NH} + 1/5\cdot Y_{Mi} + 1/5\cdot Y_{NL} + 1/5\cdot Y_{To}) + \alpha. + \beta.\cdot X_{Ha} + \varepsilon_{Ha}$

which says that Ha has 5 ‘queen’ neighbors,

$Y_{To} = \rho\cdot(1/4\cdot Y_{Ha} + 1/4\cdot Y_{Mi} + 1/4\cdot Y_{NL} + 1/4\cdot Y_{Wi}) + \alpha. + \beta.\cdot X_{To} + \varepsilon_{To}$, which says that To has 4 ‘queen’ neighbors, etc.



The ‘clustering’/spatial structure is contained in the Weight Matrix: how the ‘clusters’ are built:

- Each case/region IS its own ‘cluster’!
- ‘Clusters’ overlap: same regions can belong to > 1 ‘cluster’!
- There is **cyclical influences** between ‘members’:

‘Contagion’/interference & causal reasoning

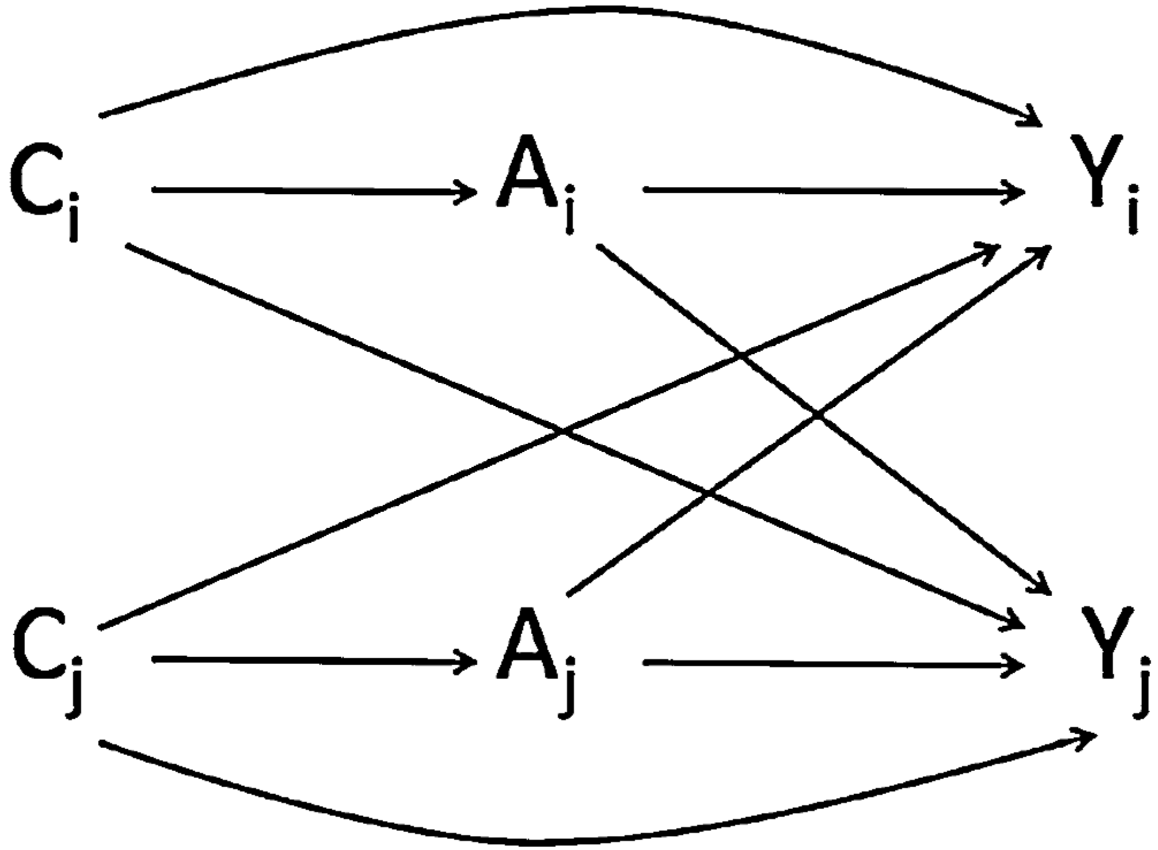
* i and j are 2 ‘individuals’: regions here.

* They ‘affect’ each other (‘contagion’): a different type of causal confounding at work.

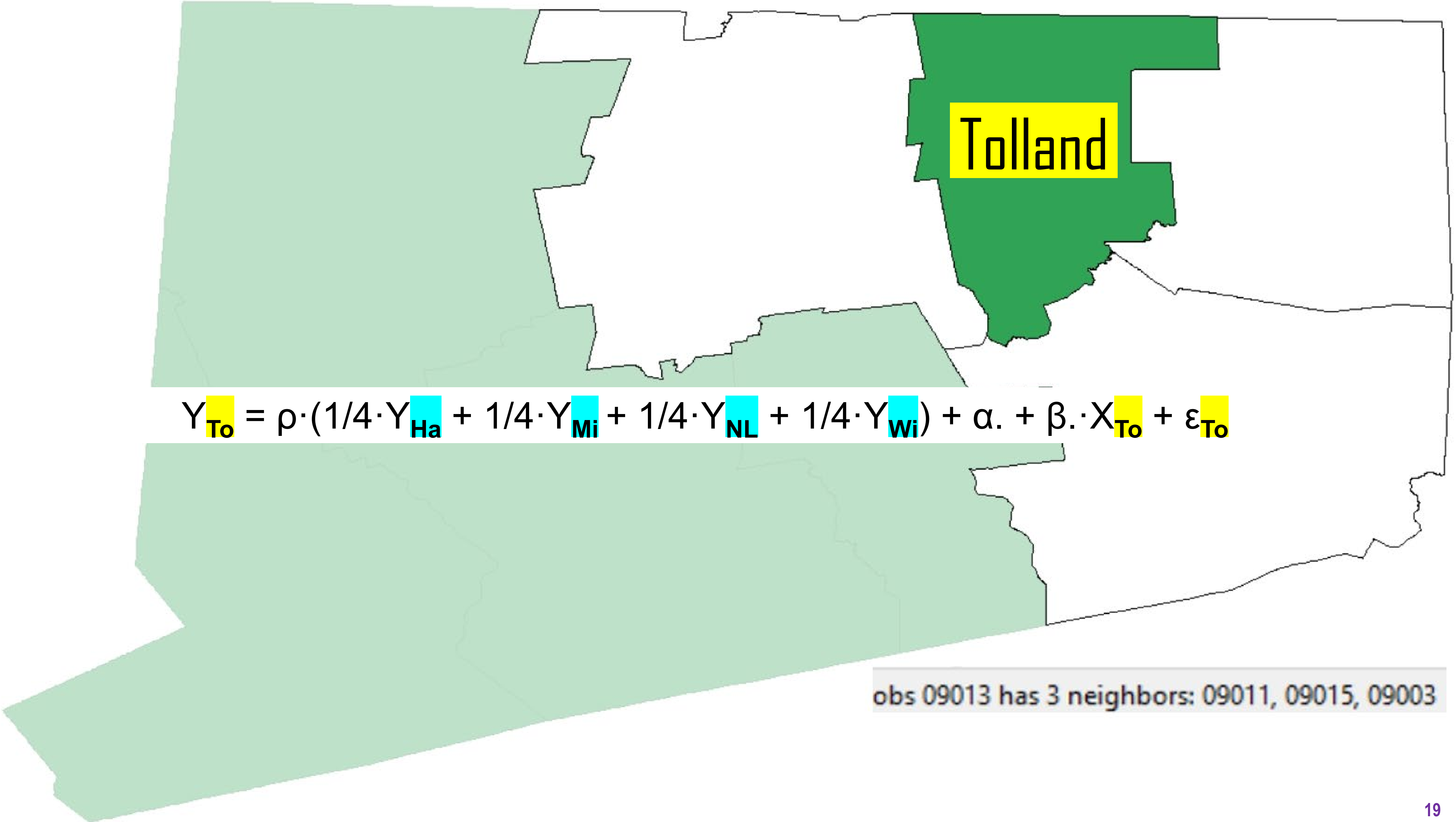
*** This truly turns patient /clinical/medical health research into public health research.

* The spatial structure adds to this individual DAG (direct acyclic graph) reasoning!

Fig. 5.a

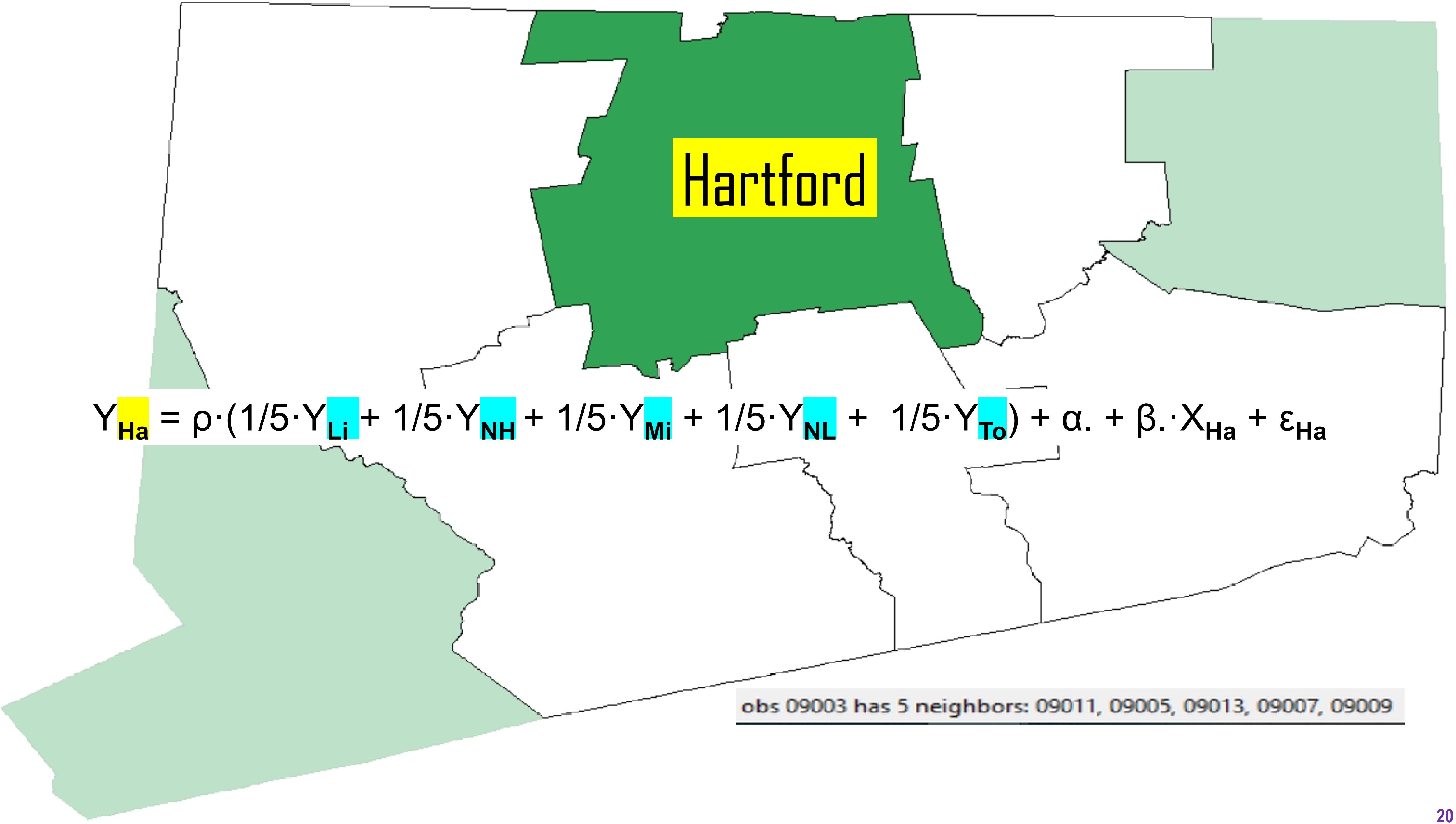


Ogburn, E. L., & VanderWeele, T. J. (2014). Causal Diagrams for Interference. Statistical Science, 29(4), 559-578.



$$Y_{To} = \rho\cdot(1/4\cdot Y_{Ha} + 1/4\cdot Y_{Mi} + 1/4\cdot Y_{NL} + 1/4\cdot Y_{Wi}) + \alpha. + \beta.\cdot X_{To} + \varepsilon_{To}$$

obs 09013 has 3 neighbors: 09011, 09015, 09003



$$Y_{Ha} = \rho \cdot (1/5 \cdot Y_{Li} + 1/5 \cdot Y_{NH} + 1/5 \cdot Y_{Mi} + 1/5 \cdot Y_{NL} + 1/5 \cdot Y_{To}) + \alpha. + \beta. \cdot X_{Ha} + \varepsilon_{Ha}$$

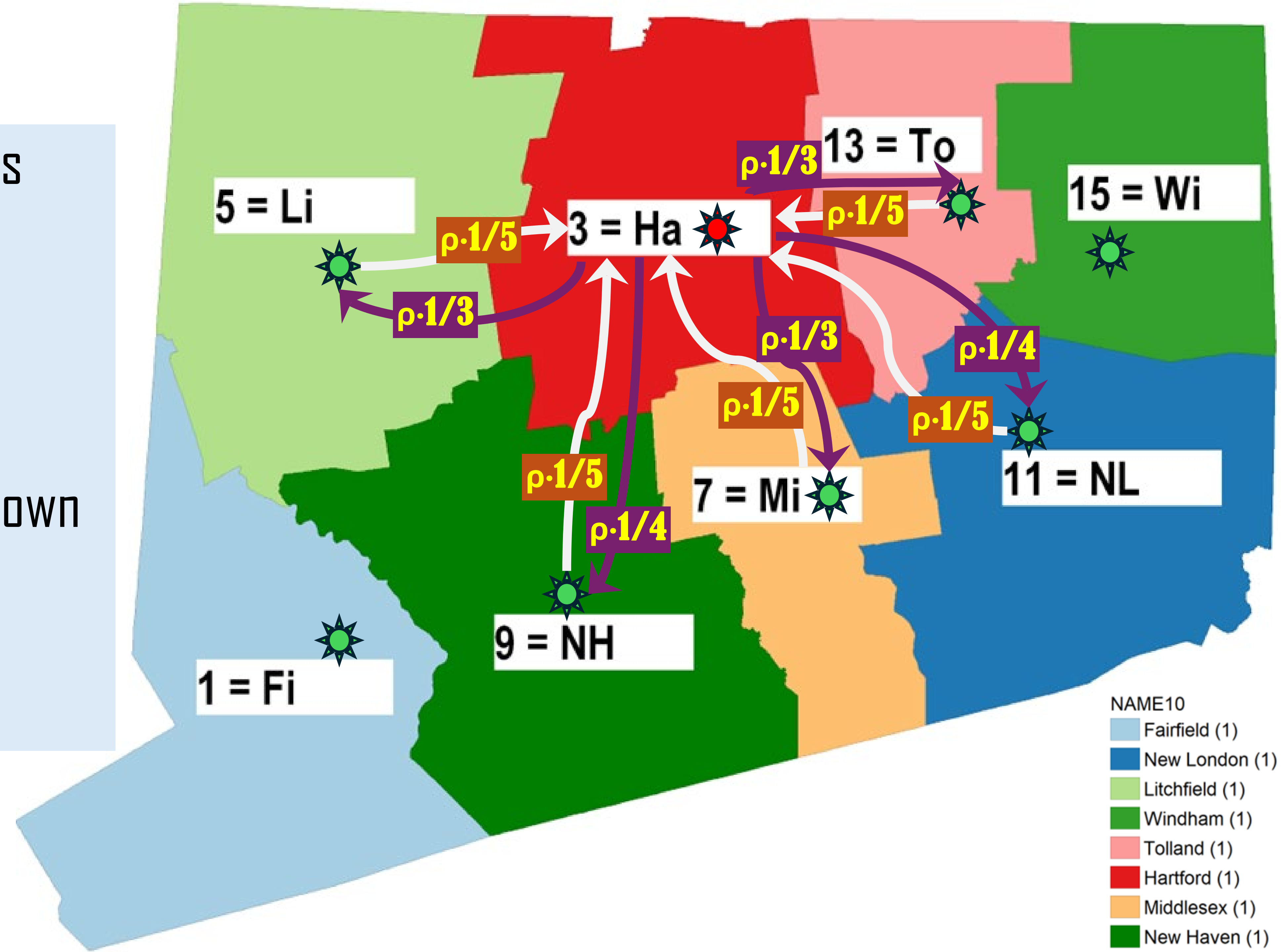
obs 09003 has 5 neighbors: 09011, 09005, 09013, 09007, 09009

Hartford's LifeExp. is pushed down by its neighbors

ρ = -0.93

t=-2.522

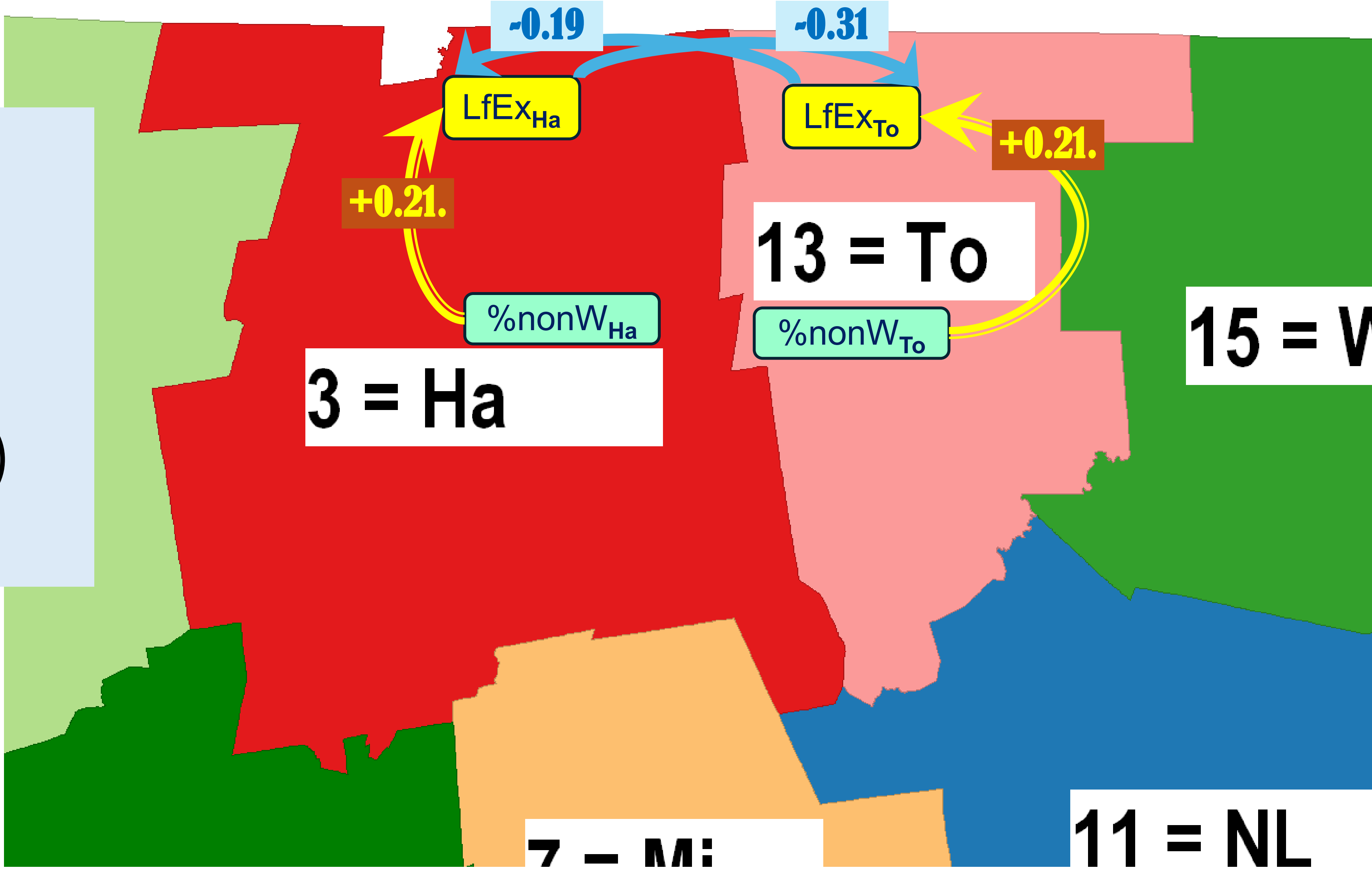
But it also pushed down its neighbors, with varying strengths, however.



Spatial 'auto'-correlation & causal reasoning

Ha: -0.19= -0.93·(1/5)

To: -0.31= -0.93·(1/3)



Counties 'individuals': regions here.

Ogburn, E. L., & VanderWeele, T. J. (2014). Causal Diagrams for Interference. Statistical Science, 29(4), 559-578.

Intermediary linguistic clarification

** A correlation is a same-case (region, person) & across/between-variables statistics:] 'mutual similarity' in two sets of numbers: knowing one region's X tells us something about that region's Y

** 'Auto'-correlation is on the other hand across/between-persons & same-variable statistics: knowing a region's neighboring regions' Xs tells us something about its own X

X	Y	
% non-White	*Life Exp.*	County
11.0	80.3	Litchfield
14.2	81.1	Tolland
14.9	79.2	Windham
15.0	81.3	Middlesex
22.7	79.8	New London
32.5	79.8	New Haven
34.7	82.0	Fairfield
35.2	80.0	Hartford
22.5	80.4	**Means**

X	Y	
% non-White	*Life Exp.*	County
14.9	79.2	Windham
22.7	79.8	New London
32.5	79.8	New Haven
35.2	80.0	Hartford
11.0	80.3	Litchfield
14.2	81.1	Tolland
15.0	81.3	Middlesex
34.7	82.0	Fairfield
22.5	80.4	**Means**

Naïve correlation

Intuition:

When ordering by 1 variable, the other variable's values 'cluster' : all high, or all low.

Upper view kind of supports it: 3 of HIGH LifeExp are in LOW %non_White (so we see a 1+3+3+1 pattern in the binary Lo/Hi crosstabulation.

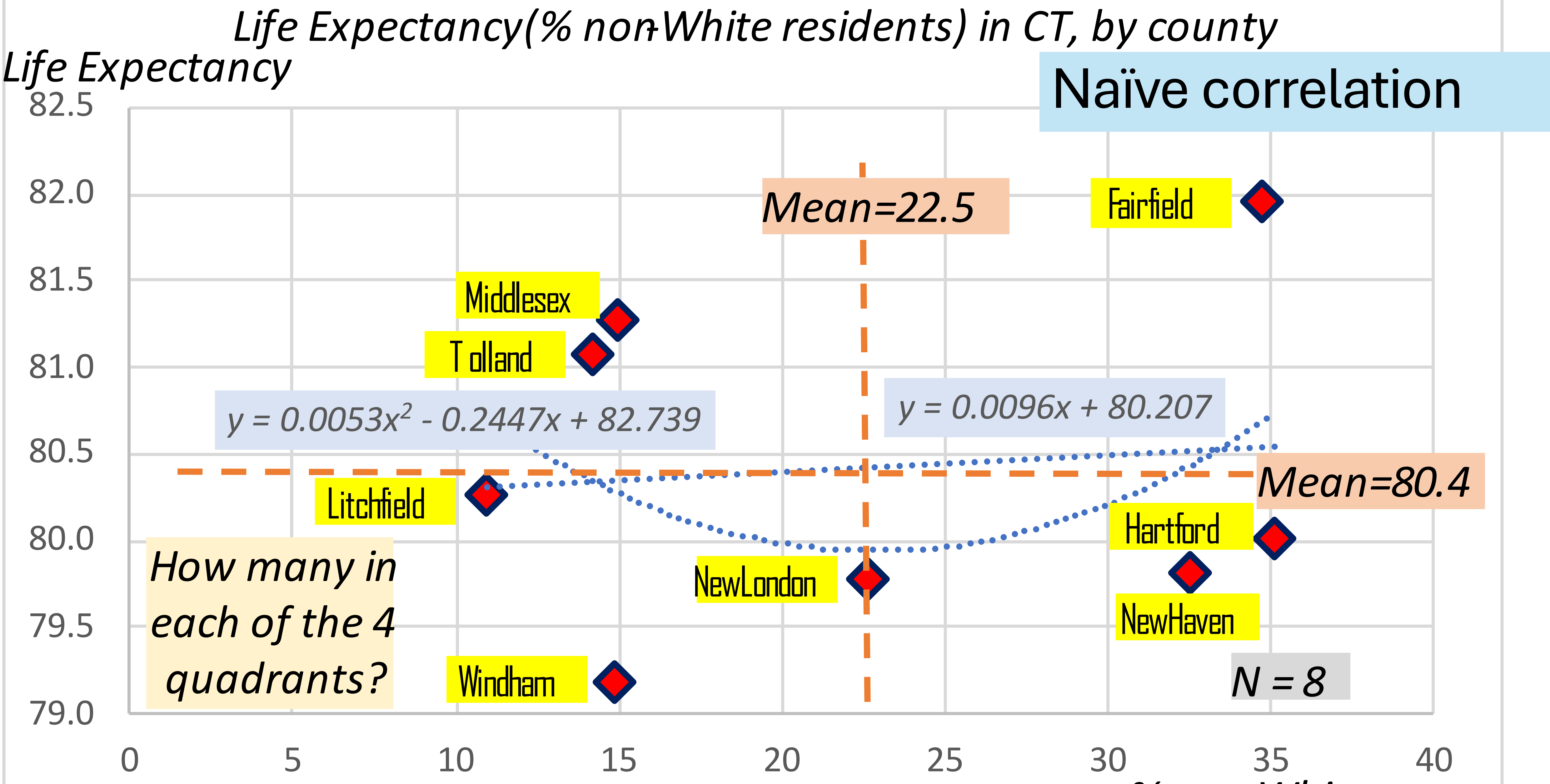
A chi-square test would not find this data pattern statistically significantly different from the null/no relation data pattern (2+2+2+2).

	LifeExp	0	1	
%nWhte	0	1	3	4
	1	3	1	4
		4	4	8

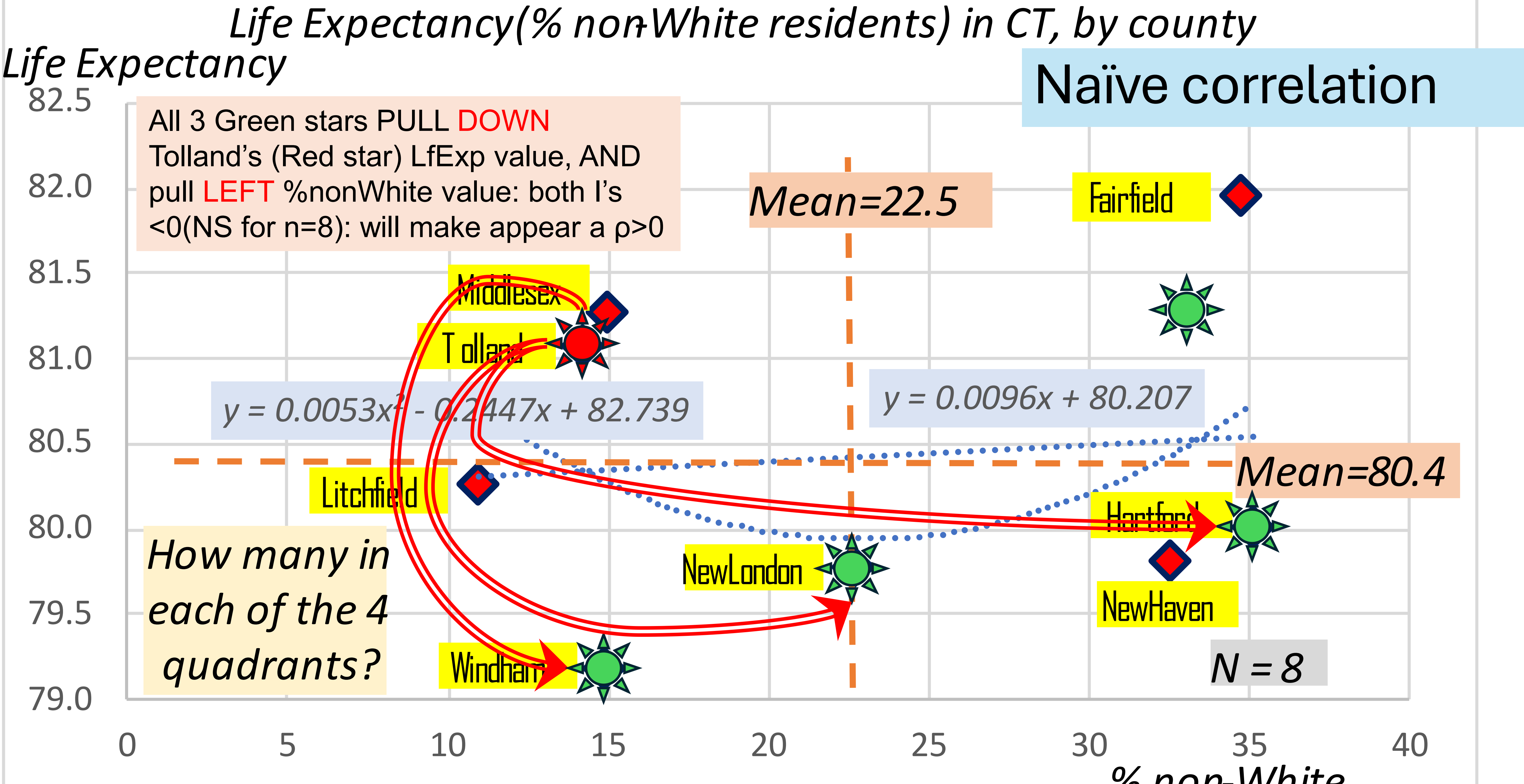
Legend:

LIGHT color = LOW values

DARK color = HIGH values



Source: CDC Life Expectancy data



Source: CDC Life Expectancy data

6/22/24

CT Small Claims for Medical Debt totals

Year	Numbers of claims	Total filed amounts	Mean amount per docket
2015	10,272	$ 15,767,136	$ 1,535
2016	12,056	$19,382,123	$ 1,608
2018	12,097	$20,786,962	$ 1,718
2019	9,185	$16,348,638	$ 1,780

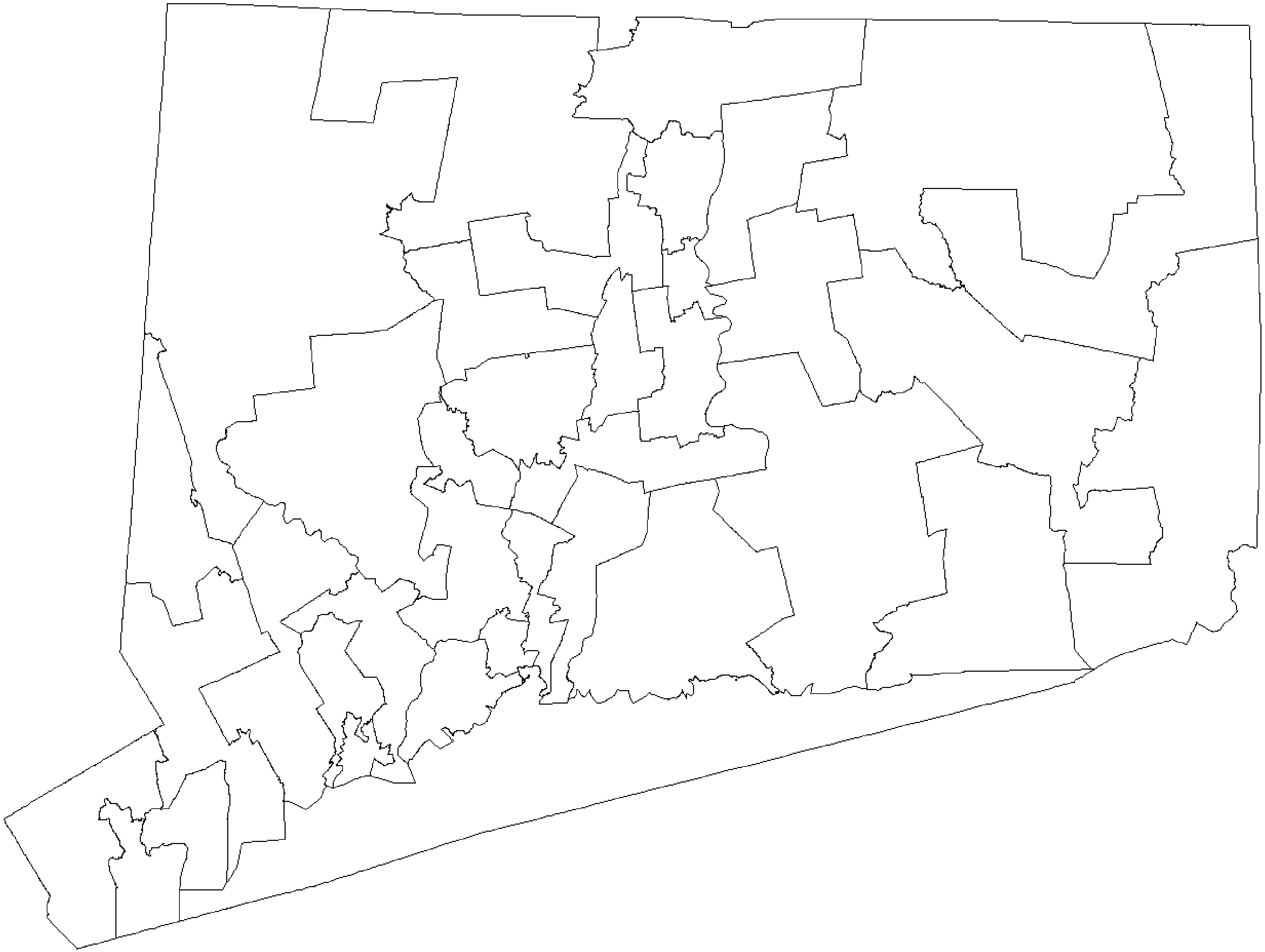
The total number of medical debt small claims in CT, total and average amounts charged per defendant/patient
Notes: The 2017 data did not cover the full year, and is not reported; the claims counts up number of unique 'dockets' or cases filed (multiple family members may appear in the same docket); amounts are shown as 'filed', not as 'awarded' the awarded amounts are 99.2% on the whole from the total amounts filed, in years 1, and 2; years 3 and 4 data did not have amounts awarded.

CT Basic descriptives across 2 geographic/regional layers

	N	Mean
Percent of all people who were nonWhite in 2020$_{CsTr}$	820	33.8
CT Senate district	36	32.5
Average annual out of pocket per person on medical care in 2019$_{CsTr}$	761	\$1,011
CT Senate district	36	\$ 908
CDC SVI Per capita income estimate, 2014-2018 ACS$_{CsTr}$	820	\$42,750
CT Senate district	36	\$42,903
Gini Index inequality of household income 2016-2020$_{CsTr}$	761	0.427
CT Senate district	36	0.379
Rate of medical debt (per 10,000 residents) in 2019$_{CsTr}$	759	28.18
CT Senate district	36	25.75

CsTr = Census Tracts

CT State Senate Districts

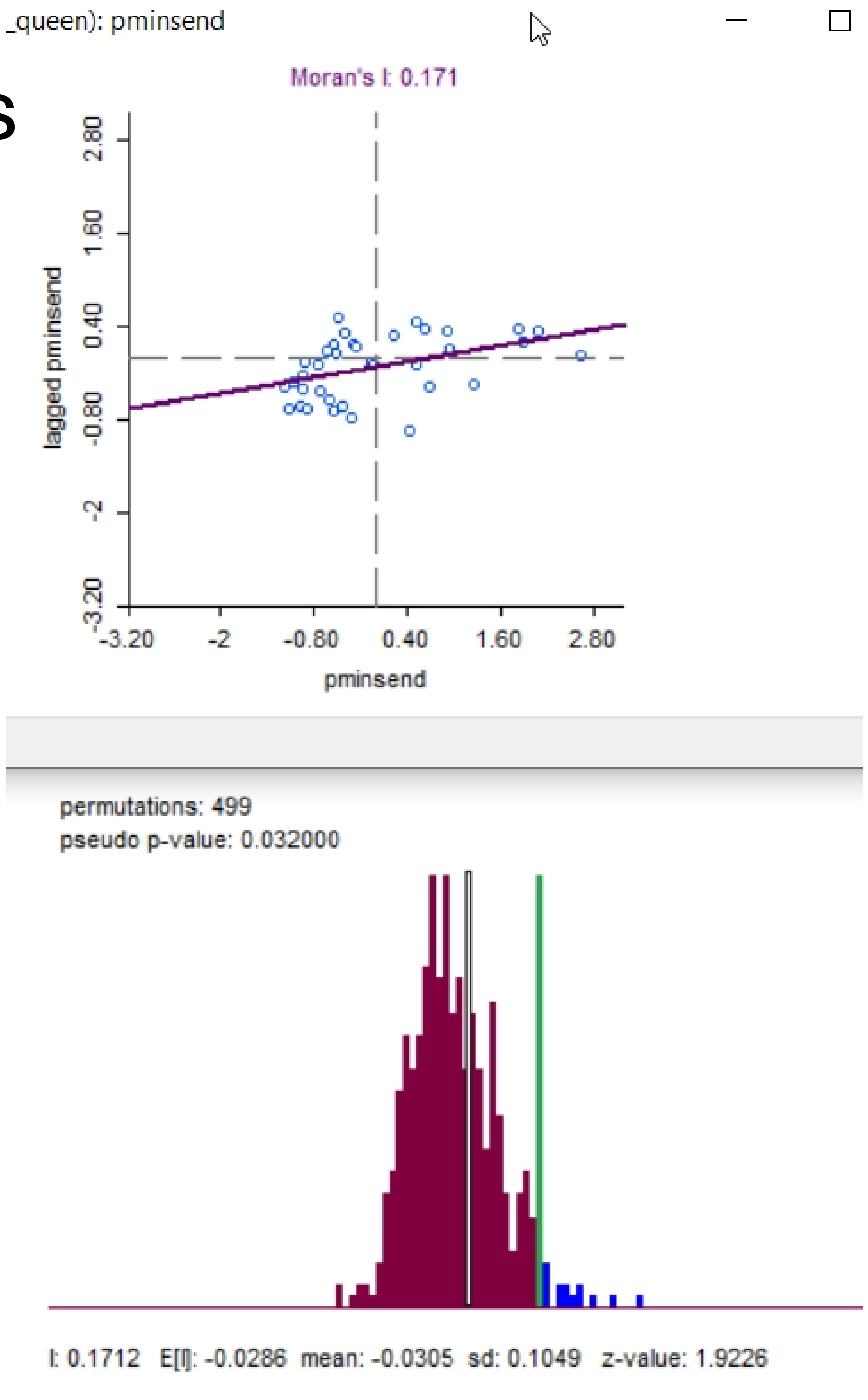


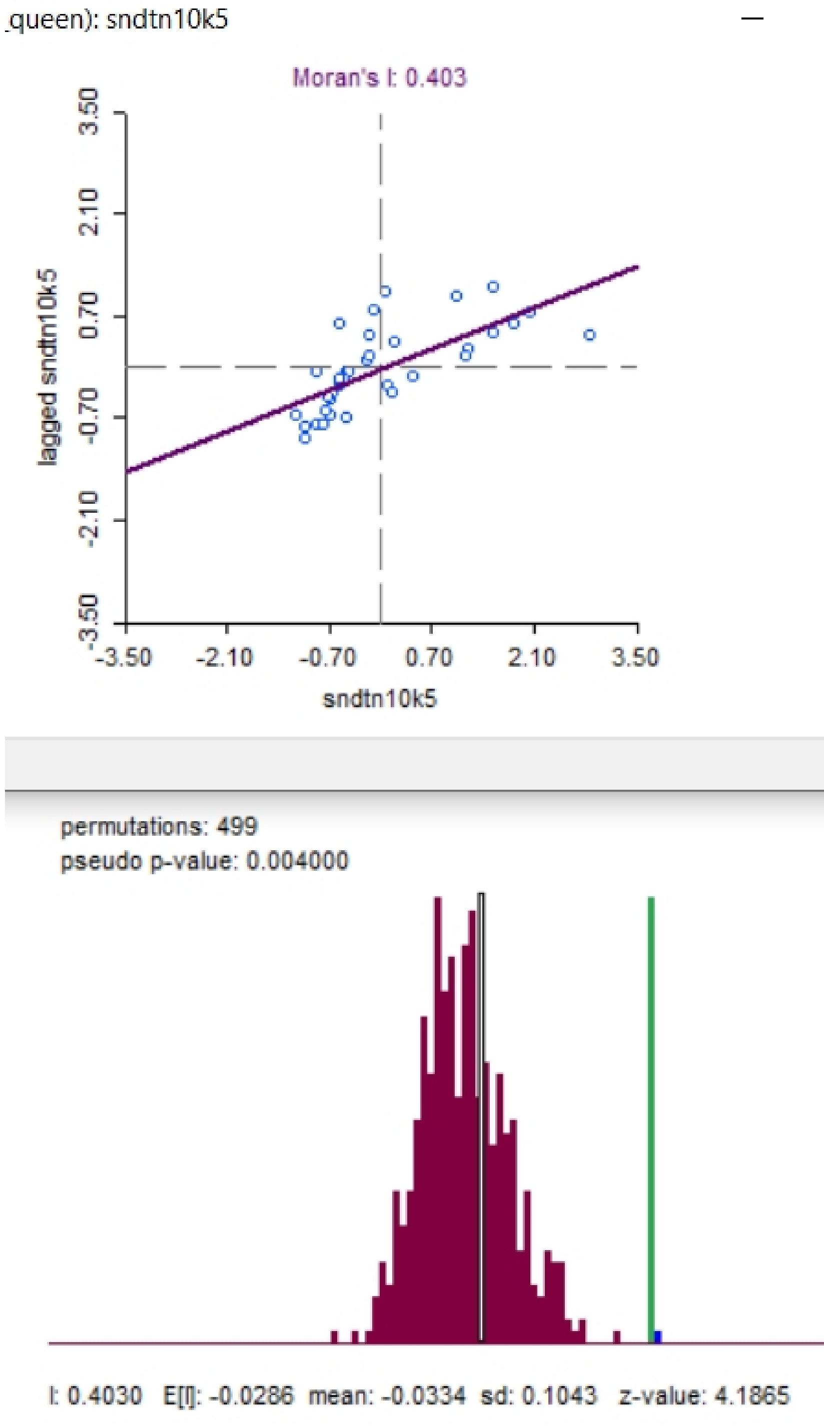
State Senate Districts ‘auto’-correlations

$I_{DebtRate} = .403, z = 4.186$

$I_{\%non\text{-}White} = .171, z = 1.923$

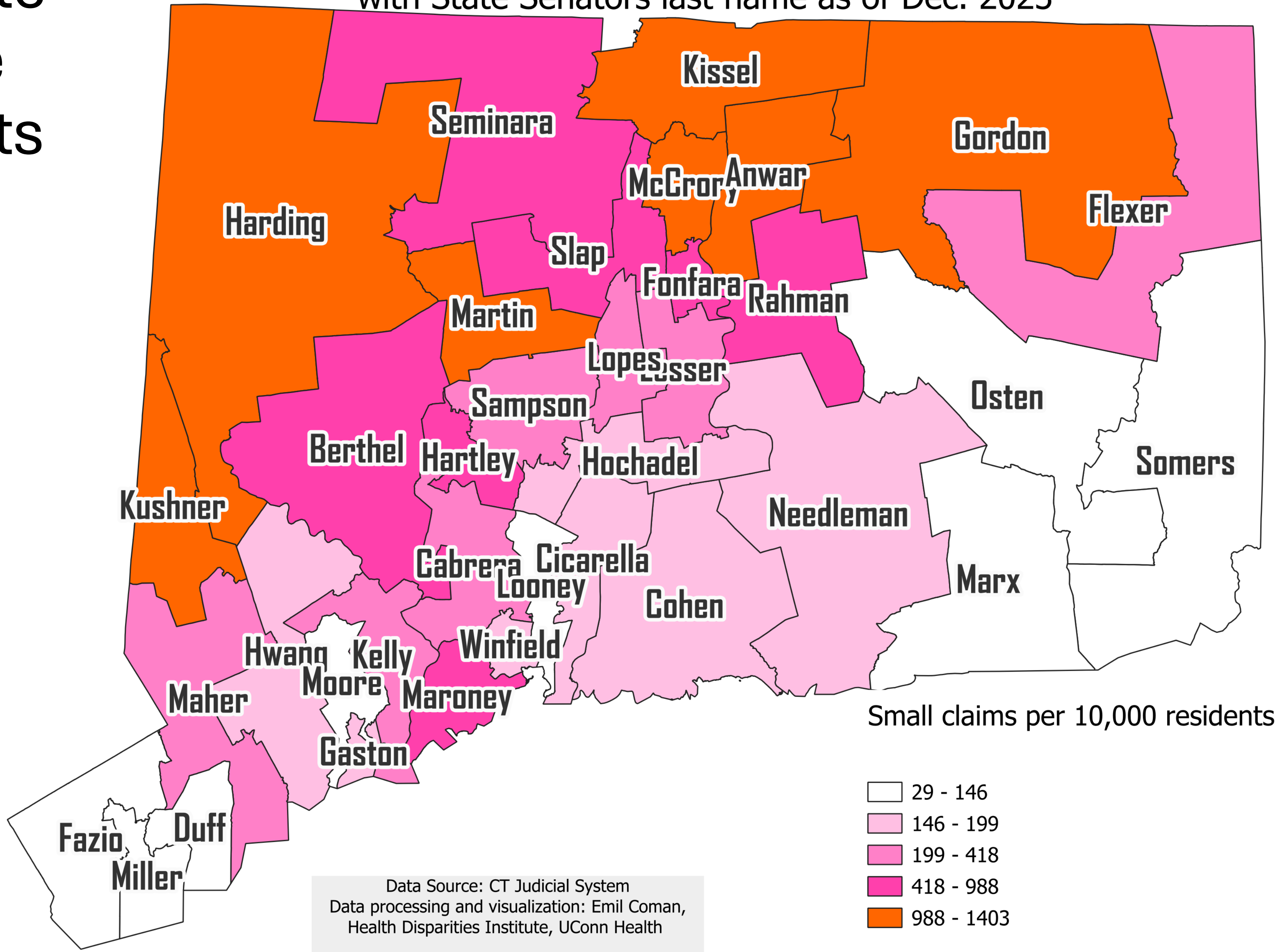
$I_{income} = .329, z = 3.343$





CT State Senate Districts

Small Claims medical debt in 2019 in CT by State Senate district, per 10,000 residents, with State Senators last name as of Dec. 2023



Data Source: CT Judicial System
Data processing and visualization: Emil Coman, Health Disparities Institute, UConn Health

2 Predictors of Medical Debt rates in 2019

N = 36, CT state senate districts

	Naïve Beta	Naïve Beta t	Spatial^{L} Beta	Spatial^{L} Beta z
% Non-White	0.077	0.432	0.130	1.037
Gini inequality	-91.96	-1.264	-108.73^{SIG}	-2.120

Notes: ^{L} - Spatial lag regressions in GeoDa; ^{SIG} - z/t > 1.96.

Seems to suggest that CT state senate districts with more income inequality have a lower debt rate.

Conclusions

1. Estimating effects with spatial data depend require the modeling of spatial ‘auto’-correlation, or non-independence.

2. Causal thinking with spatial data forces one to consider two networks: with links between cases (regions), and with links between variables.

3. Spatial data allows for aggregation and mapping of evidence aimed at legislators, or the public.