Comparative benefits of analyzing spatial aggregate data using Stata’s sp versus gsem & sem

The comparison follows two main research questions that could be of interest for researchers in other states or looking at the whole US:

RQ1. How much shorter lives do residents in regions with more minorities(non-White) live?

RQ2. How can one disentangle the effect of minorities on life expectancy from the related ones of income, or socio-demographic factors?

**1. Why spatial analytics**

1.i. Spatial ‘autocorrelation’/non-independence nearly universally ignored; try it out on your state’s (census tract) level from CDC: do residents living in higher-number FIPSs live longer? Why would they?

1.ii. Investigations of health disparities become difficult because in the US %Minorities is highly co-related with a host of regional/spatial indicators, like income, unemployment, poverty, education.

1.iii. [CDC Social Vulnerability Index](https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/SVI_documentation_2018.html); [Area Deprivation Index (ADI)](https://www.neighborhoodatlas.medicine.wisc.edu/)

1. iv. Missed opportunities

- Individual (patient e.g.) data geocodable and linkable with spatial data [Race, Ethnicity, Neighborhood Characteristics, and In-Hospital Coronavirus Disease-2019 Mortality](https://pubmed.ncbi.nlm.nih.gov/34334737/): analyses not more than 1 DV/outcome/effect at a time modeled, neither spatial nor multilevel. “The effect of ADI was significant, with residents in the most disadvantaged neighborhoods being at highest risk of death. While Black and Hispanic/Latino patients were 3–5-fold more likely to reside in such neighborhoods*, ADI, but not race, predicted increased mortality*.” - %Black or %H were not analyzed/modeled.

\* Y**i** = α. + β.·X**i** + ε**i** (classic regression)

\* For each region r, there are some regions s that are neighboring it, hence their effects need to be controlled for, which is done by incorporating an ’autocorrelation’ term ρ [1]:

Y**r** = ρ·(∑**s** W**rs**·Y**s**)**r** + α. + β.·X**r** + ε**r** (spatial regression)

where Wrs is the R x R ‘spatial weight matrix’, which has mostly zeros (in diagonals first of all), and non-zeros for those regions that are the neighbors of each r region.

\* With a small 5 regions (R = 5) setup for simplicity, and assuming only 1’s in the weight matrix where a neighboring relation exists, we can rewrite

Y**1** = ρ·(1·Y**2** + 1·Y**3** + 1·Y**4**) + α. + β.·X**1** + ε**1**, which says that regions 2, 3, and 4 are neighbors of 1

Y**2** = ρ·(1·Y**1** + 1·Y**4**) + α. + β.·X**2** + ε**2**, which says that regions 1 and 4 are neighbors of 2

Y**3** = ρ·(1·Y**1** + 1·Y**4** + 1·Y**5**) + α. + β.·X**3** + ε**3**, which says that regions 1, 4, and 5 are neighbors of 3

Y**4** = ρ·(1·Y**1** + 1·Y**2** + 1·Y**3** + 1·Y**5**) + α. + β.·X**4** + ε**4**, which says that regions 1, 2, 3, and 5 are neighbors of 4

Y**5** = ρ·(1·Y**3** + 1·Y**4**) + α. + β.·X**5** + ε**5**, which says that regions 3 and 4 are neighbors of 5



**2. Why CT**

2.i. Richest state and poorest residents

2.ii. Large health disparities

2.iii. Effects within CT vs. Within other states

3. Illustrate sp on CT census tracts data from CDC

3.i. Use shape files into Stata

3.ii. Merge with ‘actual data’ from other sources

3.iii. Create weight matrices: several

3.iv. Create lagged components for variables

3.v. Create maps

**4. Illustrate spatial** gsem **and compare to** sp

4.i. Naïve SEM/regression

4.ii. Spatial SEM

4.iii. Mediation

4.iV. Spatial Confirmatory Factor Analysis CFA

RQ1.

RQ2.

**5. Extensions:**

5.i. Nonrecursive SEM to estimate model implied spatial correlation

5.ii. Multilevel spatial SEM

5.iii. Spatial Linear Mixed Model

In some small states like CT, with only 8 counties, instead of multi-level spatial models, one might need to settle on multi-**group** models instead.

5.iv. **Other extensions: 2-level CFAs for US census tracts and counties: local meanings**

\* Naïve/a-spatial simple regression (or SEM) models indicate that residents from CT census tracts with 10%p more minorities live 0.67 years shorter lives.

\* Spatial simple regression (or SEM) models indicate that CT census tracts with 10%p more minorities live 0.30 years shorter lives.

\* Stata spregress spatial regression (or SEM) models indicate that CT census tracts with 10%p more minorities live 0.41 years shorter lives (direct effect), or 0.41 years shorter lives (total effect, including from neighboring regions).

\* Spatial mediation SEM models indicate that CT census tracts with 10%p more minorities live 0.38 years shorter lives, of which 42% (0.16) happens because higher %minority regions have lower income.

\* Spatial multiple-group SEM models indicate that the county level effects of 10% more minorities on Life Expectancy varies across the CT counties (n = 8) from a -2.13 To a -0.32 (fewer) years.

**~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~**

We demonstrate of the powers of the underutilized Stata’s spatial analytical module sp, with an eye on the broader and older path analytic modeling framework (gsem & sem, for Structural Equation Modeling, SEM).

Spatial aggregate data has become widely available, yet analysts often ignore its spatial structure (regions have neighbors, and neighboring regions are more similar than by chance). Research often reports artificial naïve/a-spatial associations that ignore this spatial non-independence. We analyze public data from the CDC, on social vulnerability and life expectancy, at census tract level, using the state of CT in USA as illustration.

We compare:

(1) The spregress modeling options against SEM models that include the outcome’s spatial lag as co-predictor;

(2); A two steps mediation models with spregress against SEM with indirect effects;

(3) The total effects of a spatial predictor on a spatial outcome estimated with spregress by adding up effects from neighbors to each region (and back), against nonrecursive SEM models that use spatial lag versions of each spatial variable as instrumental variables.

We point to several extensions of spatial modeling into the SEM approach, like spatial factor analysis and spatial ‘causal’ mediation models, and contrast Stata’s utilities against GeoDa and Mplus comparable models.

1. Anselin, L., *Spatial econometrics: methods and models* [*https://drive.google.com/file/d/1qLlFsVyh\_o1h4IEO0OxhNSD6sdX4lJ20/view?usp=sharing*](https://drive.google.com/file/d/1qLlFsVyh_o1h4IEO0OxhNSD6sdX4lJ20/view?usp=sharing). Vol. 4. 2013: Springer Science & Business Media.