INFO 7470/ILRLE 7400
Geographic Information Systems

John M. Abowd and Lars Vilhuber
March 29, 2011

Outline

- Examples from OnTheMap
- Basic concepts
- Census Bureau definitions
- Geocodes and vintages
- Coding longitude and latitude
- Geospatial functions

Demo of OnTheMap

lehdmap.did.census.gov

Census Bureau Definitions

- TIGER/Line Shapefiles
  - 2000
  - 2010

Geocodes and Vintages

- Layers, roads, features all have vintages
- At the beginning of a vintage (tabulations for a given decennial census) tabulation boundaries respect political boundaries
- Block definitions get revised, political boundaries change
- For the 2010 Census, some states chose to use block boundaries that do not respect political boundaries
- Federal law specifies which blocks are tabulated with a given political entity

Coding Latitude and Longitude

- Conventions
  - Coded in degrees
  - Eastern hemisphere positive longitude (western negative)
  - Northern hemisphere positive latitude (southern negative)
  - Six decimal places is accurate to within a few feet except near the poles

Geocoding in OTM

- Longitude and latitude from MAF and from PitneyBowes Business Insight (formerly Group 1)
- Address is standardized (more next week)
- Standardized address is coded to:
  - Rooftop (center of parcel)
  - Block face (location on a street in block definition)
  - Block internal point
  - Higher-level aggregate internal point

Computing Distance and Direction

- Vincenty formulae for great circle distances
- Computing angles from due north
- Examples on Sessions page

INFO 7470/ILRLE 7400
GIS Analysis Methods

John M. Abowd and Lars Vilhuber
March 29, 2011

Outline

- Example application to spatial autocorrelations
- Uses the LEHD Infrastructure data
- Thanks to Ian Schmutte

Basic Statistical Model

$$y_{it} = \theta_i + \psi_{\mathrm{J}(i,t)} + x_{it}\beta + \varepsilon_{it}$$

- The dependent variable is compensation
- The function J(i,t) indicates the employer of i at date t.
- The first component is the person effect.
- The second component is the firm effect.
- The third component is the measured characteristics effect.
- The fourth component is the statistical residual, orthogonal to all other effects in the model.

Matrix Notation: Basic Model

$$y = D\theta + F\psi + X\beta + \varepsilon$$

- All vectors/matrices have row dimensionality equal to the total number of observations.
- Data are sorted by person-ID and ordered chronologically for each person.
- *D* is the design matrix for the person effect: columns equal to the number of unique person IDs.
- *F* is the design matrix for the firm effect: columns equal to the number of unique firm IDs times the number of effects per firm.

Statistical Model

$$Corr(x_i, x_j) = f(\|s_i - s_j\|)$$

where x_i, x_j are the data items at locations i, j

s_i, s_j are the spatial coordinates

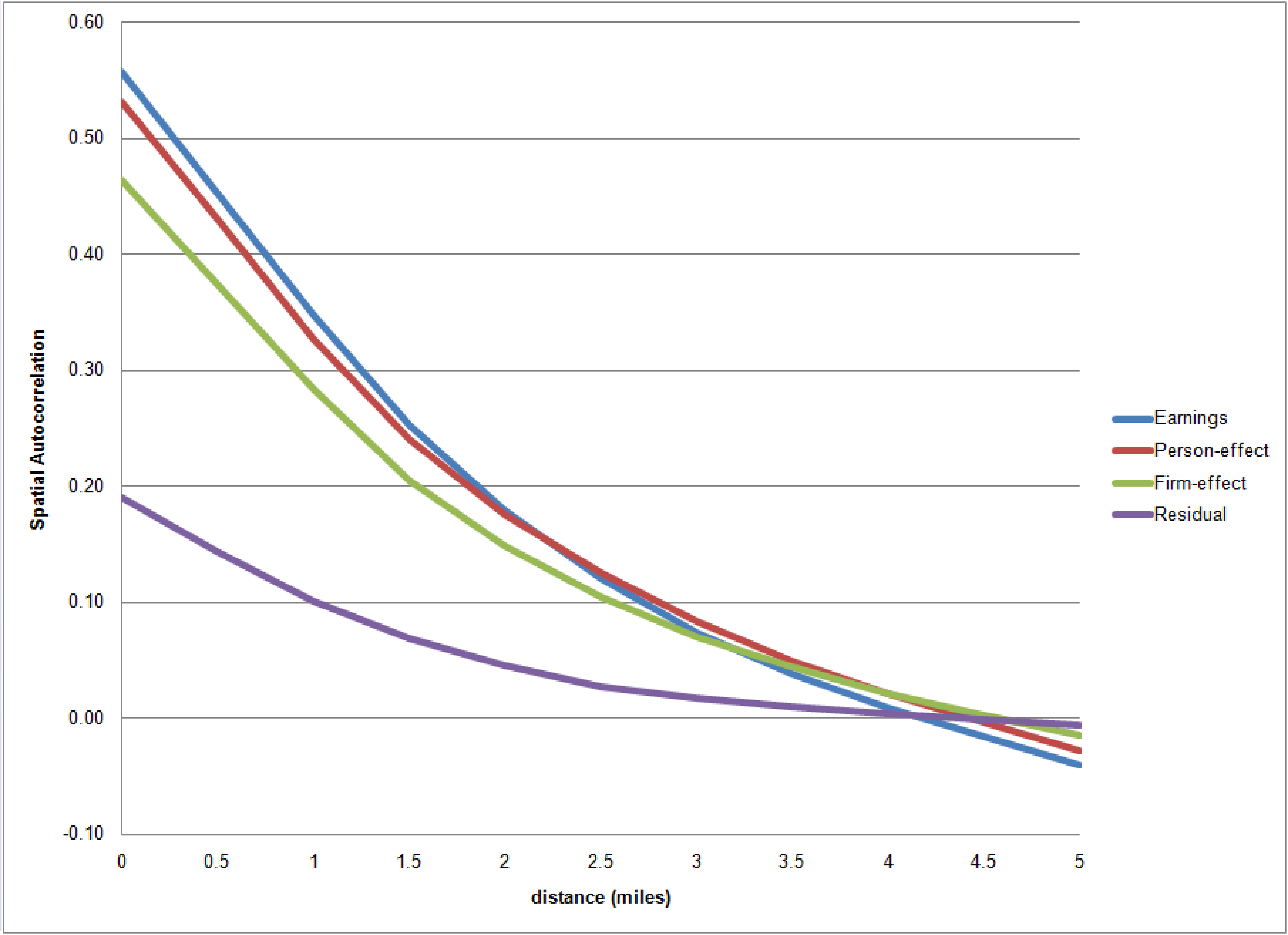
Spatial Autocovariance Function

$$\hat{f}(\delta) = \sum_{i=1}^{N} \sum_{j=1}^{N} \phi\left[\frac{|\delta - A_{ij}|}{\sigma}\right]\left(X_i - \overline{X}\right)\left(X_j - \overline{X}\right)$$

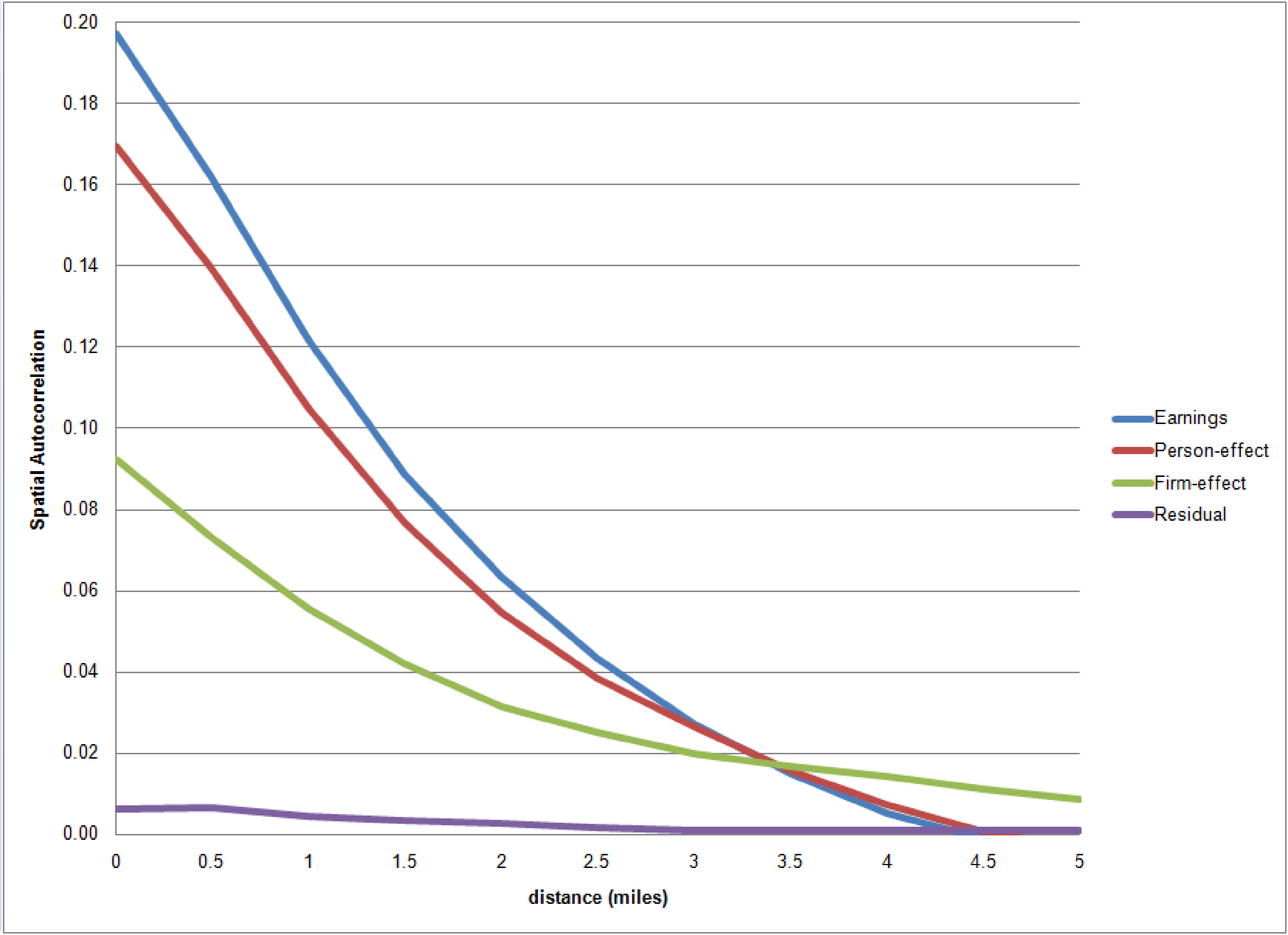
ϕ is the standard normal kernel.
δ is the distance to be measured.
A_{ij} is the distance between *i* and *j.*
σ is the kernel bandwidth

To convert this to the spatial autocorrelation, one must divide the resulting estimate by relevant product of standard deviations. With the normal kernel, this is just the sample variance.

Tract-level Spatial Autocorrelation



Block-level Spatial Autocorrelation



Discussion I

- Implement this estimator for tract-level and block-level means of all earnings and the components of the AKM decomposition.
- Compute f() at distances from 0 to 5 miles at half-mile gridpoints.
  - A_{ii} is measured as the great-circle distance between internal points of the block or tract. For the block-level estimates, the bandwidth parameter, sigma, is set to 0.5.
- For the tract level estimates, bandwidth is set at 0.7. Since the computation scales in the square of the number of observations

Discussion II

- For the block-level calculation some simplification is required.
- Randomly sample block pairs at the rate of 1/100. For a hypothetical MSA with 5; 000 blocks, which would be a fairly small one for this study,
  - this means the spatial autocorrelation function is estimated from approximately 125; 000 unique data points.
- To satisfy the disclosure avoidance restrictions required to publish these results, each point in the figures represents the unweighted average of the estimated spatial autocorrelation across 30 MSAs.
- There is some variation between the MSA-level estimates, but not enough to change the qualitative features of the plot. These plots are representative of most of the individual MSAs.