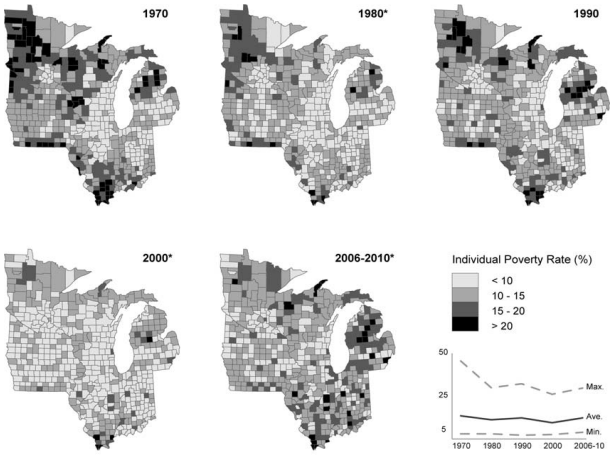
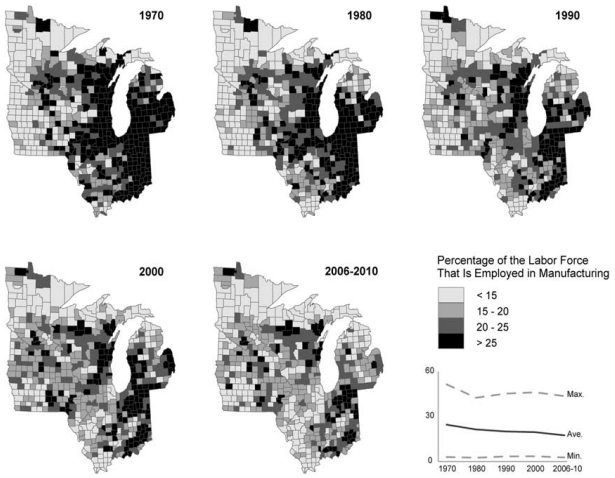
Case studies

- Look at 2 recent papers:
  - Curtis et al, 2019, The Spatial Distribution of Poverty and the Long Reach of the Industrial Makeup of Places: New Evidence on Spatial and Temporal Regimes. *Rural Sociology* 84(1):28-65
  - Elson et al, 2020, Application of kernel smoothing to estimate the spatio-temporal variation in risk of STEC O157 in England. *Spatial and Spatio-temporal Epidemiology* 32:100305
- What do I look at when I skim/read a paper?
  - Goals/questions?
  - What are the data?
  - What methods were used?
    - Could be standard methods
    - Could be something new - if so, what's the new thing?, why needed?
  - Can their methods be used for data/questions I care about?
- Subject-matter scientist more likely to focus on the results

Curtis: Poverty and industry

- Looking at relationship between poverty and industrial environment
- Goal: how has this changed over time and over space?
- Two conceptual models (hypotheses):
  - regression slope more + over time
  - bigger increase in certain regions
- Data: Each county in 6 MidWestern states, 5 times: 1970, 80, 90, 2000, 2006-10
  - % pop below poverty threshold, logit transformed
  - % pop employed in manufacturing, service, and ag (sum < 100%)





Curtis: methods

- Model as published, with slight change of names:

$$Y_{st} = \boldsymbol{X}_{st}\boldsymbol{\beta}_t + t_t + \varepsilon_s + \gamma_{st}$$

- Two sorts of subscripts: s: spatial location, t: time period
- Easier to understand by building in pieces
- Analogy: designing a car
  - Don't start with a Ferrari
  - design a skateboard, then a go-kart, then a Model T, then $\cdots$
- Start simple, add complications as needed

The skateboard model

- Use linear regression to relate
  - Y_{st}: logit % poverty in county s and time t
  - X_{st}: % manufacturing in county s and time t

$$Y_{st} = \beta_0 + \beta_1 X_{st} + \gamma_{st}$$

- Answer Q using $\hat{\beta}_1$: how strong is relationship?
- Issues with the skateboard:
  - One coefficient - don't know whether changed over time or space
  - Ignores correlation over time, over space, or combination

The go-kart model

- Allow the regression coefficients to vary over time

$$Y_{st} = \beta_{0t} + \beta_{1t} X_{st} + \gamma_{st} = \boldsymbol{X}_{st}\boldsymbol{\beta}_t + \gamma_{st}$$

  - Possible because we have data across space for each time
- Easy to implement
  - Just requires setting up the appropriate X matrix

The Model-T model

- Include temporal and spatial correlations
- Raw data shows clear spatial correlation and some temporal correlation in Y_{st}
  - Better to evaluate with residuals
  - Pattern in Y_{st} may be a consequence of correlated X_{st} and indep errors
- From space-time material: 3 possible approaches
  - Metric: convert time lag to an equivalent spatial distance
  - Separable: one component for temporal correl., second for spatial
  - Joint model: every combination of (s, t) and (s^*, t^*)

The Model-T model

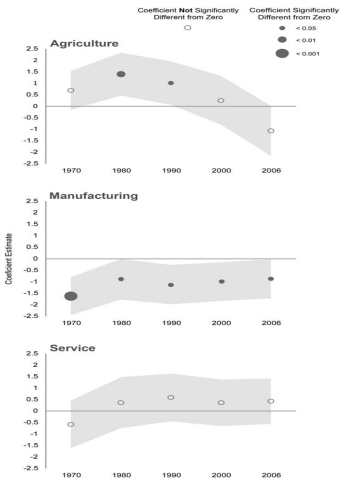
- They chose separable model: practical with such a large data set
- Temporal correlation: autoregressive, order 1 = AR(1)

$$t_t \mid t_{t-1} = \rho t_{t-1} + \tau_t$$

- Spatial correlation: CAR with 2nd order queen's neighbors

$$\varepsilon_i | \varepsilon_{-i} = \lambda \Sigma_{j=-i} c_{i,j} s_j$$

  - Second order: neighbors and neighbors of neighbors
  - Gives larger groups of "neighboring" counties
- Not easy to implement
  - Usual ML is extremely slow: need to invert large Σ many times
  - INLA: integrated nested Laplace approximations
  - Much, much faster, especially for conditional correlations (AR(1) and CAR)
  - Can't implement SAR models (at least now)
- This model answers Q1: does relationship change over time?
- Results shown by tables of coefficients and plots

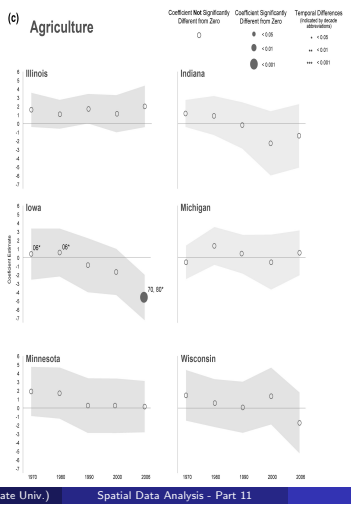


The Ferrari model

- Allow coefficients to vary spatially as well as temporally
- Various ways to do this
- They use a "regime" approach:
  - divide locations into spatial groups
  - allow coefficients to vary between groups
- Natural groups are states because may have different official policies

$$Y_{st} = \boldsymbol{X}_{st}\boldsymbol{\beta}_{Rt} + t_t + \varepsilon_s + \gamma_{st}$$

- One subscript change in the model: β's now depend on Region and time
- This model answers Q2: Do coefficient patterns over time vary spatially?
- Again, reported both by tables of coefficient values and plots



What's the new thing? What else could be done?

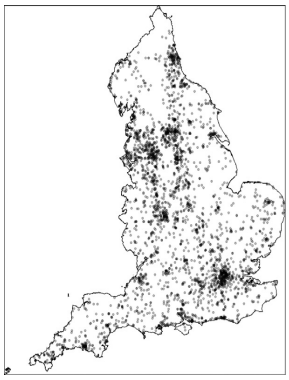
- New: Combination of ideas:
  - time- or space-time varying coefficients
  - with correlation in both space and time
- What else?
  - Here, coefficients for each time- or region-time-group are unrelated
  - What if they vary more smoothly?
  - e.g. 1970 coefficients closer to 1980's than others
  - Plots of cofficients support this idea
- Temporally / Spatially varying coefficient models
  - Allow coefficients to be correlated
  - By putting an AR(1) or CAR structure on the β's

Elson: spatial risk of disease

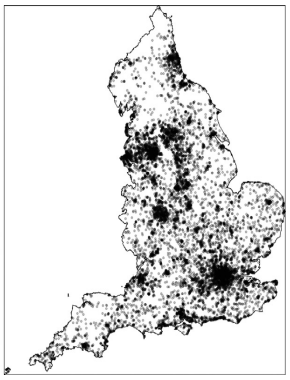
- Background:
  - STEC O-157: Shiga-toxin producing E coli. In US, called E. coli O-157 H7.
  - Bacterial food contaminant, results in intestinal upset, may kill
  - Major cause of food recalls in US.
- Goals/questions:
  - Primary: Is infection rate constant across England?
  - If not, where are areas with higher/lower rates?
  - Secondary: Effect of Travel abroad / Travel?
  - Secondary: Differences between lineages
  - Secondary: Do patterns change over time?

Elson: data

- Cases: N = 3592
  - GI patient, tested for STEC O-157
  - If positive, collect additional information
  - Locations (UK post code = $\approx$ 15 houses) of all positive tests
  - 2009 - 2015
  - Can spread within households, only considered 1st infection.
  - Care about rate = # / population, not # cases
- Controls: N = 14,368
  - Random sample of individuals in England
  - 4 times as many controls as cases



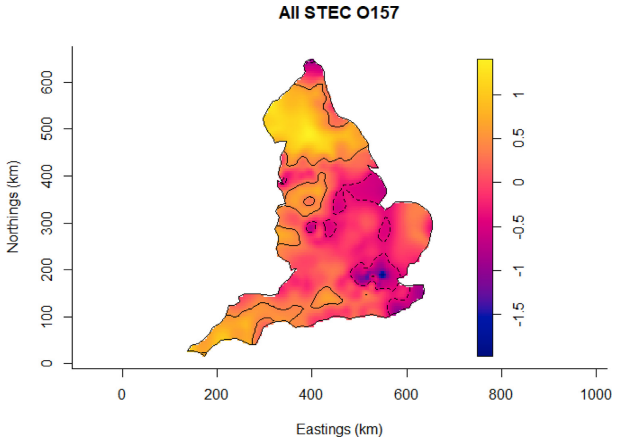
Cases

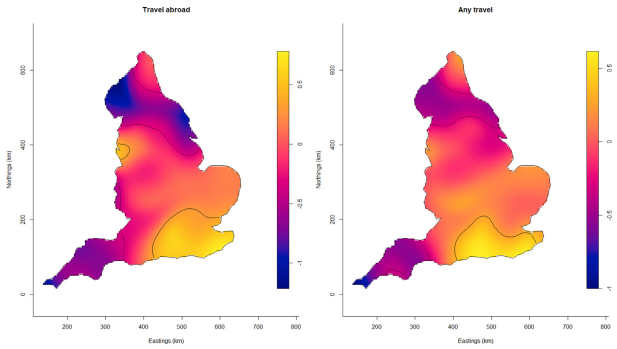


Controls

Elson: methods

- Look at intensity: IPP ideas
  - But don't have an image of population - only sample of locations
  - Apply kernel smoothing twice
- Estimate intensity of cases $= \lambda_{case}(\boldsymbol{s})$
- and relative intensity of controls $= \lambda_{control}(\boldsymbol{s})$
  - Have data on all confirmed cases
  - And a simple random sample of controls
- Interested in large scale pattern (not clustering / inhibition)
- Estimate $\log \lambda_{case}(\boldsymbol{s})/\lambda_{control}(\boldsymbol{s})$ on a grid across England
  - Control sampling fraction known: 14,368 / 52,640,000
  - so could estimate rate per person $= \log \lambda_{case}(\boldsymbol{s})/\lambda_{population}(\boldsymbol{s})$
  - Instead, scale relative to overall case/control ratio (4)
  - So $\log RR = 0 \Rightarrow$ at overall average rate
  - $\log RR = 1 \Rightarrow =$ cases 2are .72 times as likely
- Detail to care about: better to use same bandwidth for both smooths
- Randomization to assess whether unusually different from uniformity





Elson: ST methods

- Estimate spatial pattern of rate for each month
- Combine information spatially and temporally
- Two bandwidths: one for space and one for time
- Use animation to visualize changes over time

Overall

- Extension to what we've seen in class
- But methods and software available
  - relrisk() in spatstat
  - Kelsall, J.E., Diggle, P.J., 1995. Non-parametric estimation of spatial variation in relative risk. Stat. Med. 14: 2335–2342
- Are there other approaches?
  - The controls started as polygons with population
  - Could think about IPP model with X = log population

$$\log \lambda(\boldsymbol{s}) = \beta_0 + \beta_1 \log \textit{population}(\boldsymbol{s})$$

  - Or force $\beta_1 = 1$
  - β_0 is then the average log *RR*
  - Really want $\beta_0(\boldsymbol{s})$