Space-Time Models for Retail Credit

Bob Stine
Department of Statistics
The Wharton School of the University of Pennsylvania

April 2 2007

Acknowledgments

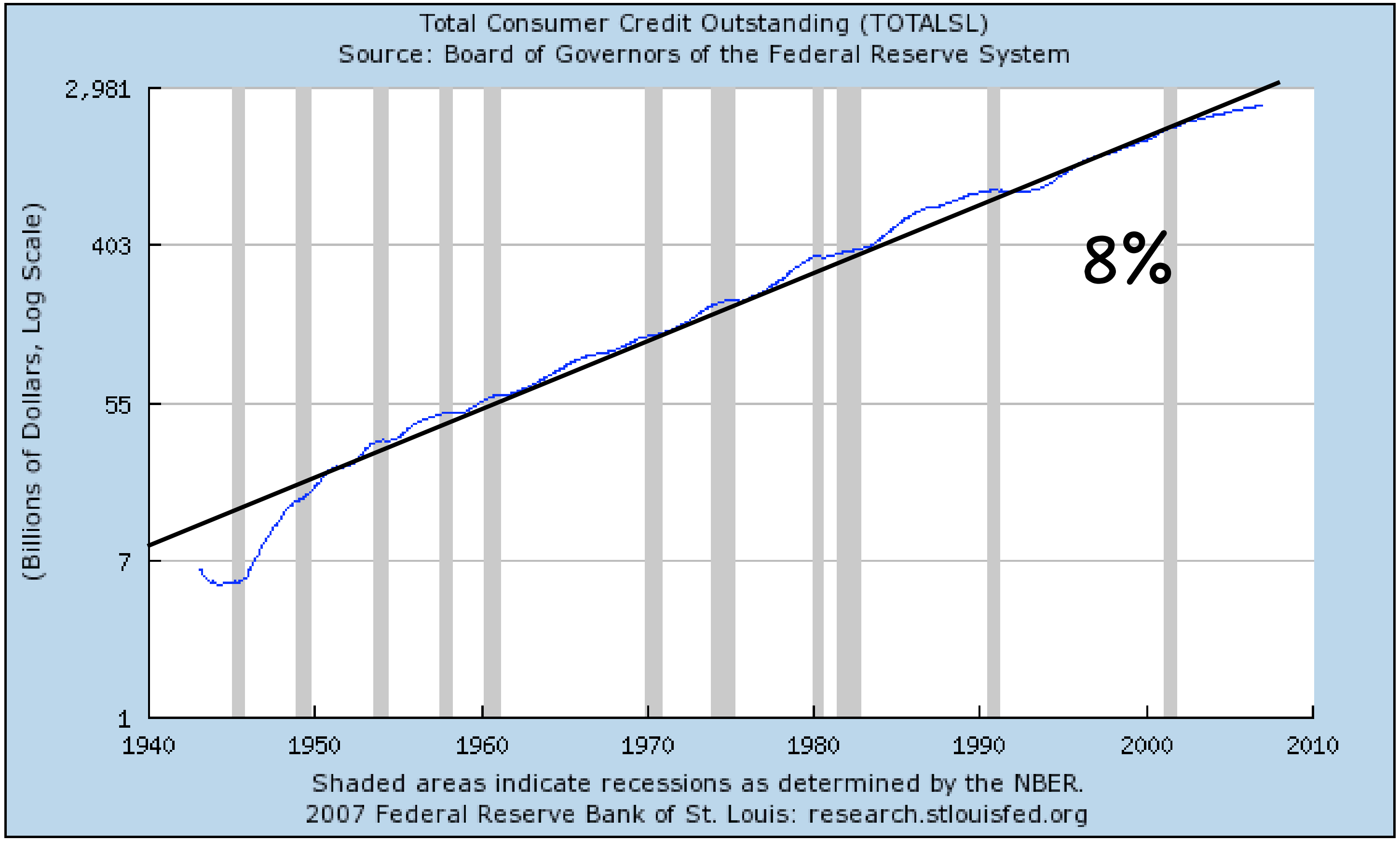
- Bill Lang and others at Federal Reserve Bank of Philadelphia
  - Collaboration to understand use of regional retail credit data
  - Source of banking data used in my analysis
- Nick Souleles from the Finance Department at Wharton

Overview

- Puzzles
  - Impact of local macroeconomic conditions
  - Consequences of spatial-temporal variation
  - Connection to stress-testing of models
- Data analysis
  - National trends
  - Local variation and dependence
  - Lots and lots of pictures...
- Models
  - What sort of model would capture the evident structure found in this analysis?

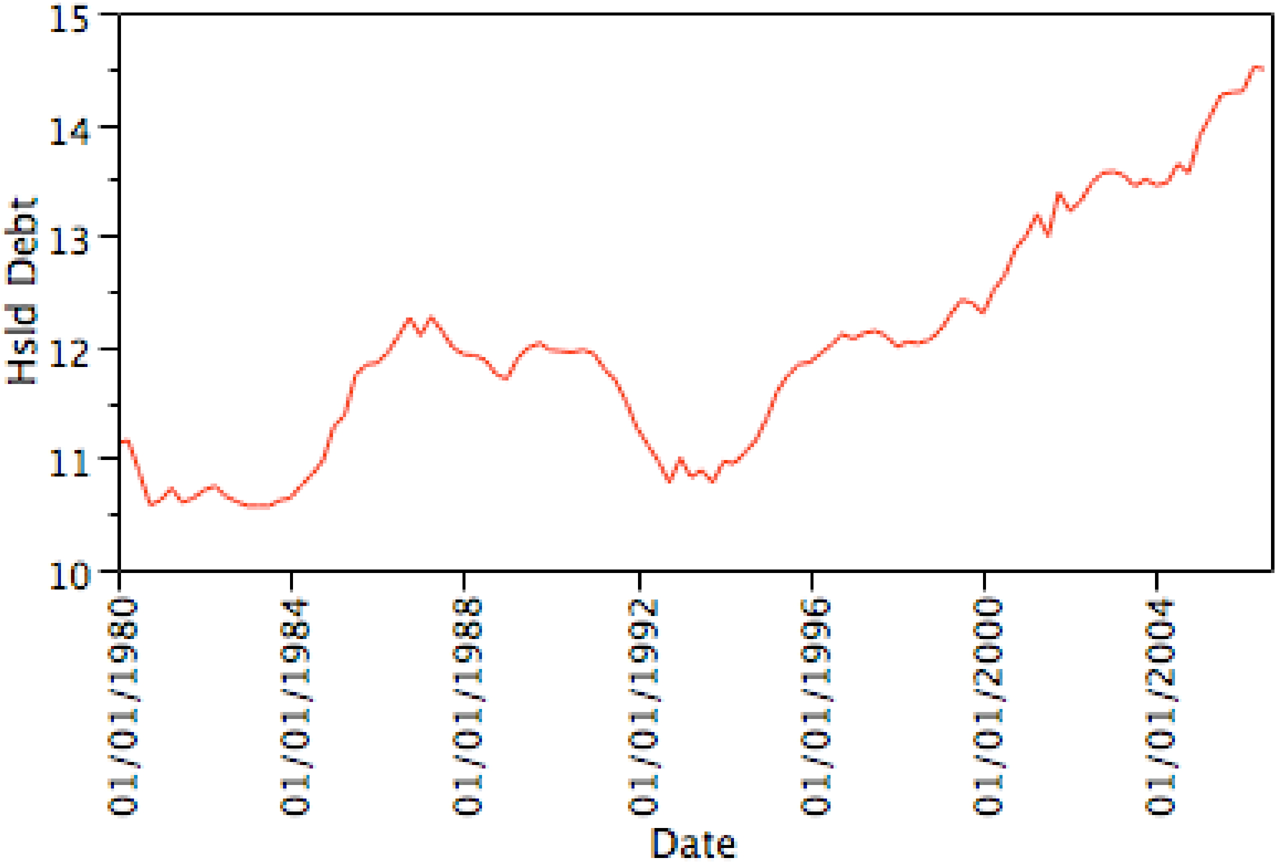
US Retail Credit Market

Consumer Credit



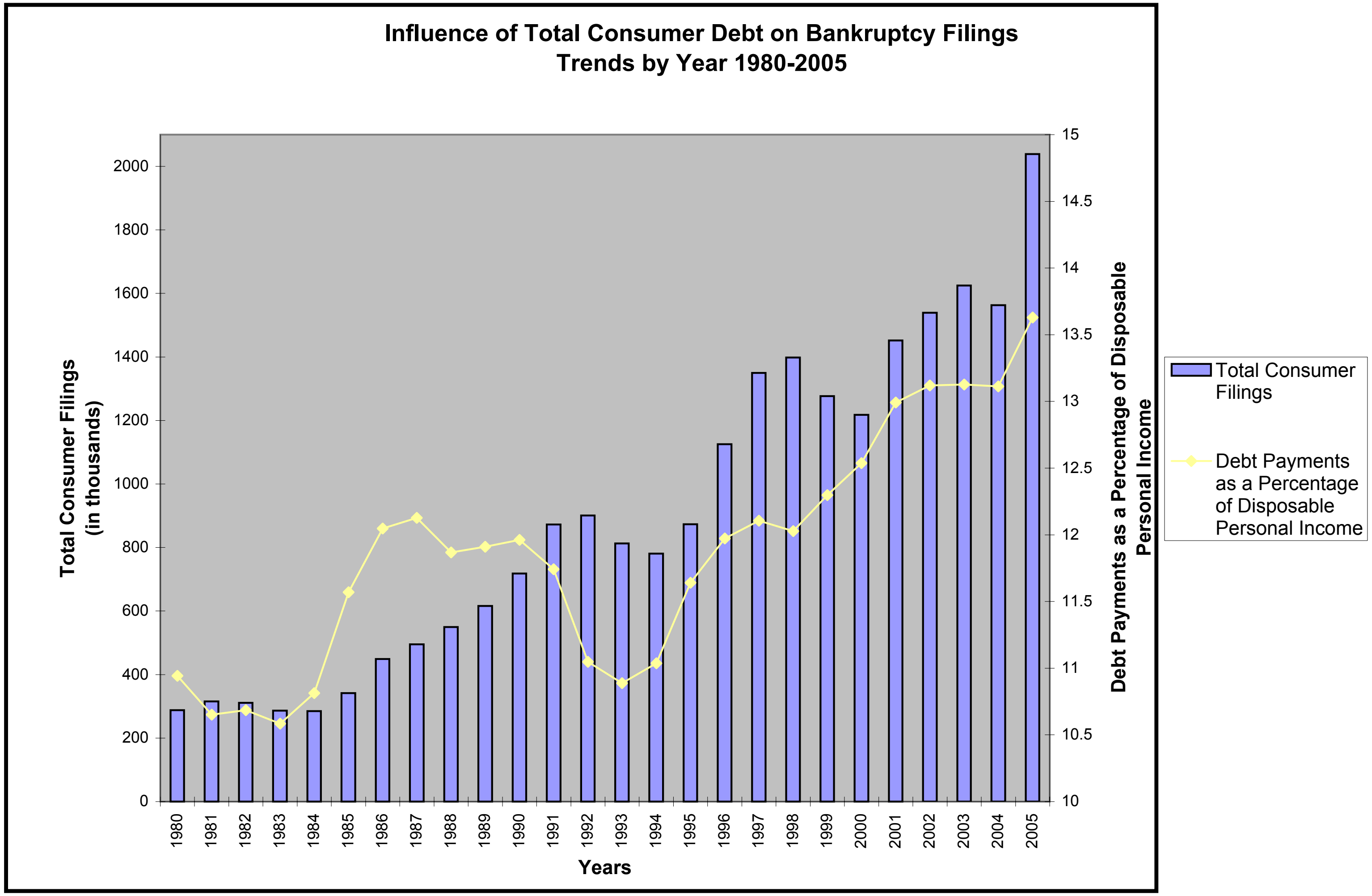
Remarkably steady growth.

Clouds on Horizon



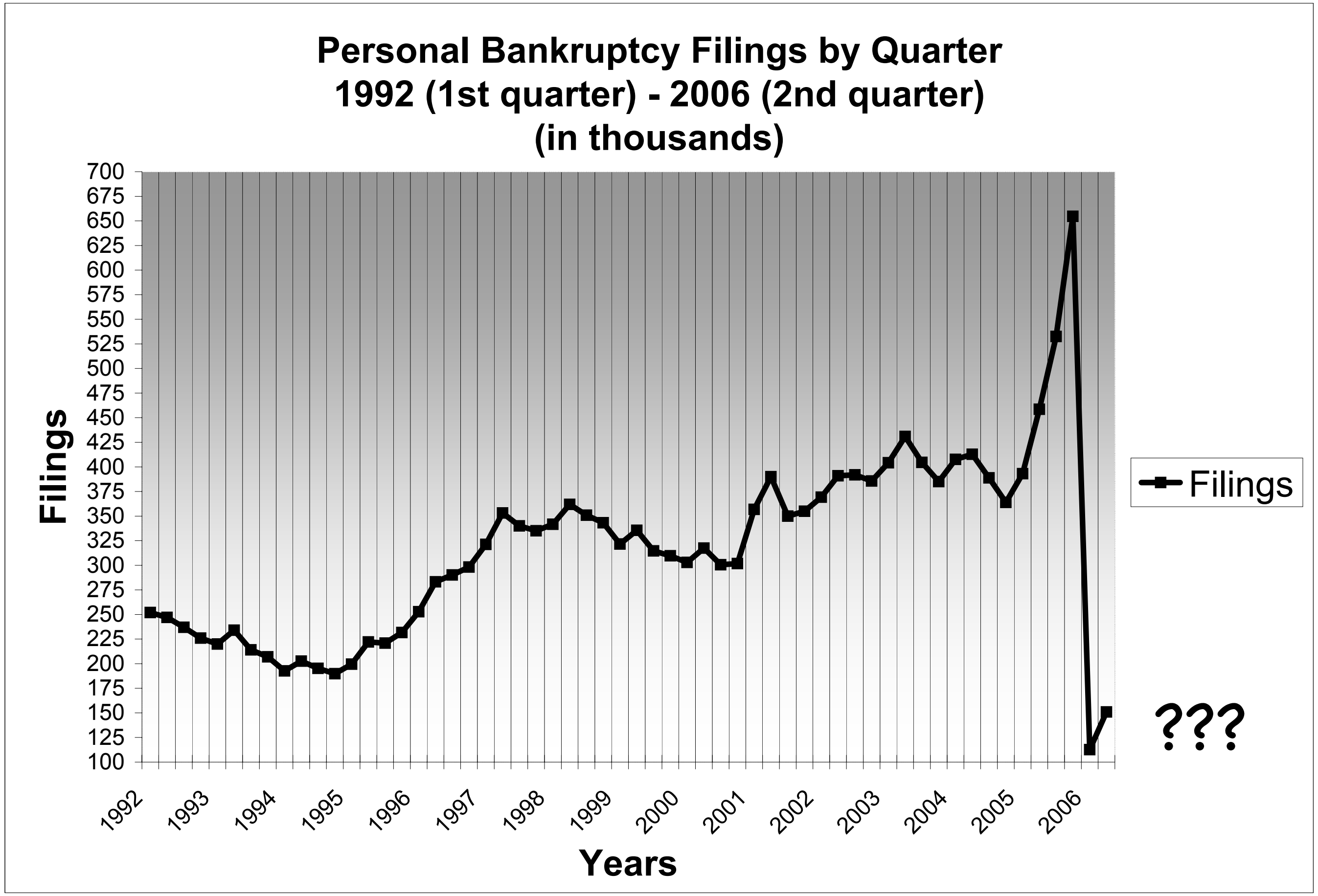
Household debt service payments as a percentage of disposable income (Federal Reserve)

Trends in Bankruptcy



American Bankruptcy Institute

Not Anymore?



American Bankruptcy Institute

Recent News

- Mortgage lending
- Higher than expected default rates in the sub-prime real estate market
- HSBC acquired loans from other originators in addition to those screened by Household division

FAULTY ASSUMPTIONS

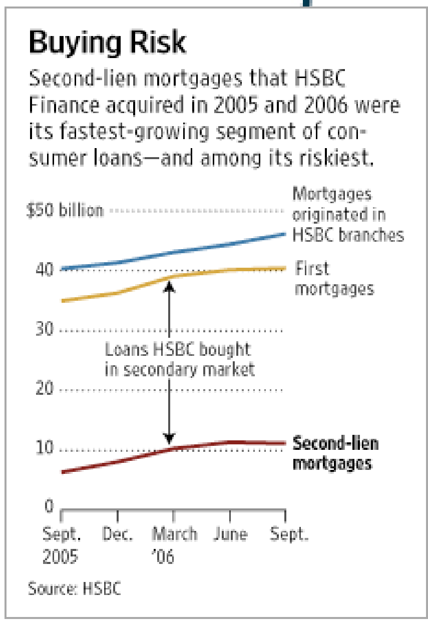
In Home-Lending Push, Banks Misjudged Risk

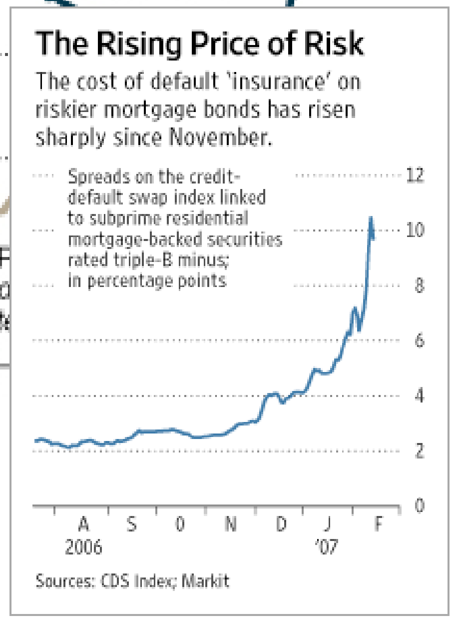
HSBC Borrowers Fall Behind on Payments; Hiring More Collectors

By CARRICK MOLLENKAMP

February 8, 2007; Page A1, Wall Street Journal

When the U.S. housing market was booming, HSBC Holdings PLC raced to join the party. Sensing opportunity in the bottom end of the mortgage market, the giant British bank bet big on borrowers with sketchy credit records.





Questions to Consider

Question 1

- What is the spatial variation of credit behavior and macroeconomic conditions in the US?
  - Often hear numbers like "the" unemployment rate or level of disposable income...
  - How much variation around the overall numbers is present?
  - What is the spatial distribution of the variation?

Question 2

- Do local economic conditions improve the fit of models that predict retail credit risk?
  - Models for risk routinely incorporate "bank data" that includes past default rates, utilization, ...
  - Do economic variables such as local employment levels add value beyond information in the bank data?

$$P(\text{default}|\text{bank},\text{macro}) = P(\text{default}|\text{bank})?$$

Question 3

- Does spatial variation in economic conditions produce a form of stress-testing?
  - Stress-test
    Does an overall model fit well when applied in times of economic distress?
  - Back-testing prescribed in Basel II regulations for building models of credit risk
  - Tricky to prescribe realistic conditions for test
  - Does spatial variation provide a natural framework for exploring model accuracy in periods of high economic stress?

Questions

1. Is there adequate spatial variation to support modeling credit risk?

2. Do local macroeconomic variables add value beyond usual bank information?

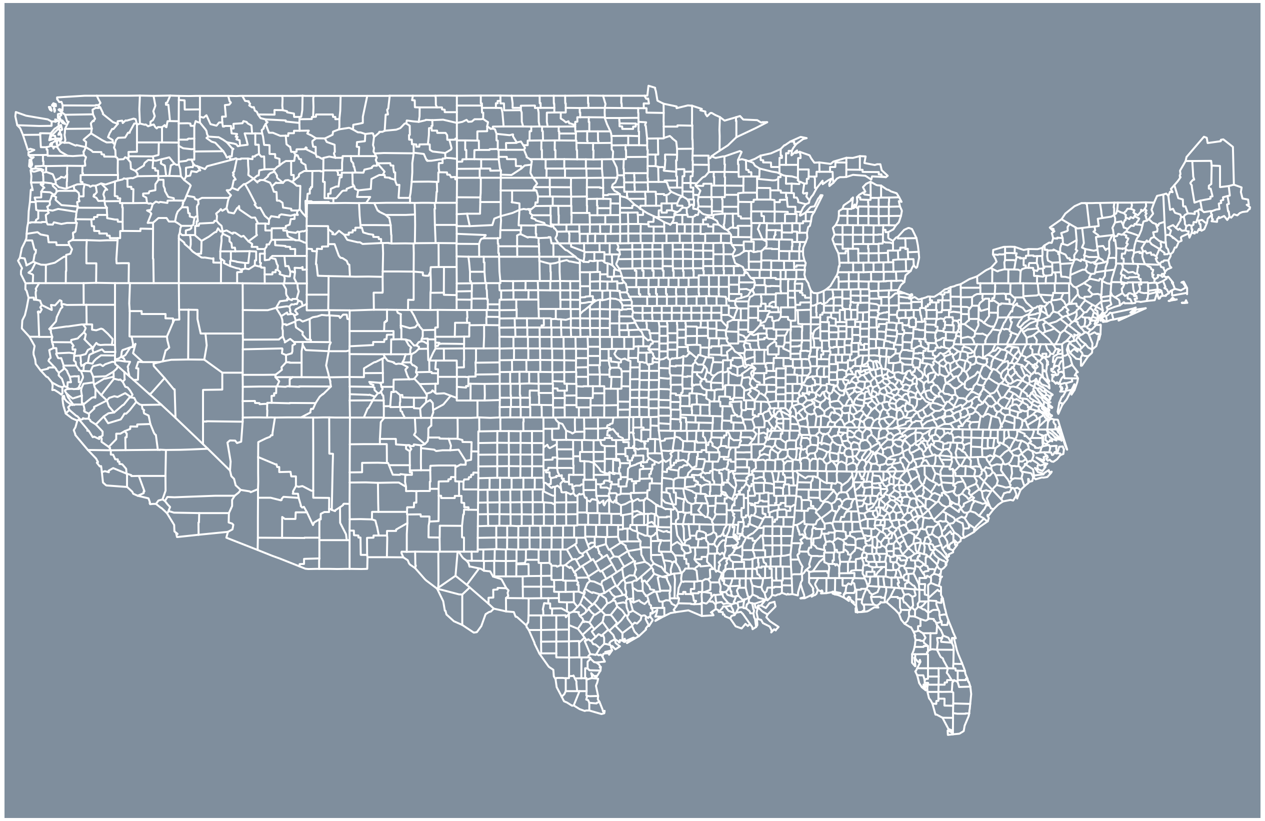
3. Do models perform consistently under local economic stress?

Data

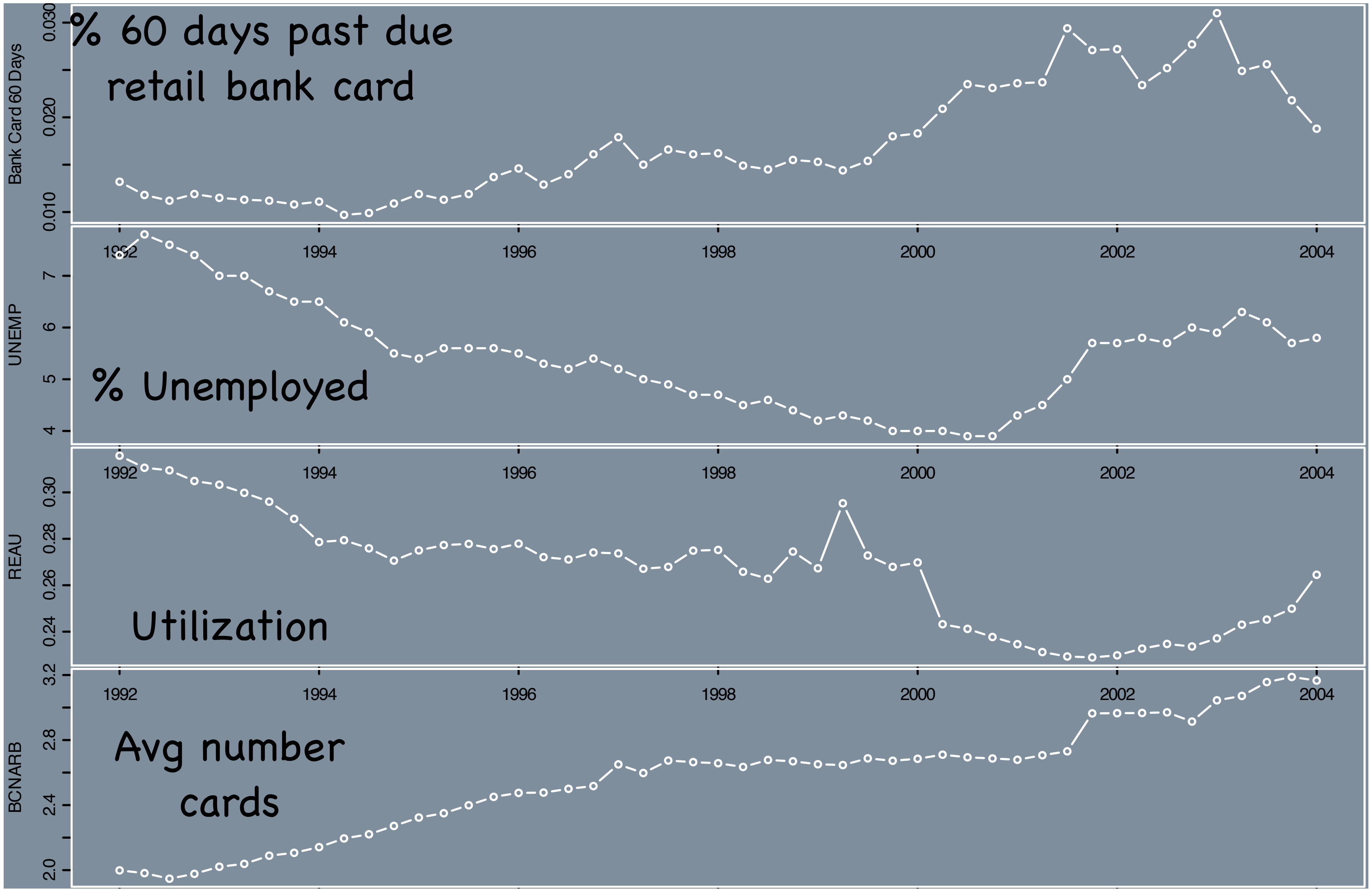
Getting the data is 90% of the work,
but none of the talk...

Regions

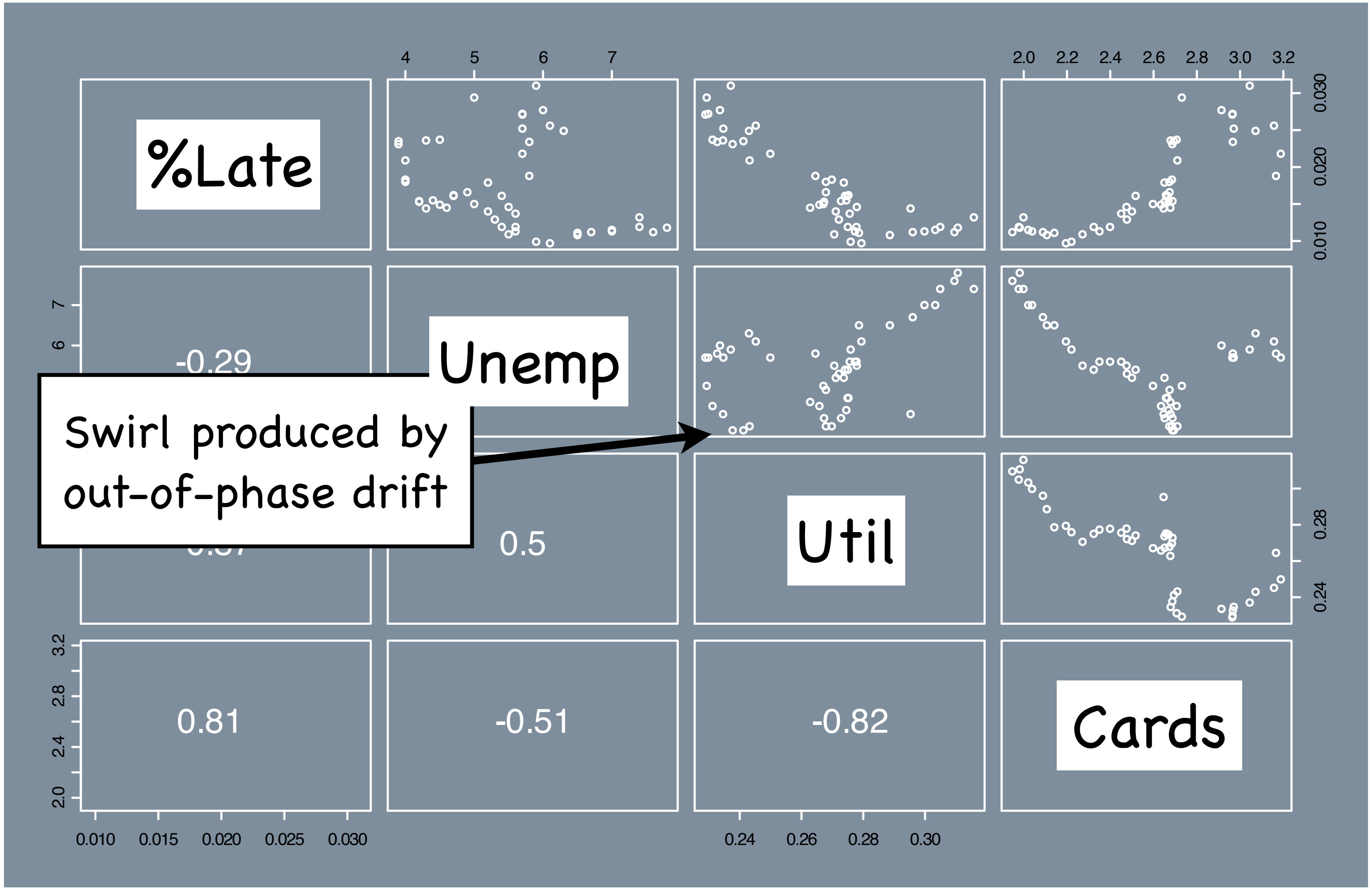
- County
  - Continental US has 3,000 counties
  - Diverse range of shapes and sizes



National Trends



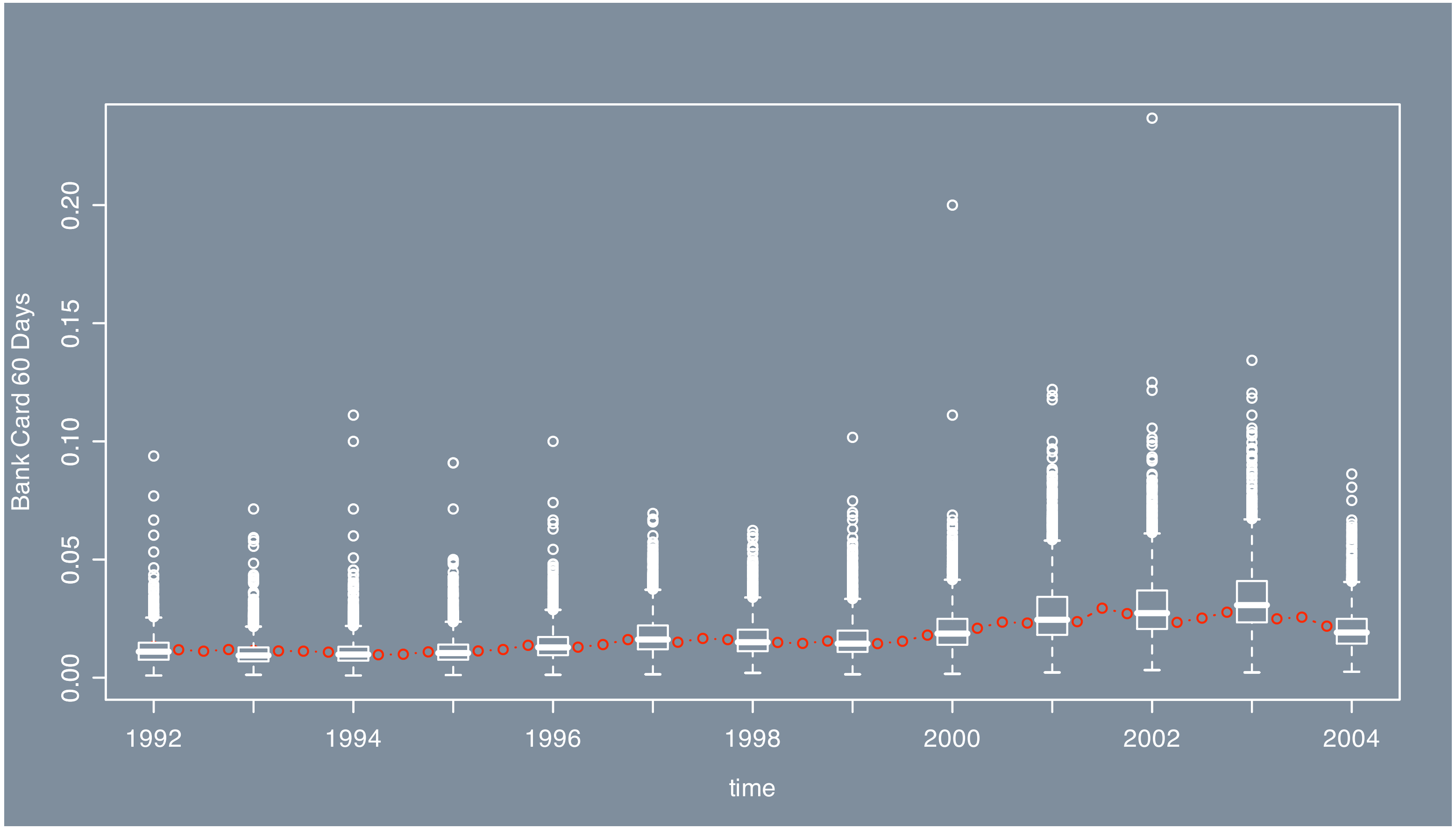
Association



Aggregation generates large correlations.

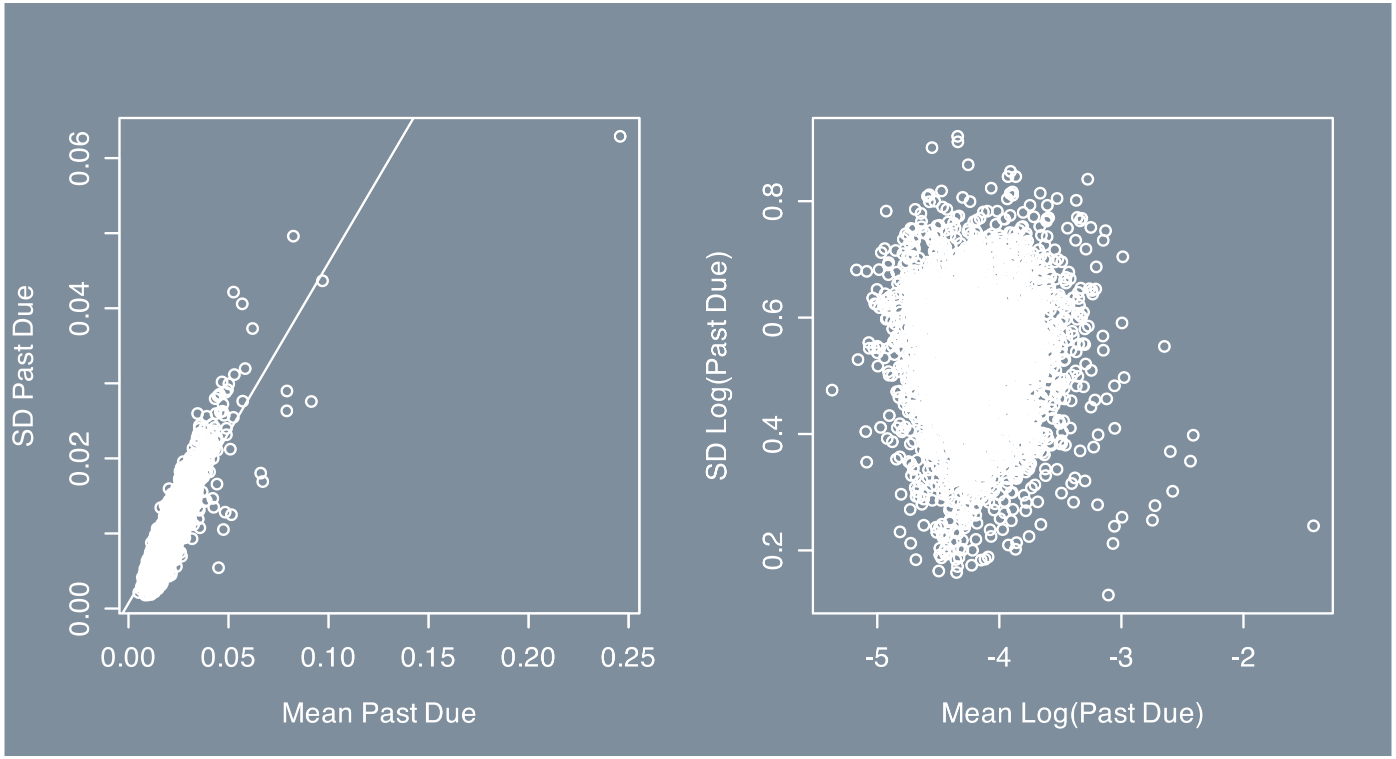
Regional Variation

% 60 Days Past Due



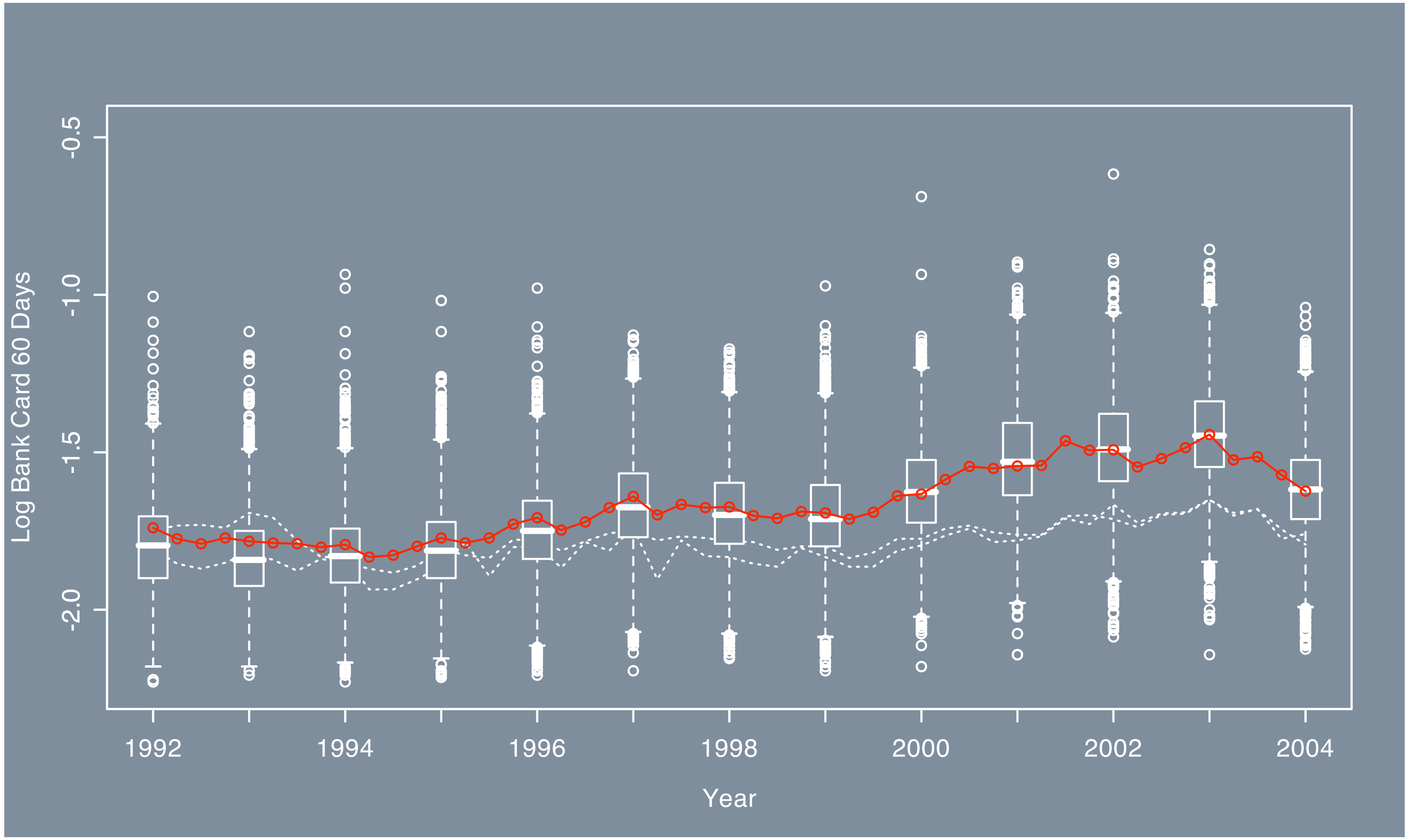
Transformation

- SD ≈ a + b Mean
  - Mean and SD for each county over 49 quarters
  - Log transformation stabilizes the variance

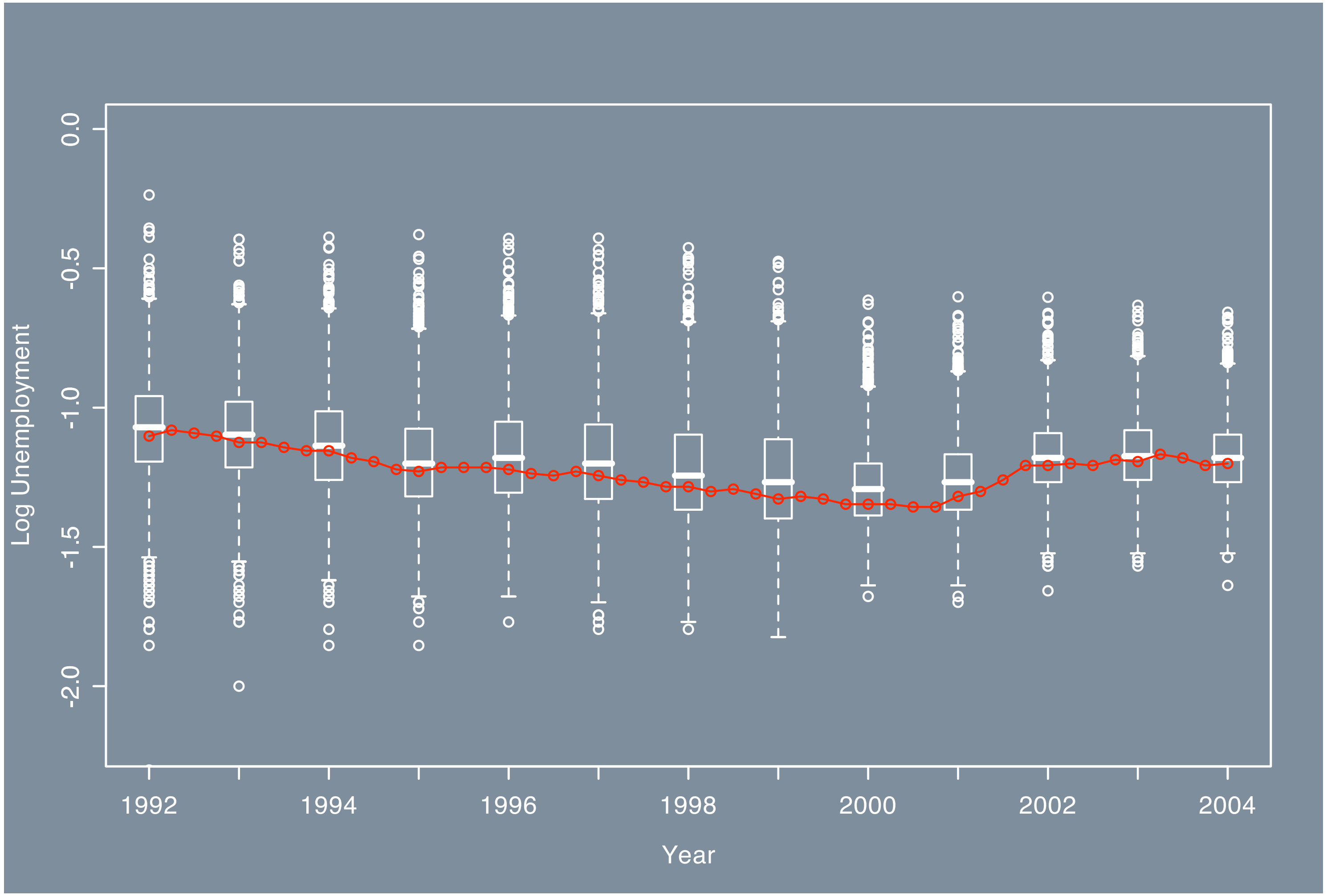


Regional Variation

Log % 60 Days Past Due

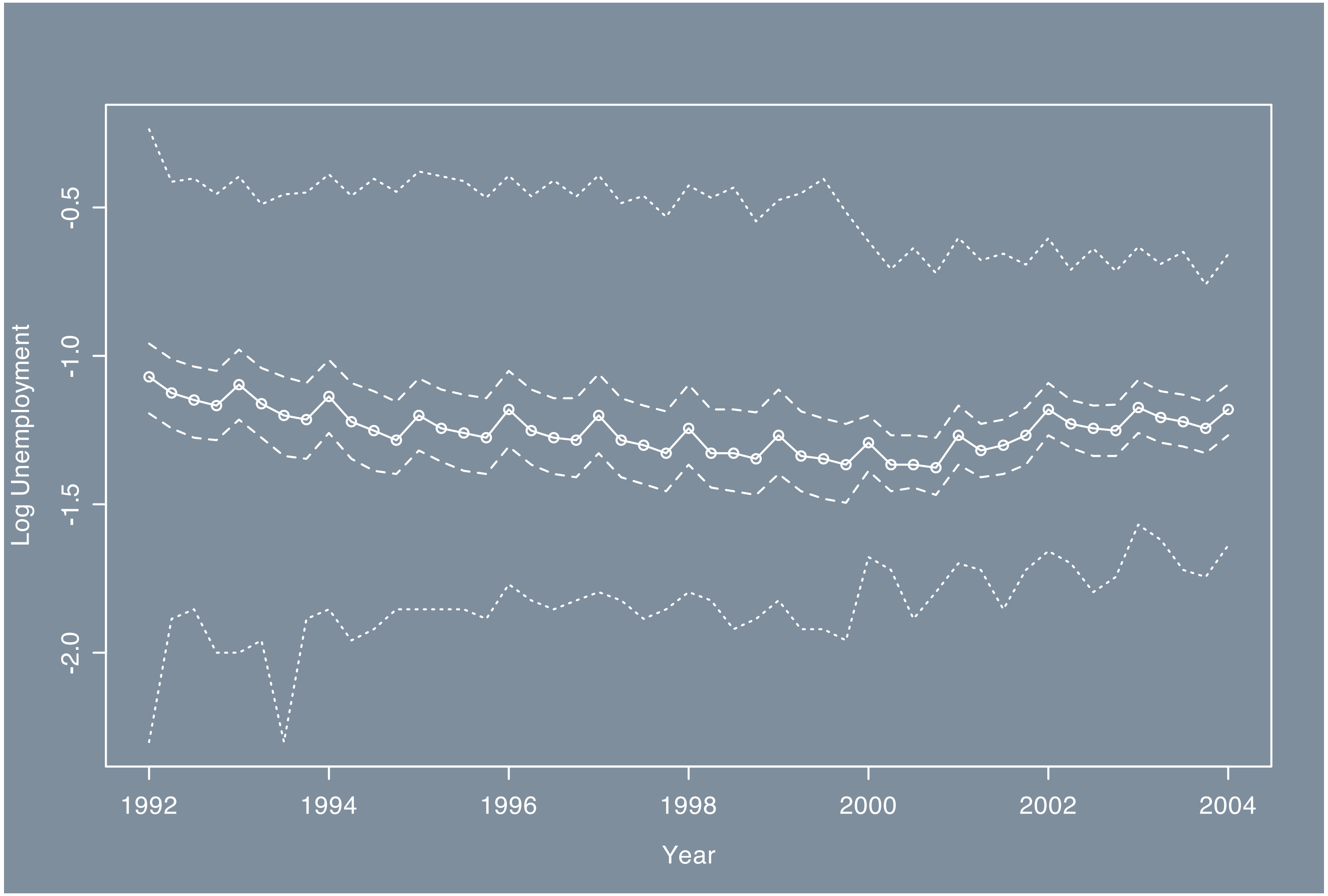


Variation in Unemployment



Quite large, with diminishing variation

Seasonal Variation

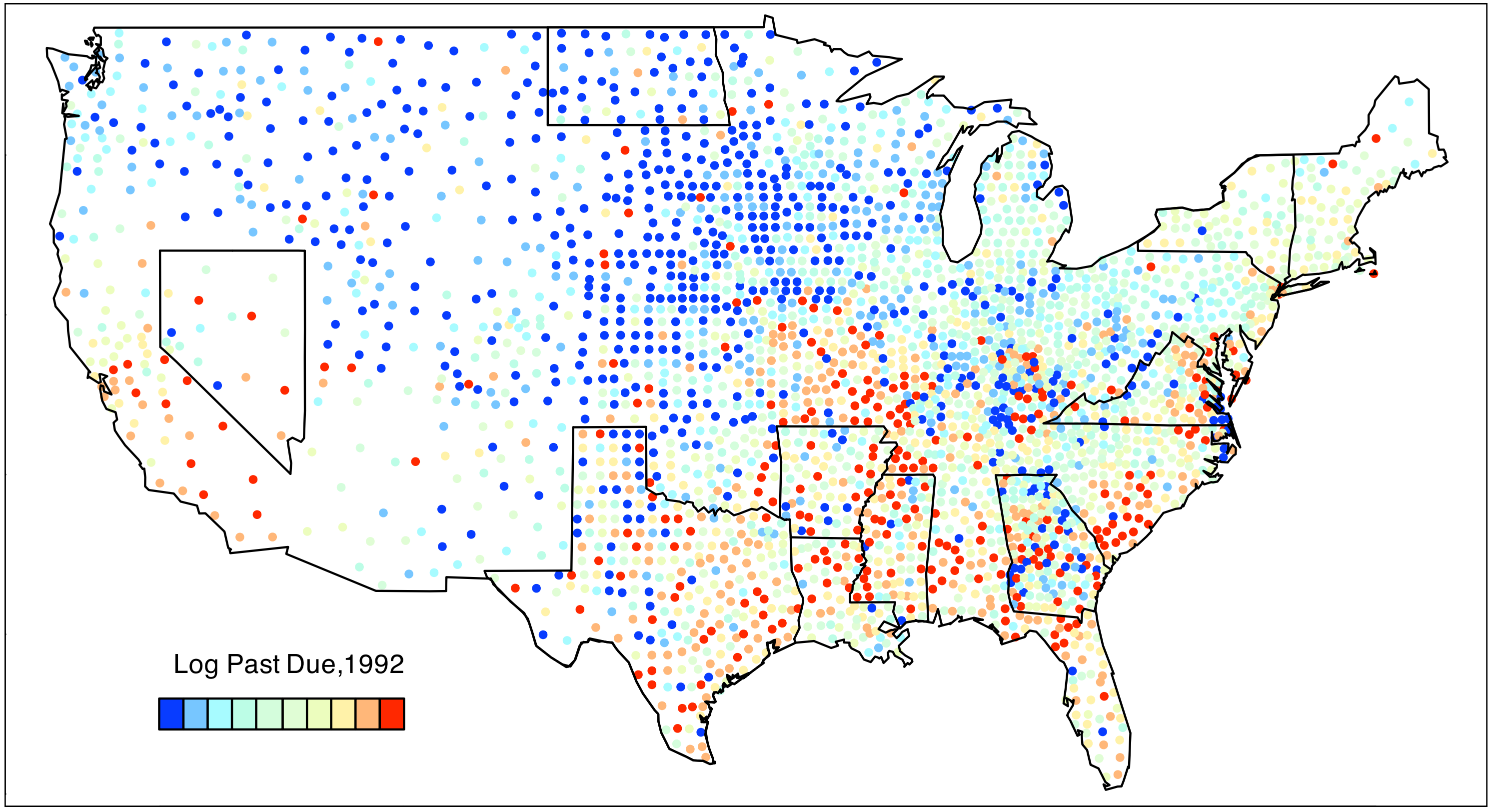


Annual “ripple” in unemployment, peaking in Q1

Spatial View of Data

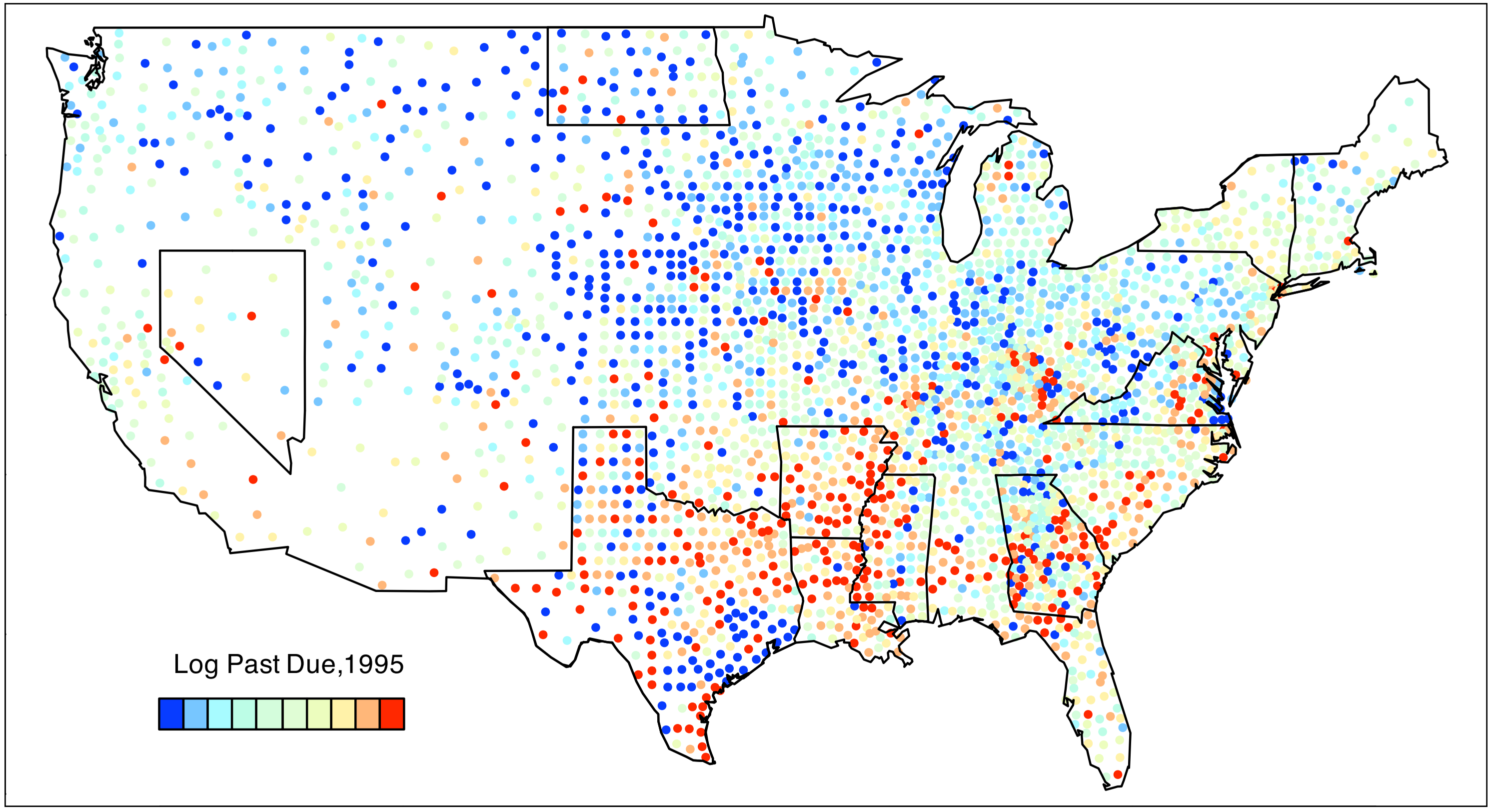
Geographical Patterns

Concentrations of high and low default



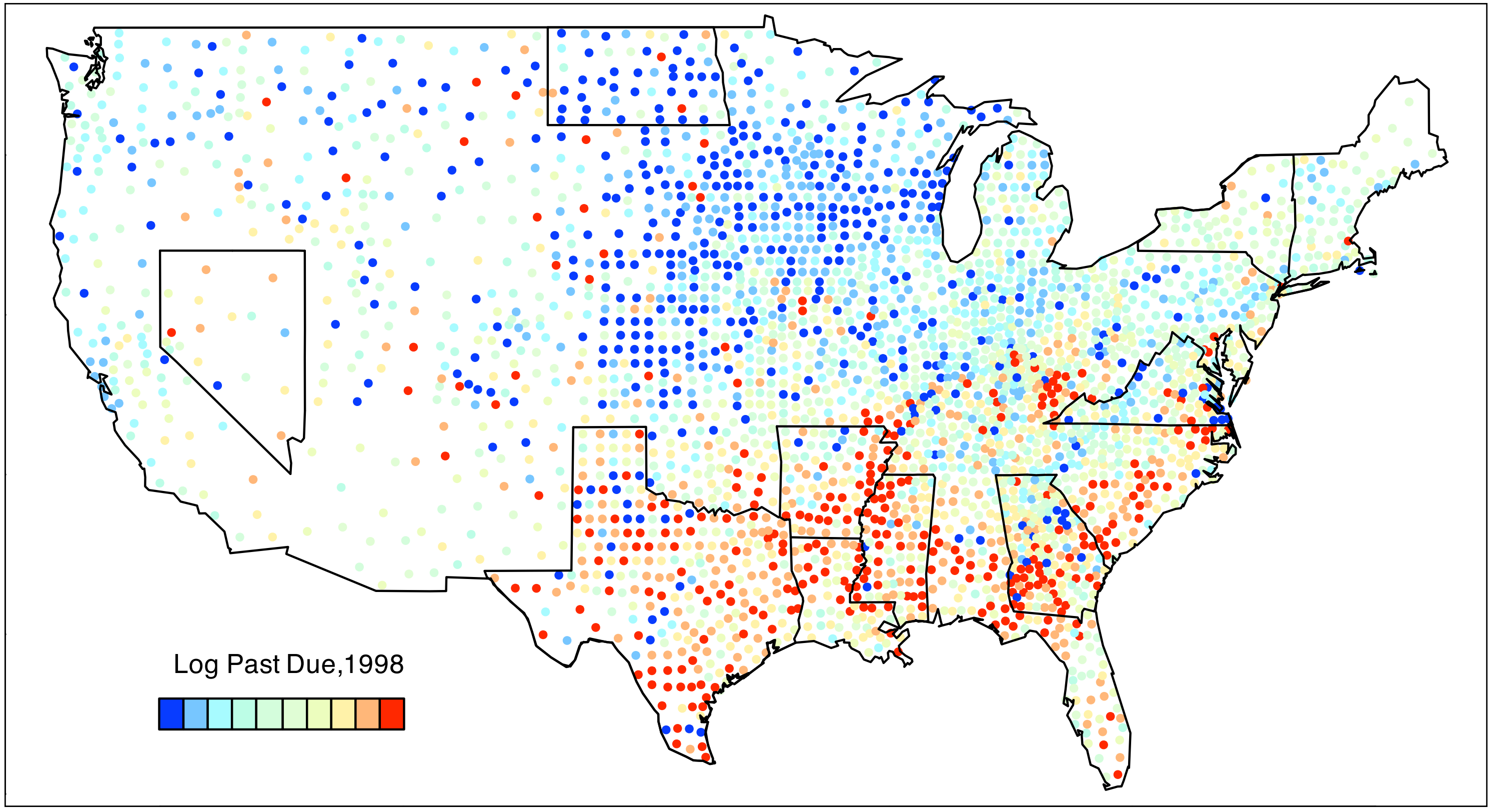
Geographical Patterns

Concentrations of high and low default



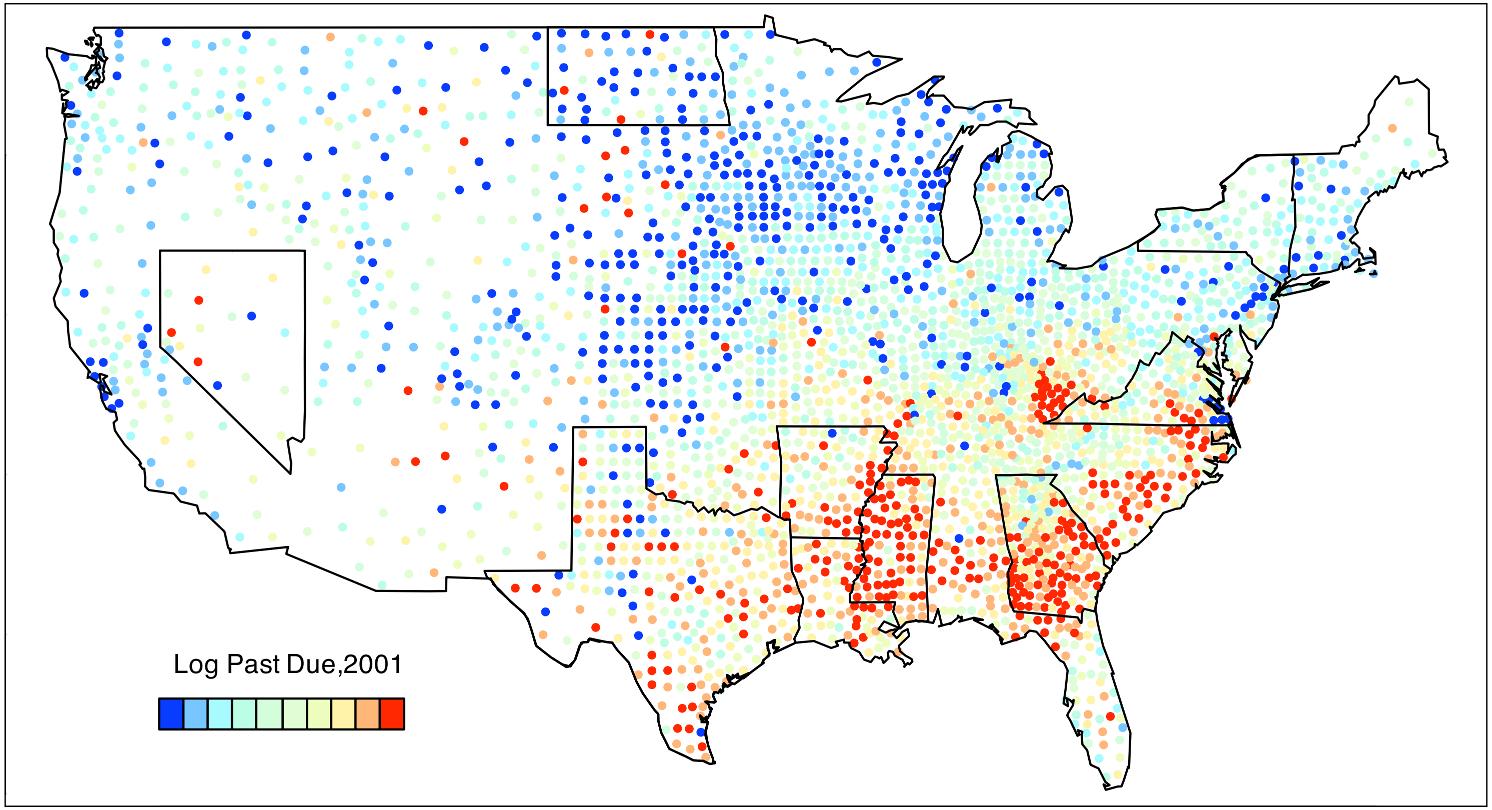
Geographical Patterns

Concentrations of high and low default



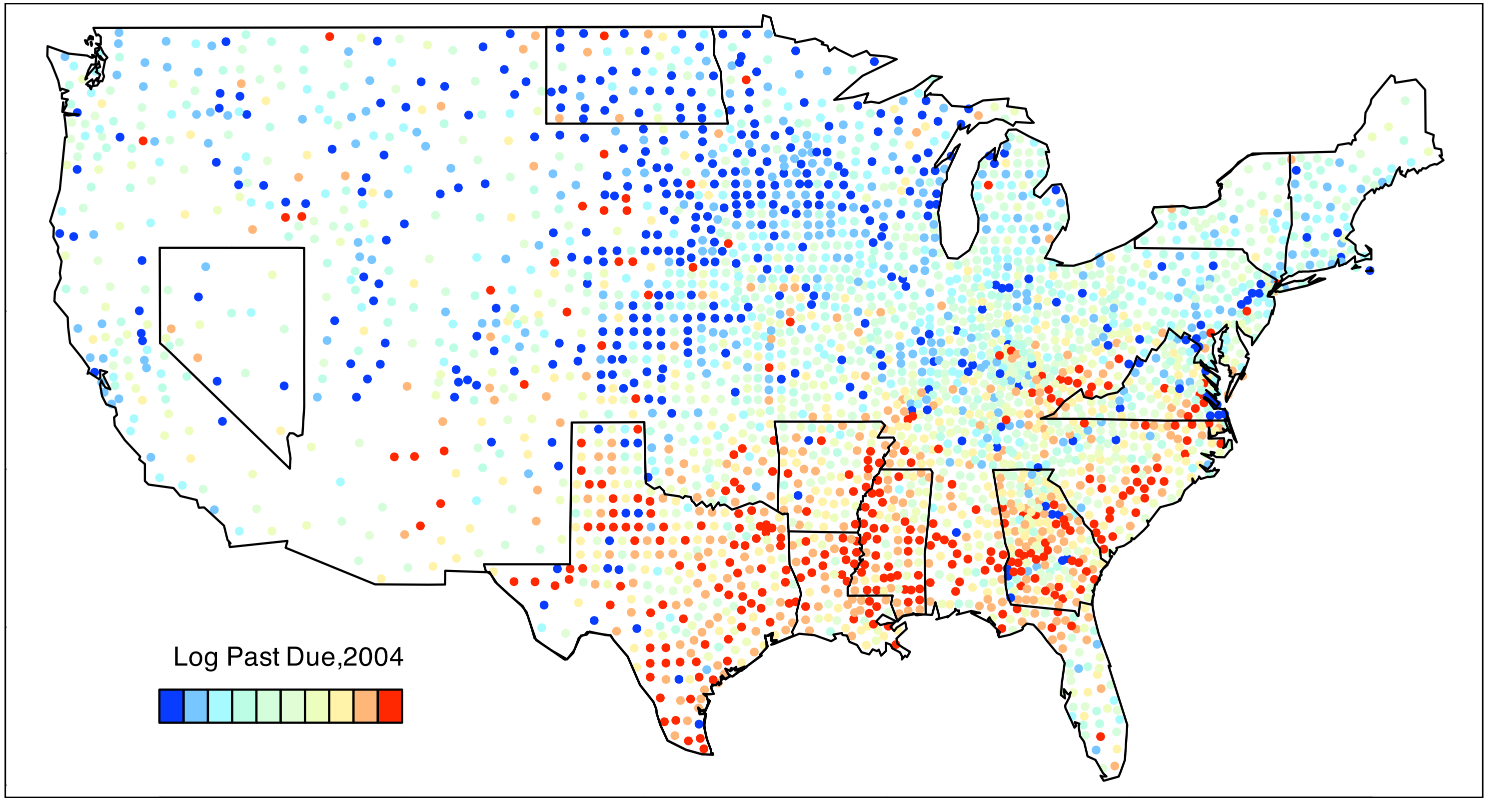
Geographical Patterns

Concentrations of high and low default



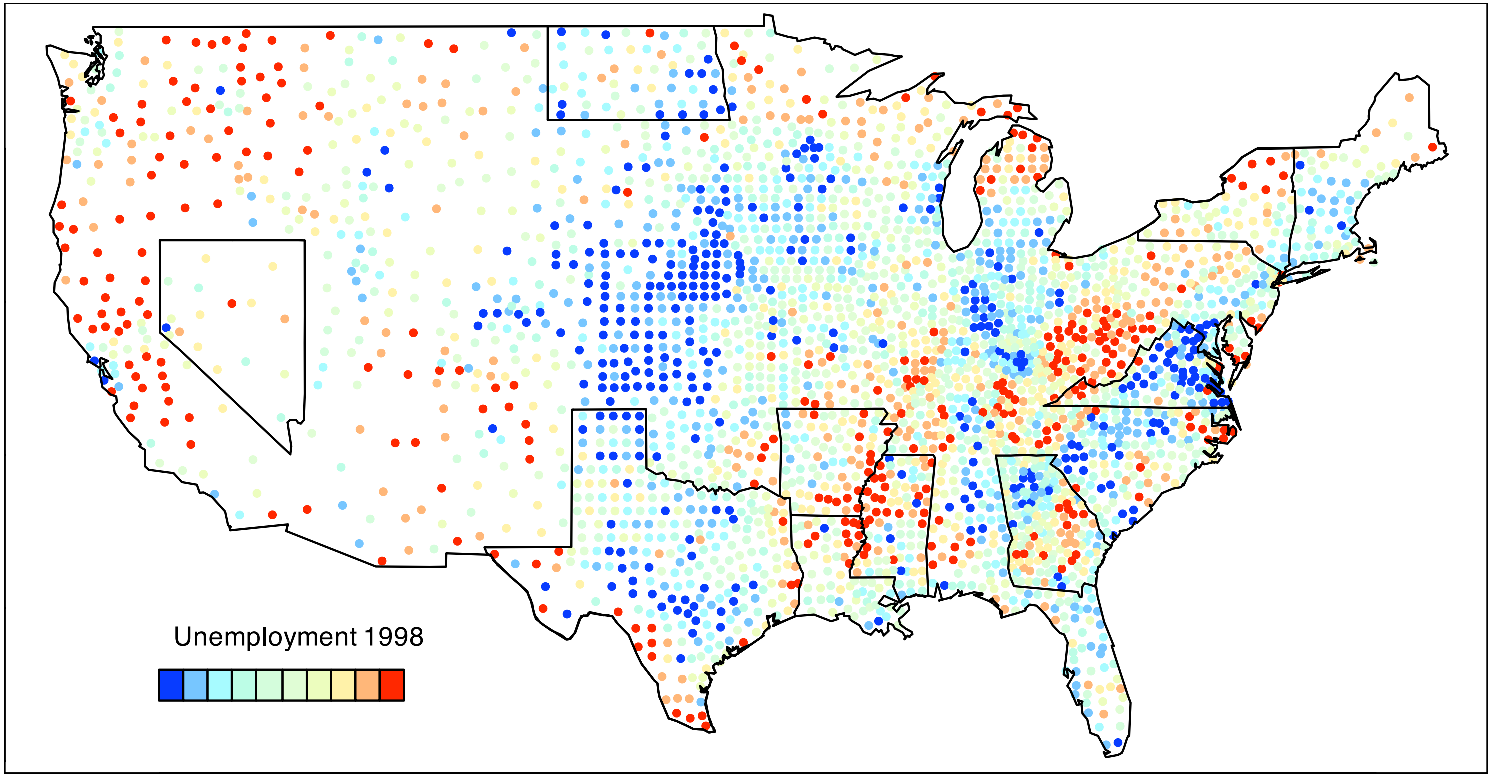
Geographical Patterns

Concentrations of high and low default



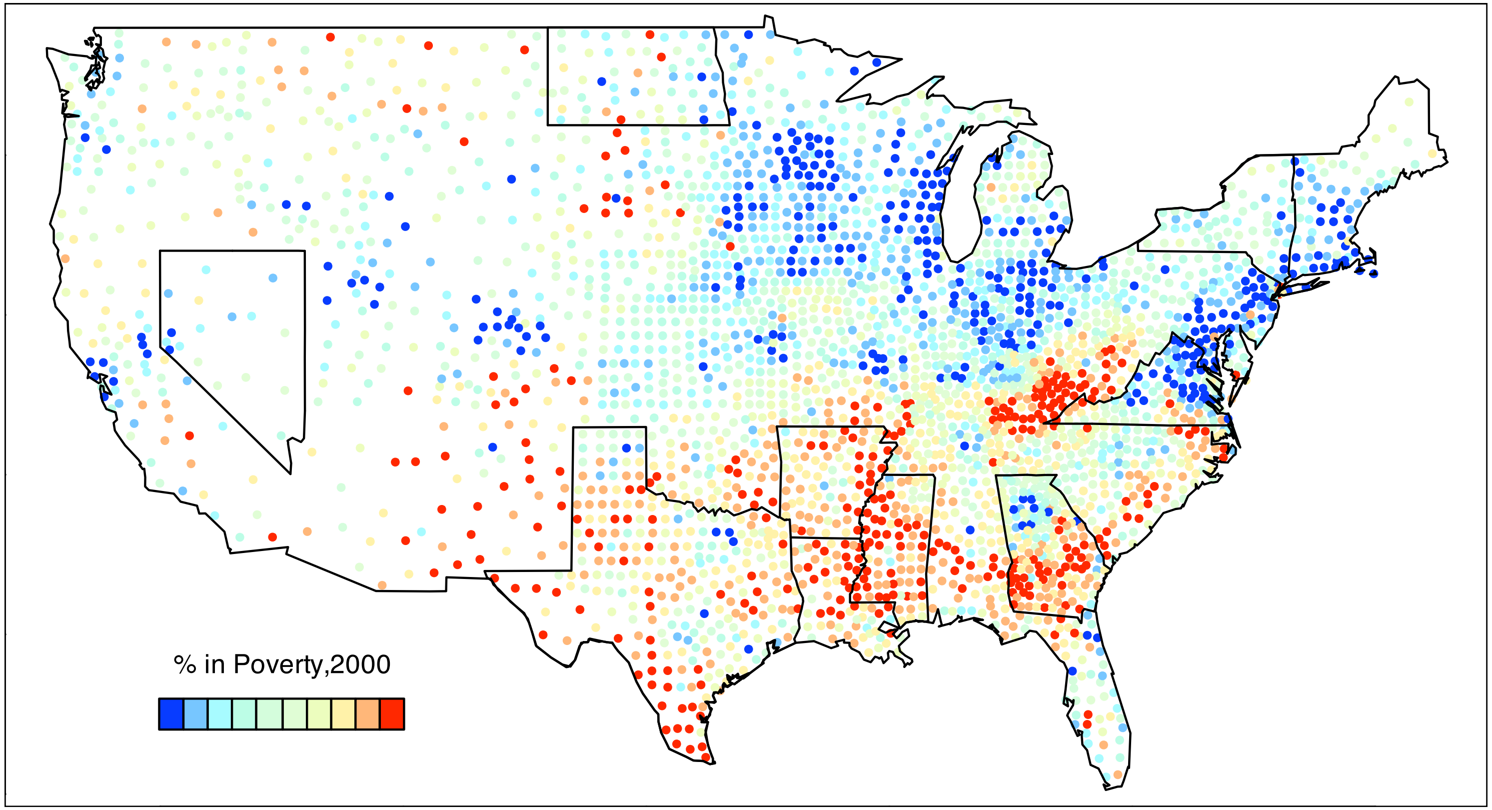
Unemployment

- Substantial regional concentrations in Mississippi valley, Appalachia, West



Geographical Patterns

Poverty also concentrated in southeastern US



Return to Questions

1. Is there adequate spatial variation to support modeling credit risk?

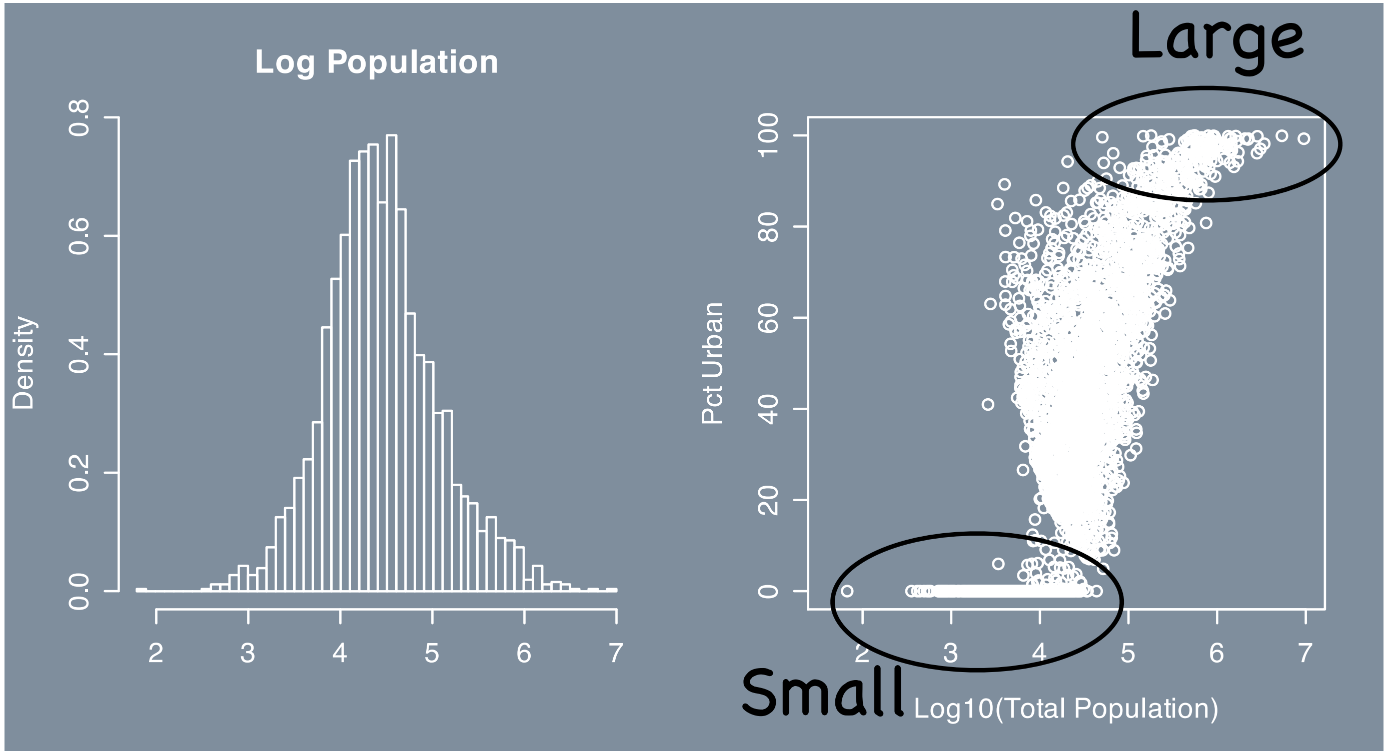
   Yes. Maybe too much!

2. Do local macroeconomic variables add value beyond usual bank information?

3. Do models suffer under local economic stress?

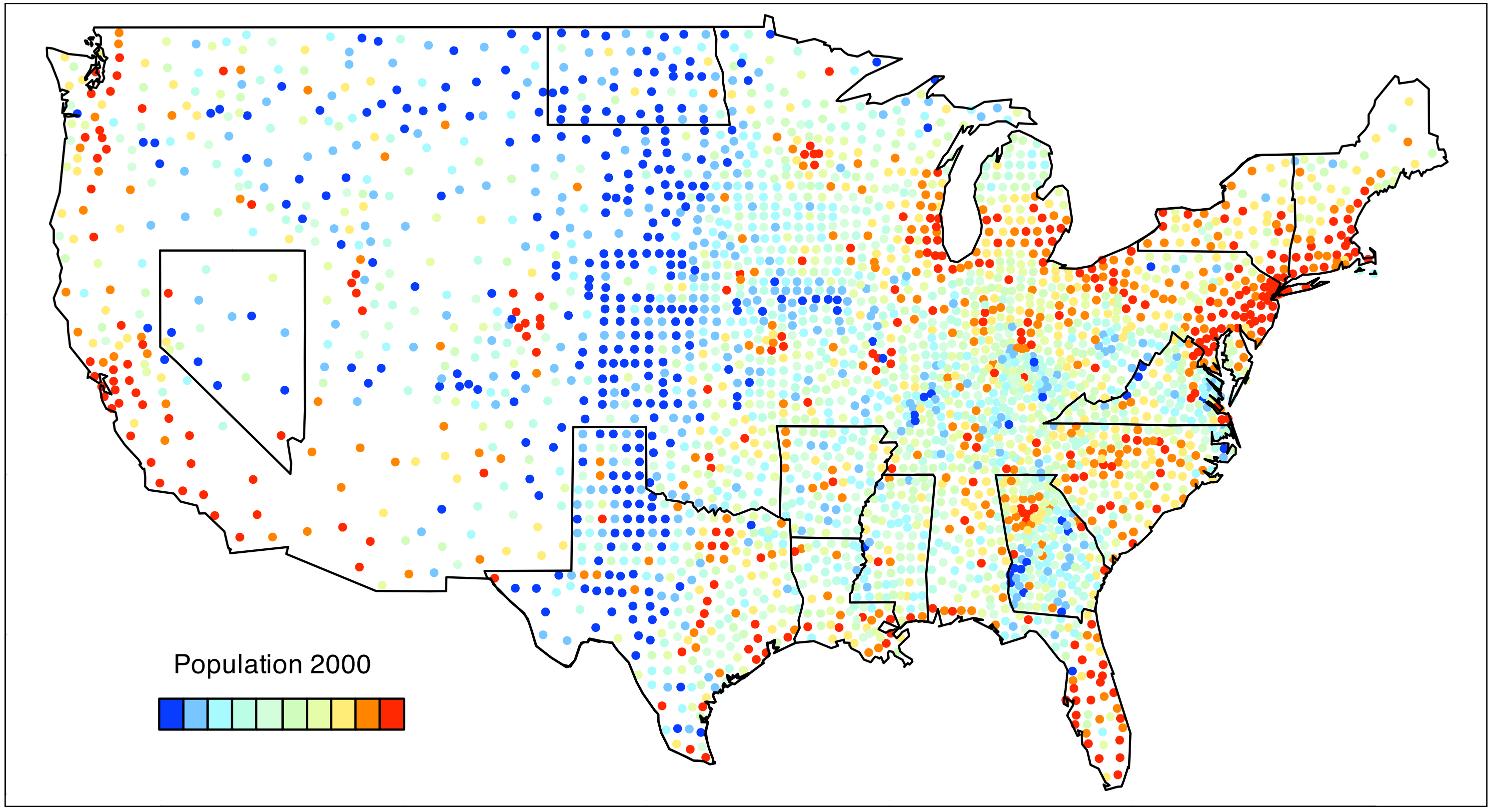
Variation in Population

- Skewed, ranging from 67 to 9.5 million
  - Log transformation brings rough symmetry
- Further confounding
  - Smallest are rural   Largest are urban



Spatial Clustering

- Concentrations of high population
  - Evident urban clusters
  - Confounding: geographic location and population



Models

Models

- Predict percentage late payments
  - Log scale, one point in time
- Baseline model ignores covariates
  - $\text{Log}(\text{Late}_t) = b_{0,t} + b_{1,t}\ \text{Log}(\text{Late}_{t-1}) + e_t$
- More complex models include bank variables plus macroeconomic variables
  - Add lagged covariates of several types

$$\text{Log}(\text{Late}_t) = b_{0,t} + b_{1,t}\ \text{Log}(\text{Late}_{t-1}) +$$

"bank" $\quad b_{2,t}\ \text{Log}(\text{Util}_{t-1}) + b_{3,t}\text{Log}(\text{Cards}_{t-1}) +$

"macro" $\quad b_{4,t}\ \text{Log}(\text{Un}_{t-1,\ t-2,\ t-3,\ t-4}) + b_{5,t}\ \text{Log}(\text{Pov}) + e_t$

Concern

$$\text{Log}(\text{Late}_t) = b_{0,t} + b_{1,t}\,\text{Log}(\text{Late}_{t-1}) +$$
$$b_{2,t}\,\text{Log}(\text{Util}_{t-1}) + b_{3,t}\text{Log}(\text{Cards}_{t-1}) +$$
$$b_{4,t}\,\text{Log}(\text{Un}_{t-1,\ t-2,\ t-3,\ t-4}) + b_{5,t}\,\text{Log}(\text{Pov}) + e_t$$

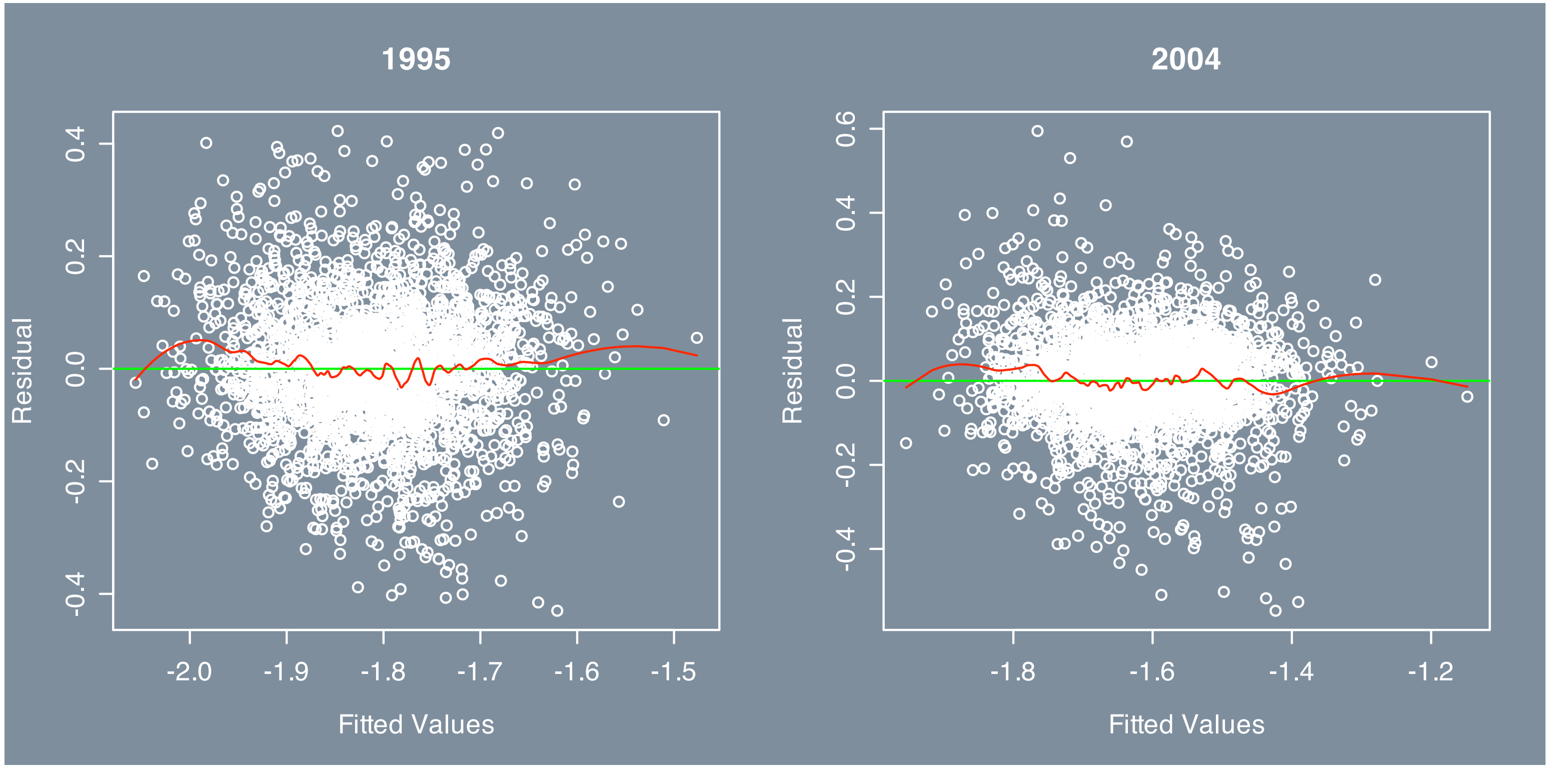
- Does the model capture the patterns evident in the spatial plots?
- Do I have 3,000 degrees of freedom for the error?
  - 3,000 df simplifies modeling since I can afford to fit a new model at each time.
  - Dependence leads to questions of the validity of claims of statistical significance.
  - Analogous to autocorrelation in time series models

Check regression diagnostics BEFORE looking at the model results...

The shown results illustrate patterns seen at other times.

Residual Plots

- Models are well-calibrated, with fitted values linearly related to response.
- Larger residuals randomly scattered

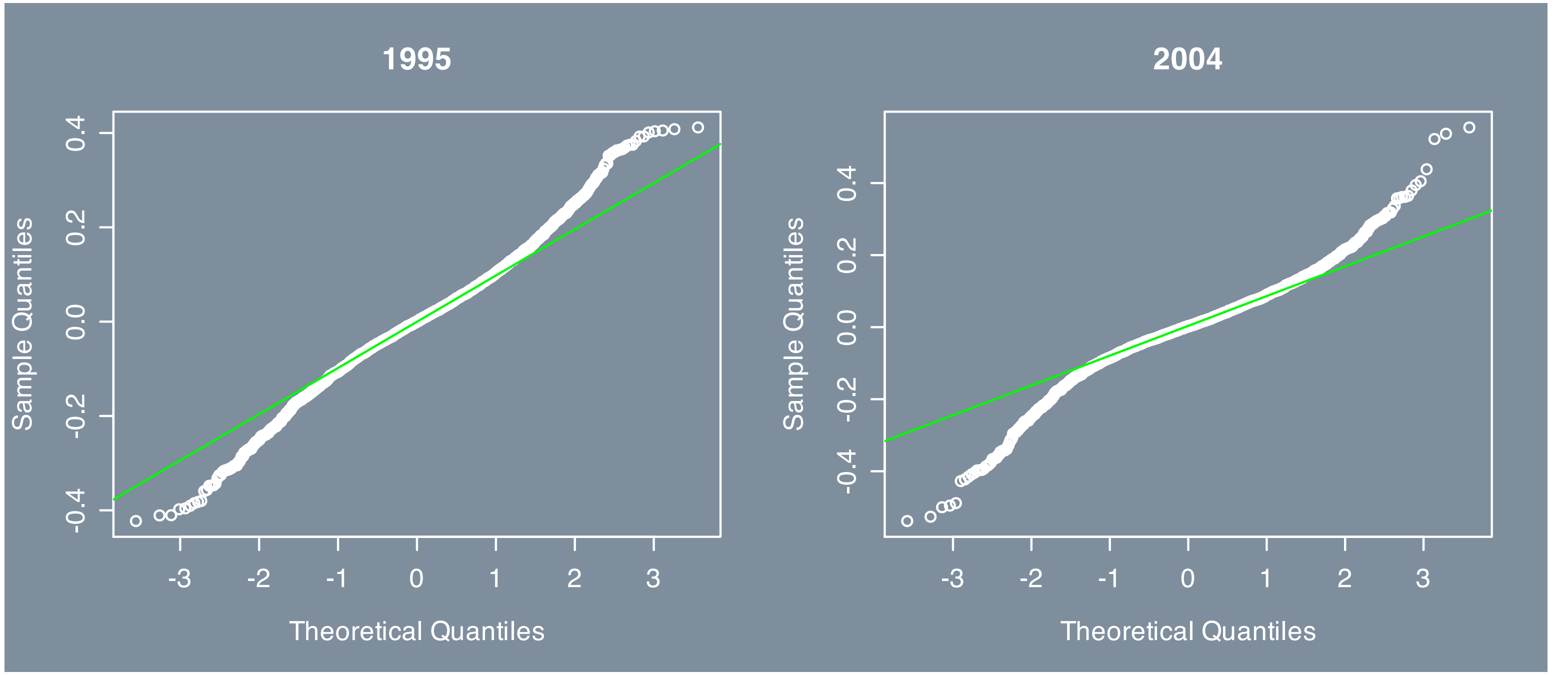


So far, so good. In addition

✓Partial residual plots
✓Partial regression (leverage) plots
✓Correlation with prior residuals

Check the distribution of the errors...

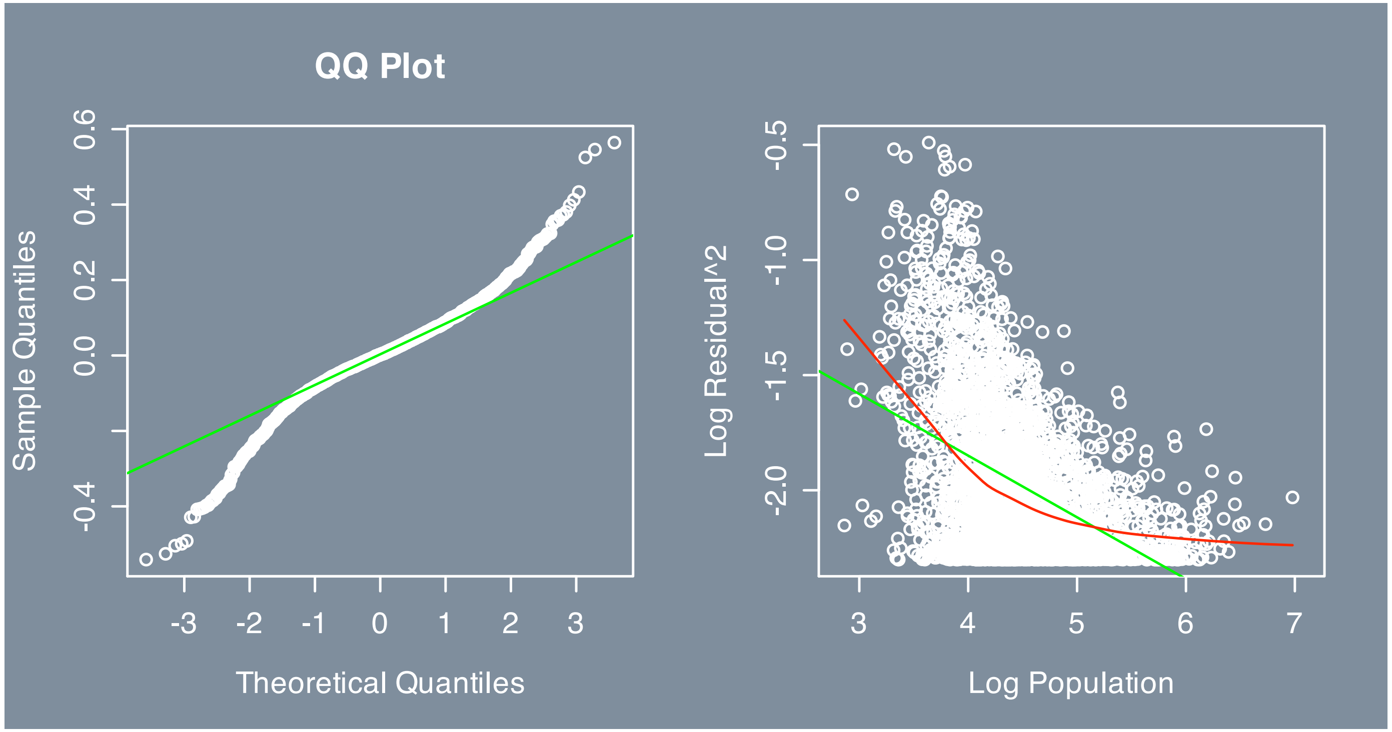
Uh-Oh!



Fat-tails: due to spatial heterogeneity?

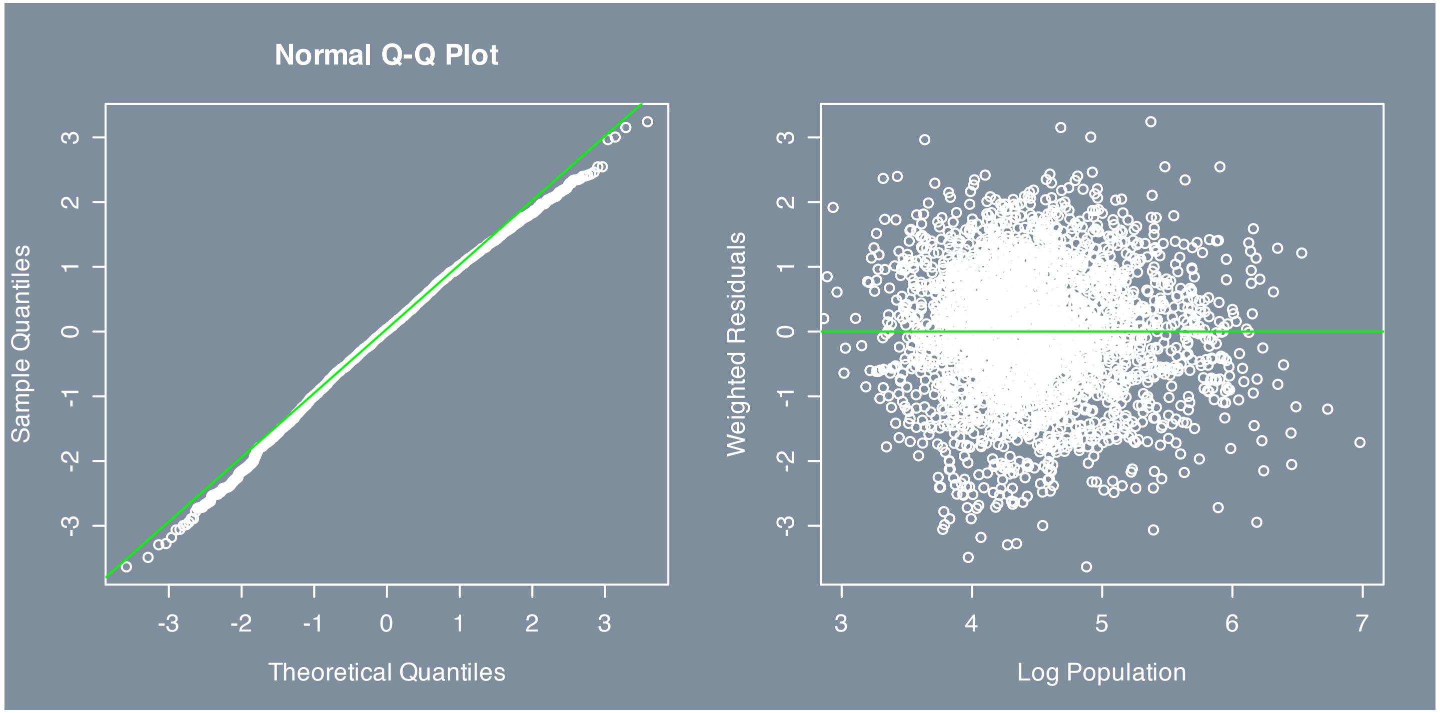
Simpler Explanation

- Residual variance related to pop size
  - $Var(e_t)$ does not fall off with population as rapidly as usual calculation would suggest



Stabilize Variance

- After weighting by the estimated variance function, residuals are much nicer.
  - Much easier than spatial adjustments



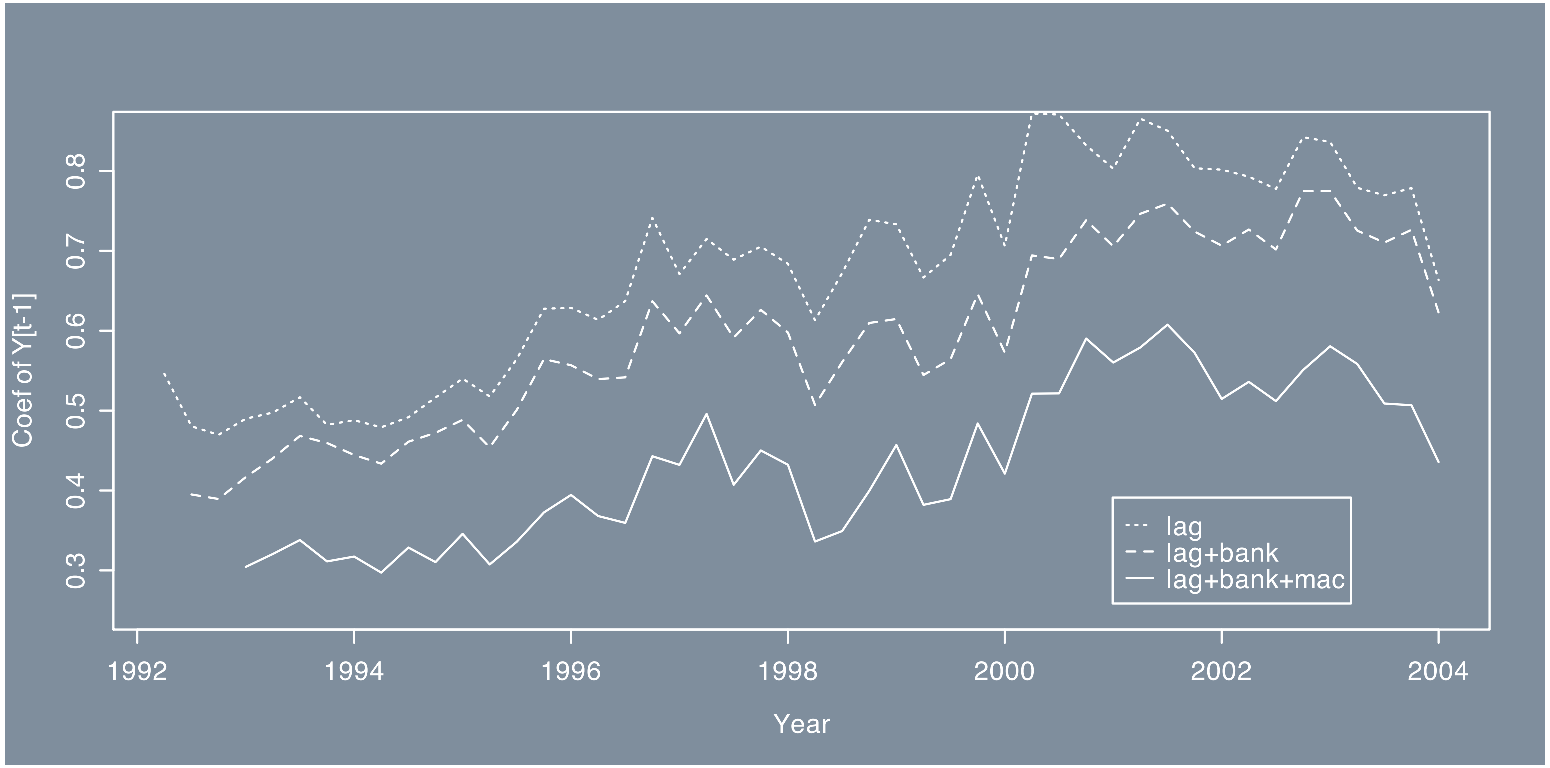
Now that we have a
reasonable model,
take a look at its
properties...

Goodness of Fit

- Fit improves over time
- Macro variables are statistically significant
- Gain worth the effort?

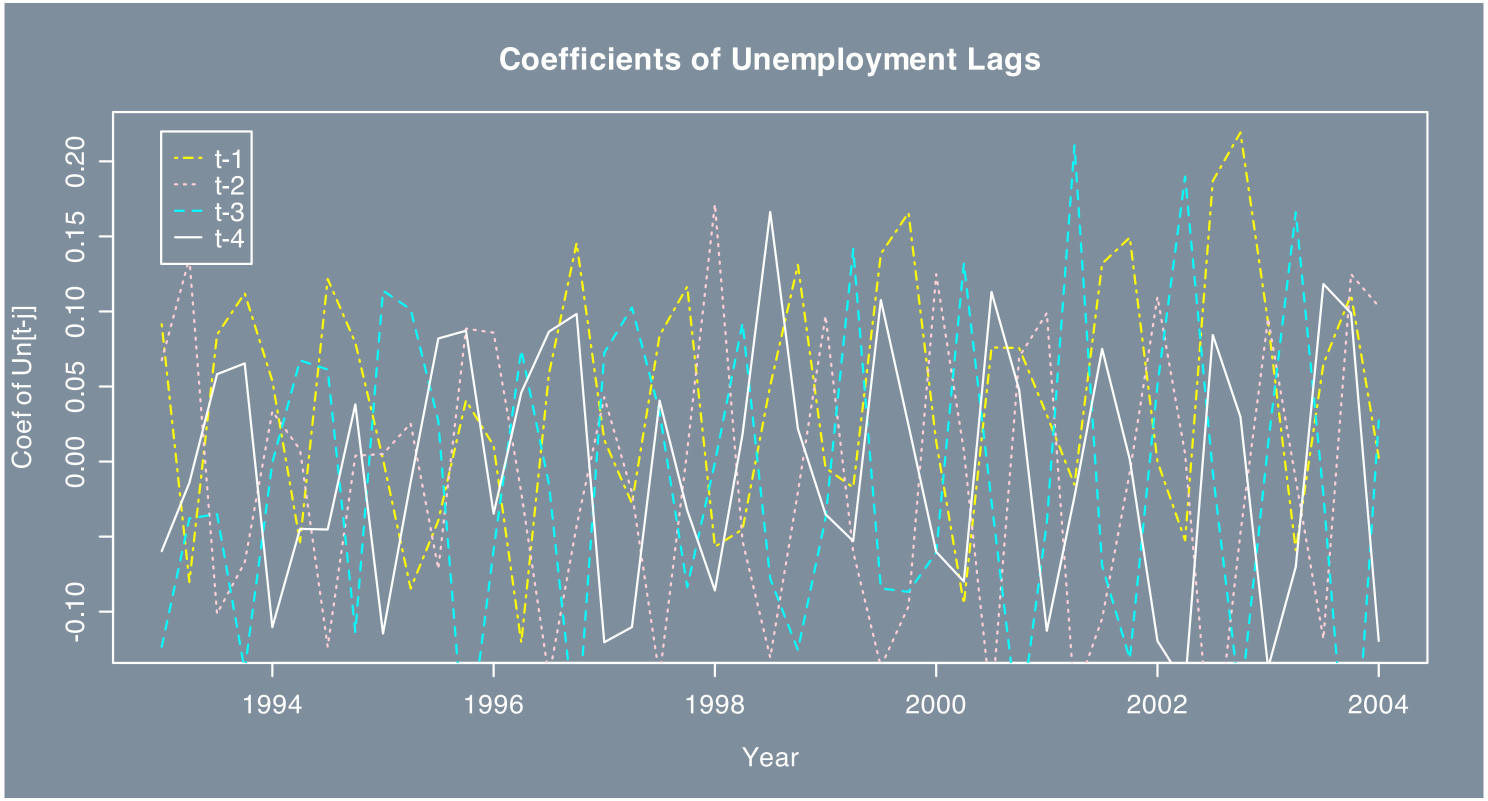
Coefficients over Time

- Estimated coefficient "drifts"
  - Size of effect of lagged endogenous grows
- Less drift when use macro variables



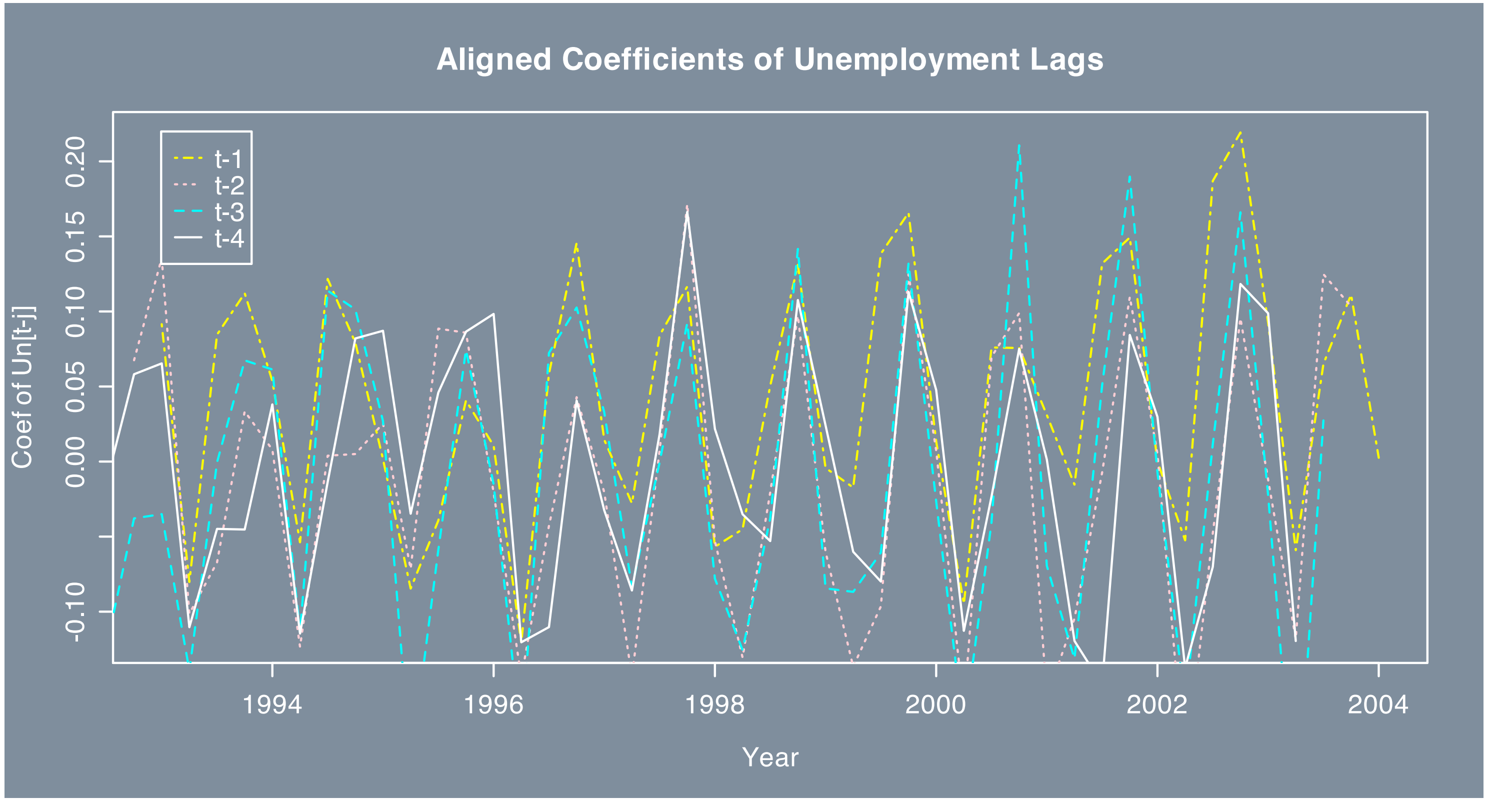
Coefficient Changes

- Estimated coefficients for lags of unemployment fluctuate wildly
- Noise?



Seasonal Structure

- Align the coefficients to fixed point in time
- Byproduct of underlying data?
  - Recall seasonal oscillation in unemployment



Borrowing Strength?

- I estimated the model M_t with data for a specific quarter, ignoring models in prior quarters M_{t-1}, M_{t-2}, ...
- Extension
  Smooth the models by capturing the dynamics of the drifting estimates and goodness of fit.
- Caveat
  Must capture seasonal effects like that in unemployment rather than smooth over. Smoothing forces similarity.

Questions

1. Is there adequate spatial variation to support modeling credit risk?

Yes. Maybe too much!

2. Do local macroeconomic variables add value beyond usual bank information?

Yes. The gain is small (albeit significant) and these variables stabilize the model structure

3. Do models suffer under local economic stress?

Spatial Effects

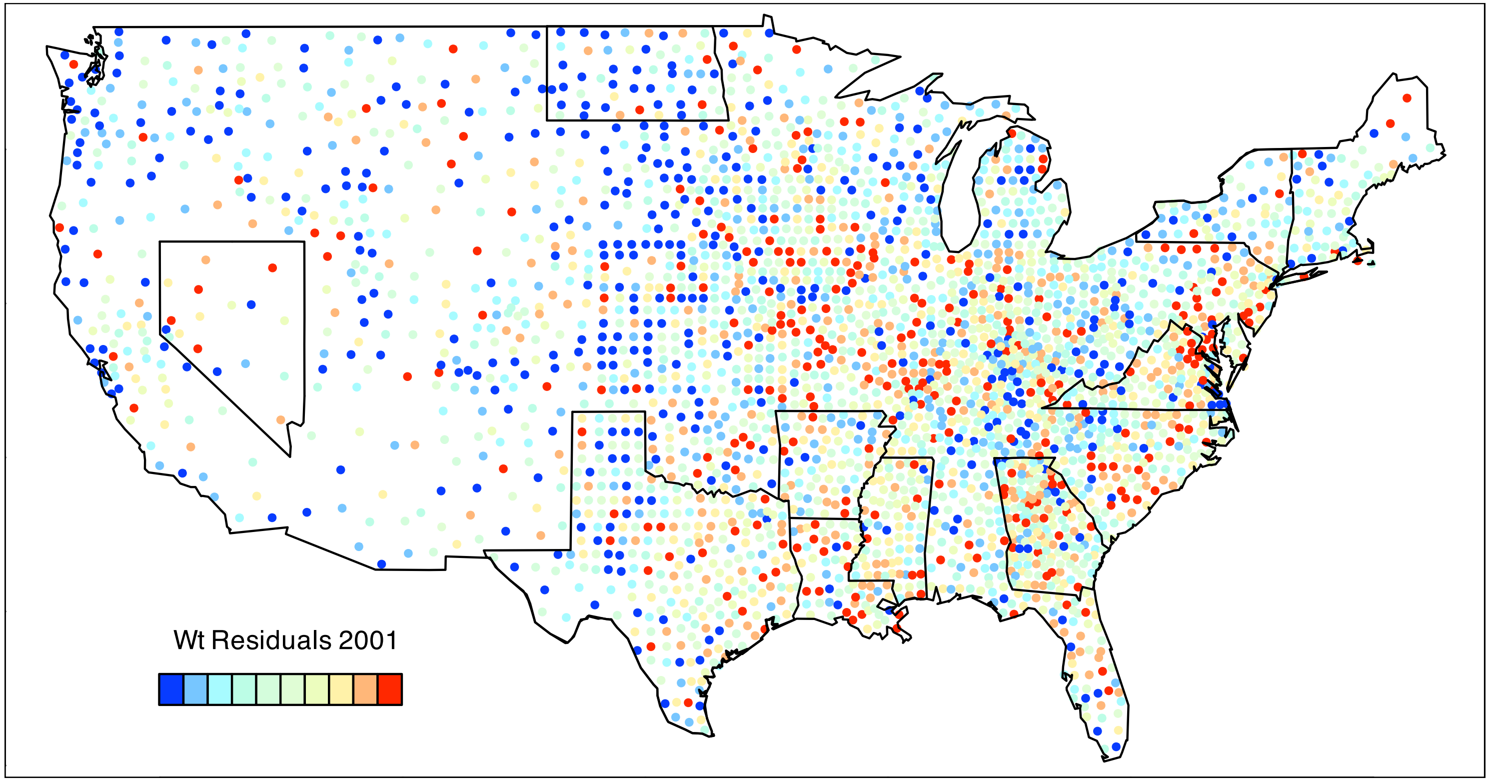
Plan

- Removed the heterogeneity due to population size, but still need to check for spatial dependence among residuals.

- Methods
  - Descriptive: variogram
  - Markovian: Markov random field

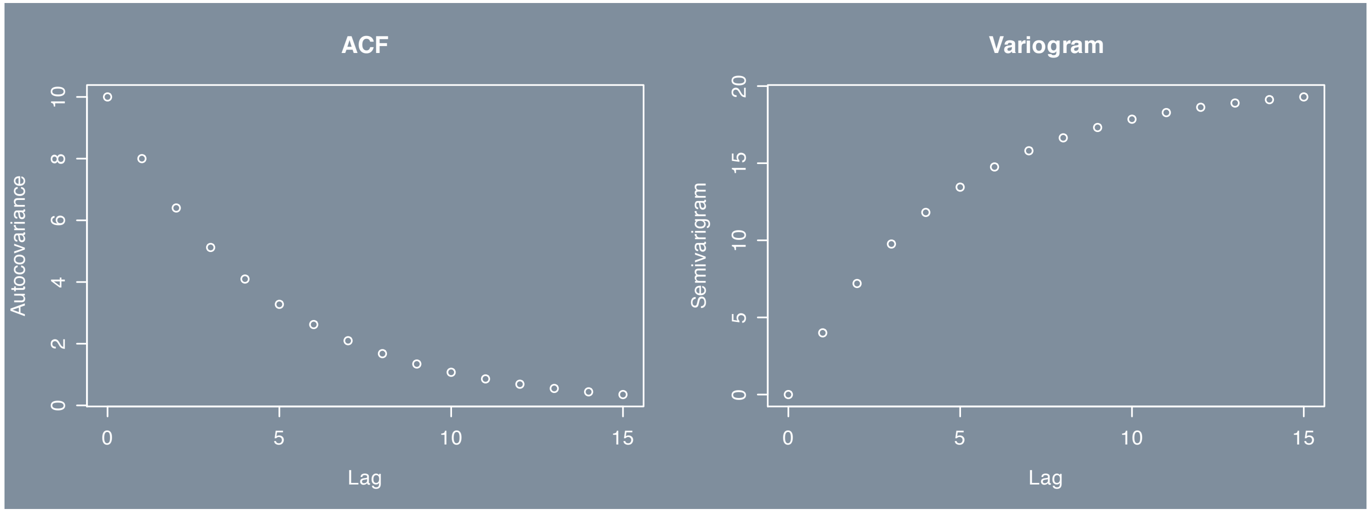
Spatial View of Residuals

Substantial clustering or natural variation?



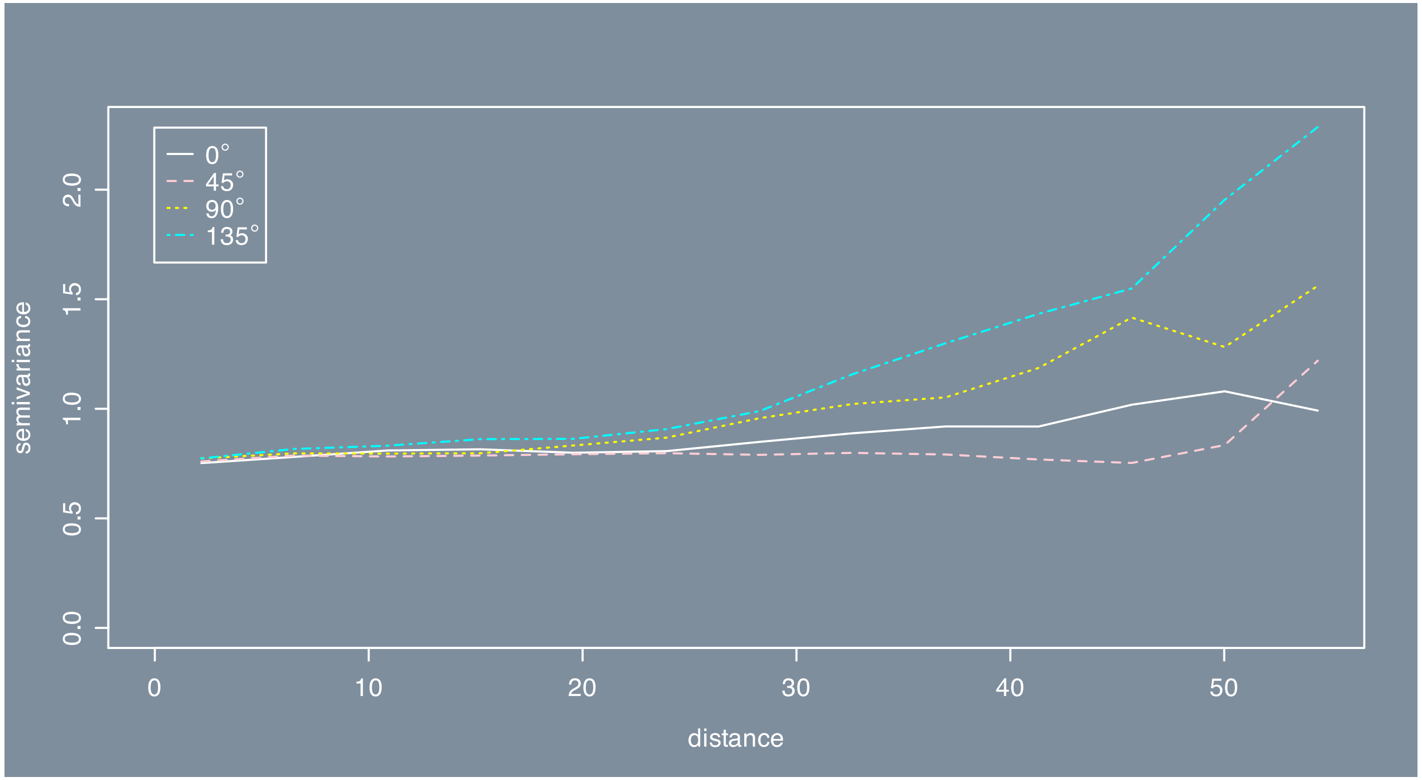
Variogram

- Alternative to the familiar autocovariance used in time series analysis
  - $ACF(j) = Cov(Y_t, Y_{t-j})$
  - $VG(d) = Var(Y_t - Y_s)$ for $|t-s| - d$
- Connection
  - $VG(d)/2 = Var(Y) - Cov(Y_t, Y_s)$
    $= ACF(0) - ACF(t-s)$



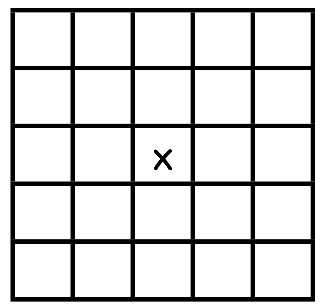
Directional Variogram

- Plot $Var(e_u - e_w)$ vs distance $|u-w|$ (units?)
- Define neighborhoods along 4 angles
- Little dependence except NW to SE



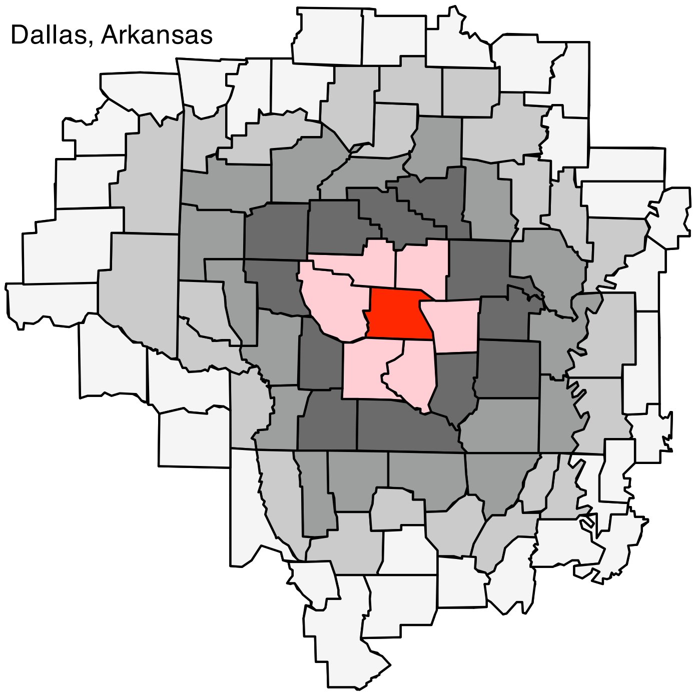
Spatial Models

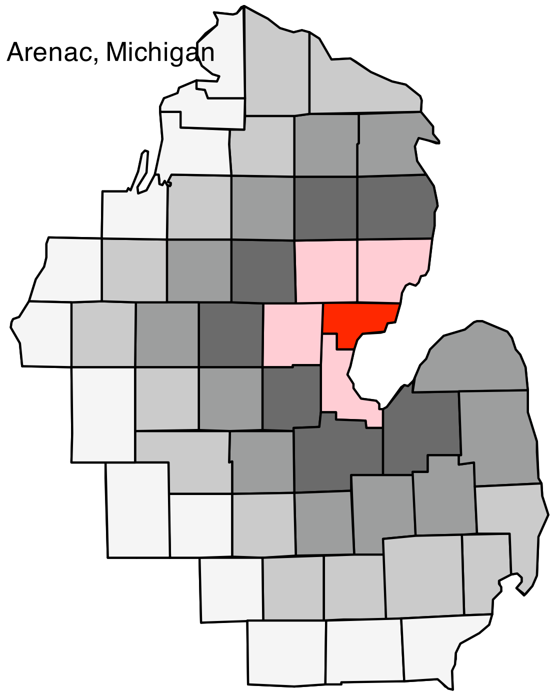
- Broad class
  - More models than those available in time series analysis because the dependence is not naturally "one-sided"
    - e.g., two types of first order autoregressions
    - 2-D plane only partially ordered
- Neighborhood
  - Which observations are "close"?
  - Easily defined in computer graphics because data live on well-defined grid
  - Less clear for map-based geographical units



Neighborhoods

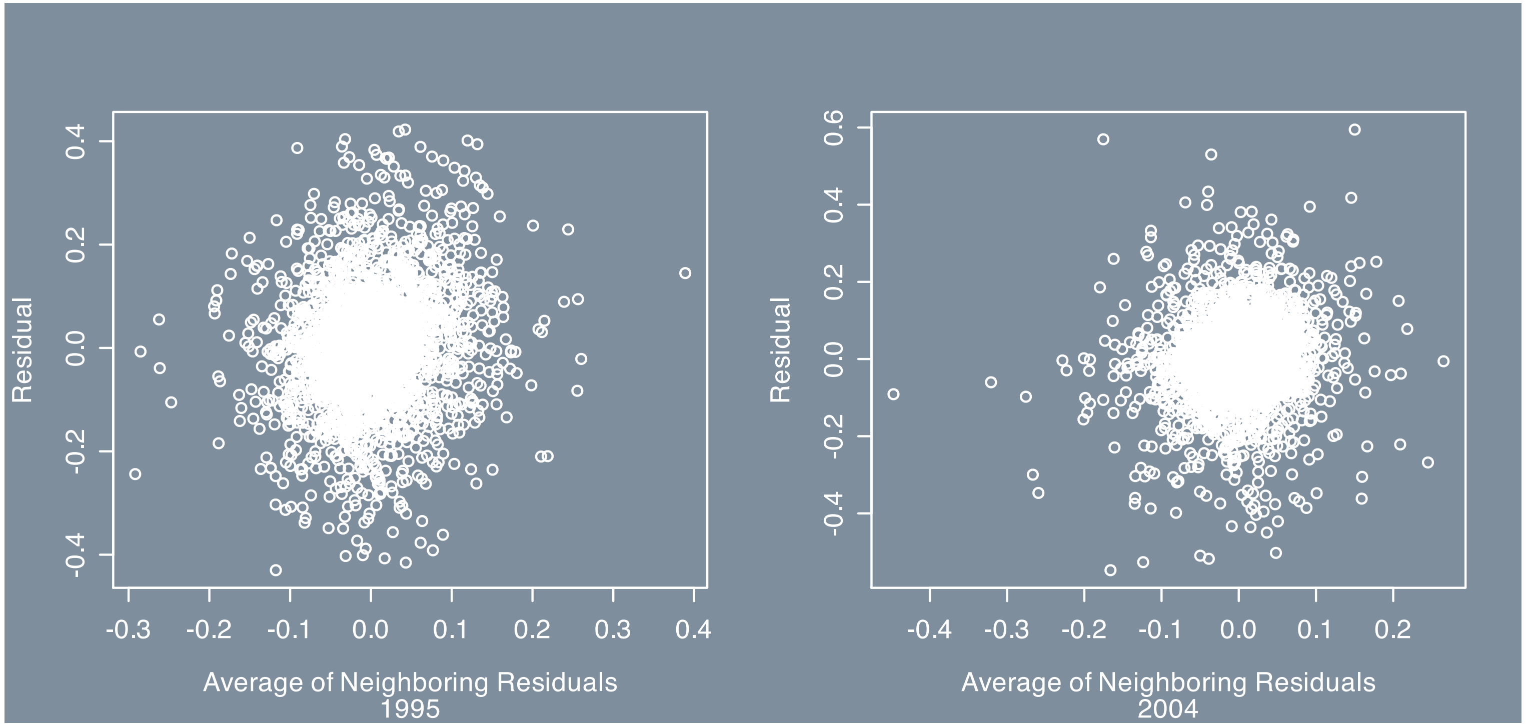
- Model dependence using adjacent counties
- Layers define neighborhoods
  - CAR model assumes P(Y|all) = P(Y|neighbors)
  - Conditional on neighbors, independent of others





Link to Neighbors

- Relationship of residuals to average residual over neighboring counties
- No evidence of dependence (r ≈ 0.15)



Where are we?

- Model fit to data for one quarter works well in sense that it
  - explains substantial variation
  - uses natural predictors
  - produces random unexplained variation

But...

- Who's to say that the structure of the model itself should be homogeneous over space?
  - After all, the model parameters drift over time. Why should it be the same over regions?

Spatial Locations

- Spatially local fits require "small" samples
  - Want small enough area so that do not lose ability to localize spatial properties
  - Cannot spread over time since we know these models change over time
- Sample 300 spatially separated points around the US
  - None is adjacent to another
  - Use 5 "layers" to define each neighborhood
  - Each neighborhood has 50 to 100+ counties

Finding Spatial Deviations

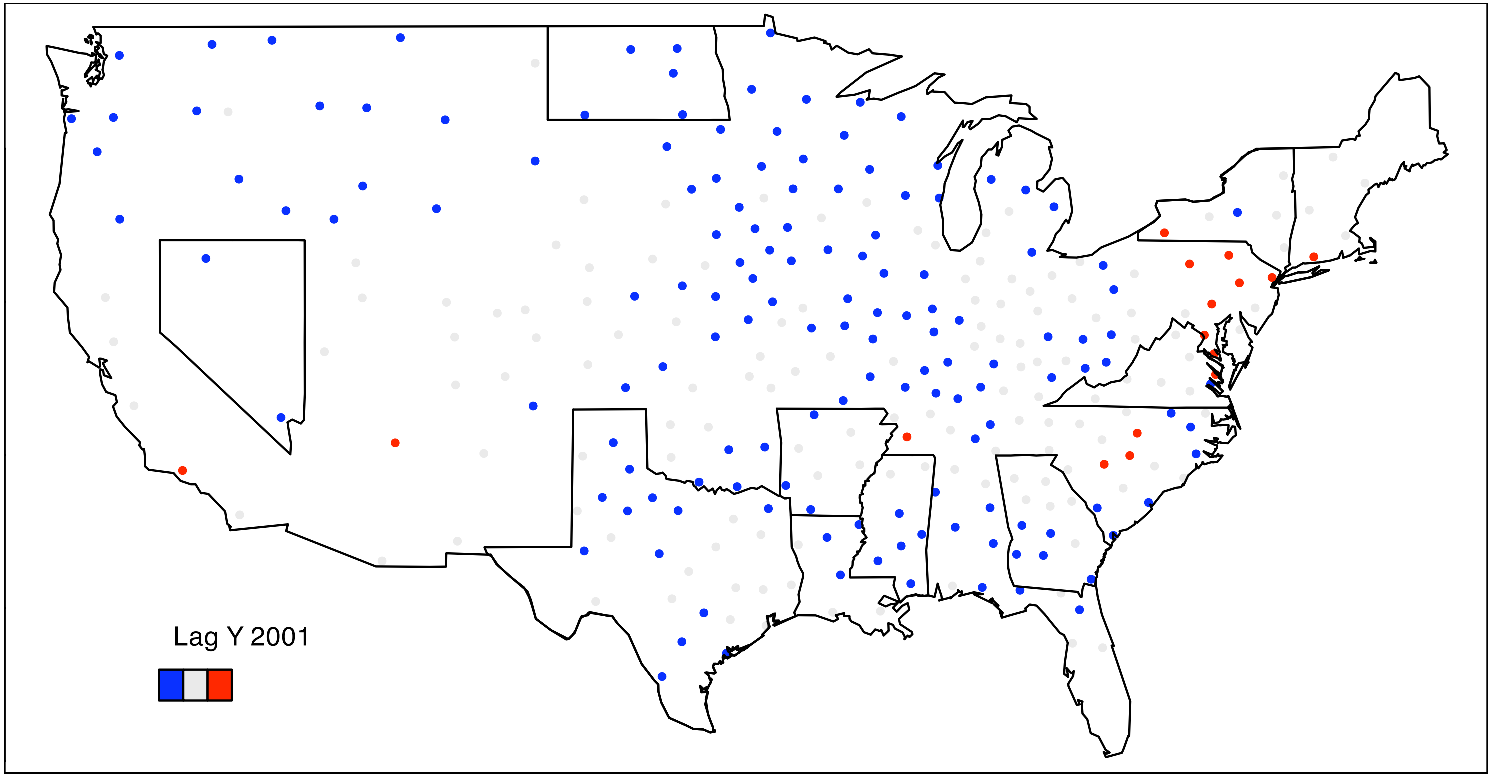
- Estimate deviations from overall estimates
  - Replace Y_t by e_t as the response
- Only interested in meaningful deviations from the overall fit
  - 300 fits offer many chances for accidentally estimating large deviations from overall WLS fit
  - Akin to over-fitting in variable selection
- Solution
  Use an approach that avoids the problem in variable selection: shrinkage
  - Ridge (ie, Bayes estimator under normal prior)
  - Adaptive estimator (Polyshrink)

Shrinkage

- Estimates of deviations from the overall coefficient of the lagged response

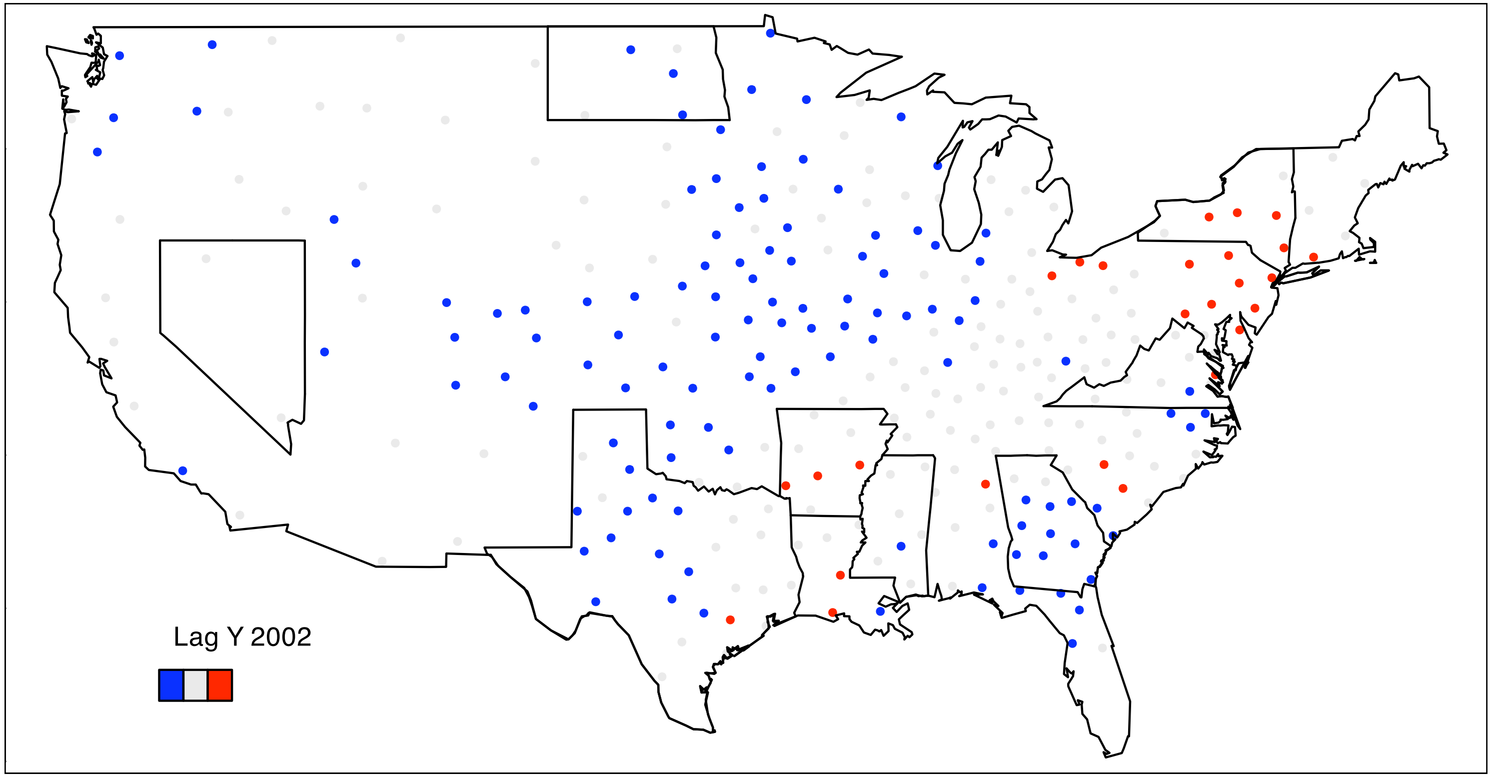
Deviations from Overall

- Polyshrink estimates of the coefficient of the lagged endogenous variable, 2001 Q1



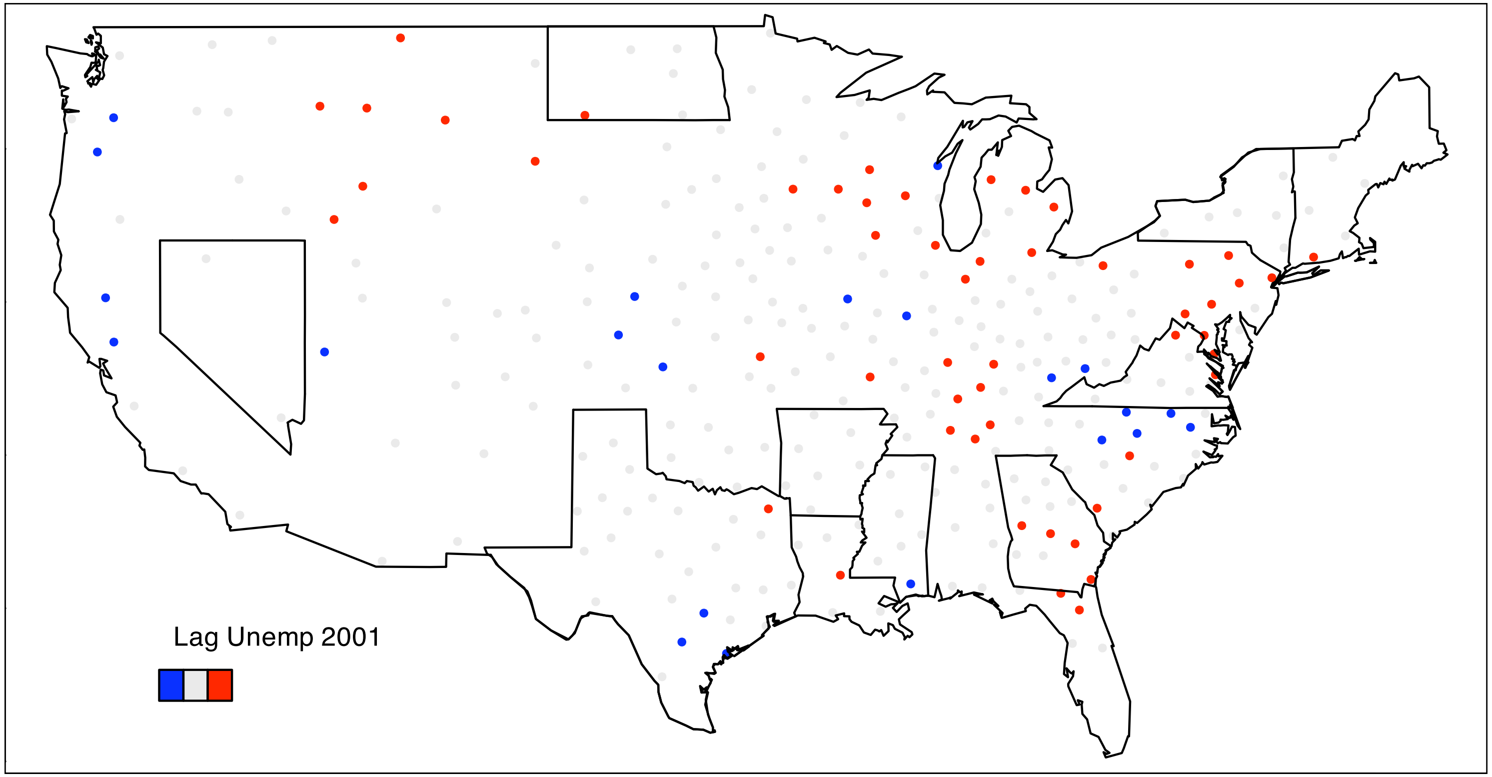
Deviations from Overall

- Polyshrink estimates of the coefficient of the lagged endogenous variable, 2002 Q1



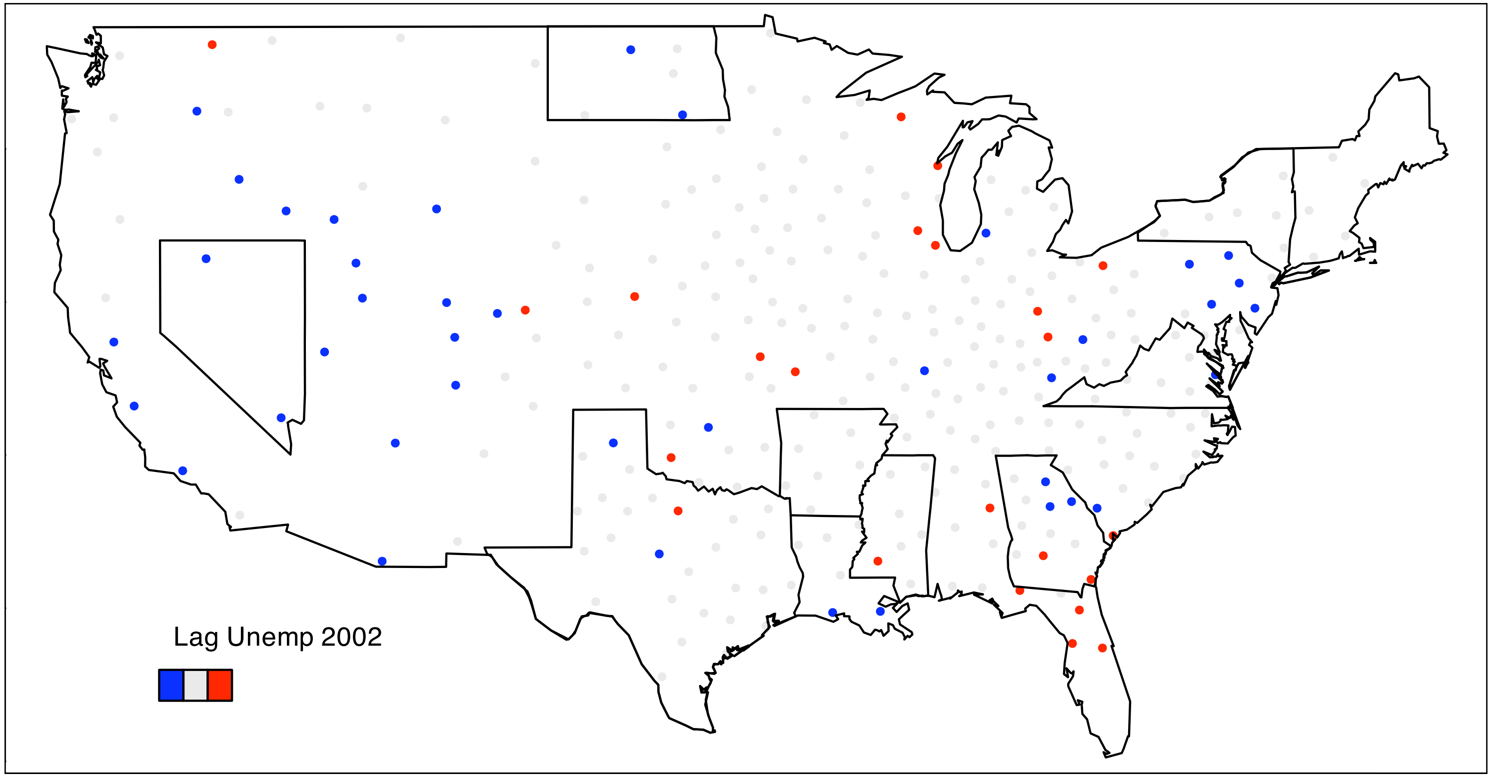
Deviations from Overall

- Locations of the selected "seed" counties
- Find more increases than with lagged y



Deviations from Overall

- Locations of the selected "seed" counties
- Less evident clustering or flow



Questions

1. Is there adequate spatial variation to support modeling credit risk?

Yes. Maybe too much!

2. Do local macroeconomic variables add value beyond usual bank information?

Yes. The gain is small (albeit significant) and these variables stabilize the model structure

3. Do models suffer under local economic stress?

Models certainly vary spatially as well as over time. Explanations?

What next?

- Incremental
  - Multivariate response (mortgage, installment)
  - More population/demographic information
  - State-level aggregation
- Modeling
  - "Global" model that describes the evolution of parameters over time and spatial clusters.
- Methods
  - Hierarchical Bayes?
    Accommodate 3,000 counties over 50 periods?

Thanks!