Spatial Temporal Models for Retail Credit

Bob Stine
Department of Statistics
The Wharton School, University of Pennsylvania
stat.wharton.upenn.edu/~stine

Outline

- Introduction
- Exploratory analysis
  - Trends and maps
- Measuring spatial association
  - Nonparametric clustering using SVDs
- Models
  - Spatial, temporal and spatio-temporal
- Next steps

- Collaborators
  - Sathyanarayan Anand
  - Federal Reserve Bank of Philadelphia

Key Points

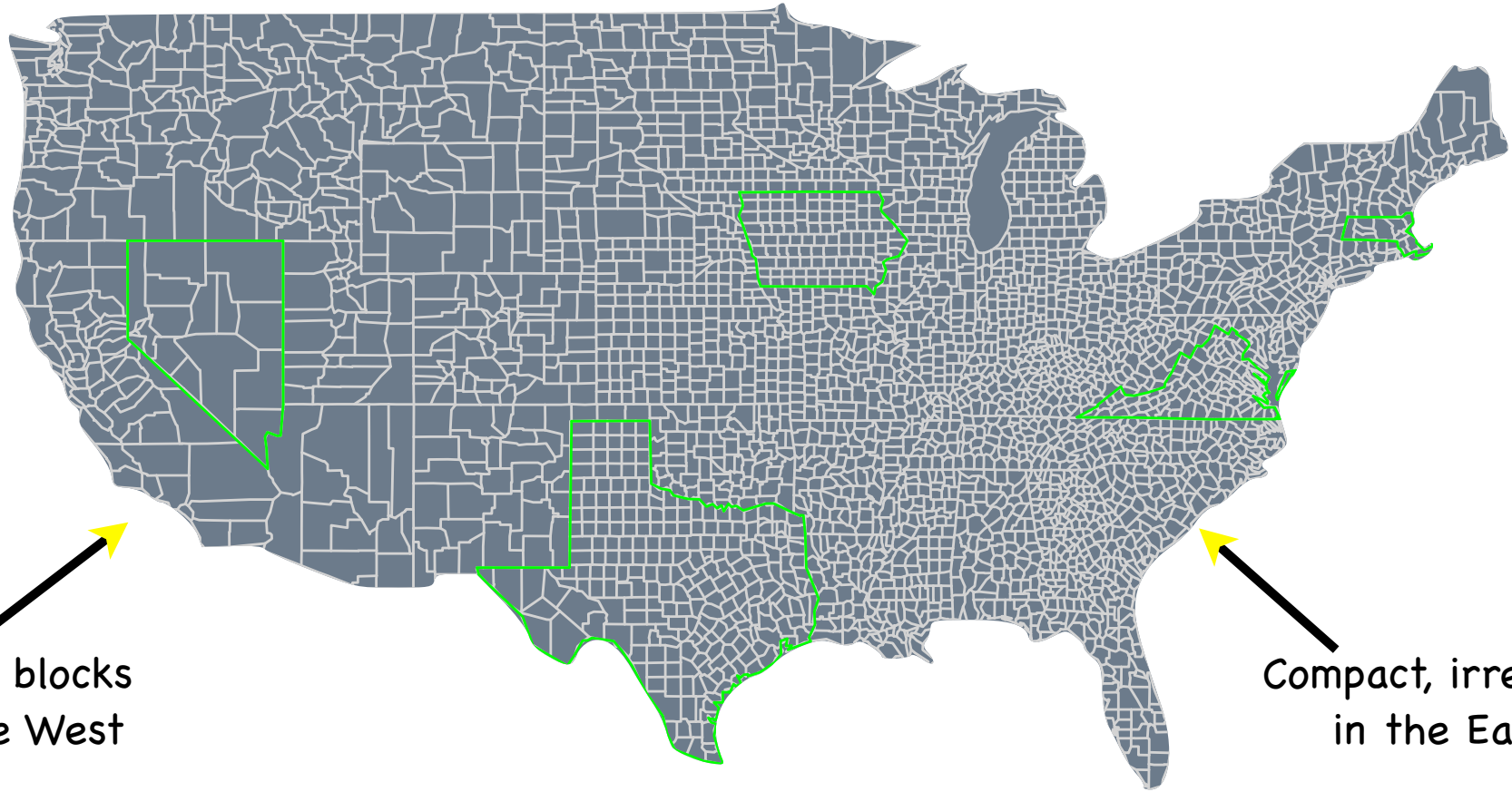
- Exploratory analysis
  - Finds spatial association in various types of default (mortgage, installment, revolving)
- Analysis of spatial patterns — Mortgages
  - Correlation risk
  - Three spatio-temporal patterns
  - Nonstationarity motivates simple models.
- Models — Cards
  - Models with broad correlations predict better than those more narrowly defined
  - Correlations in data impact claims of precision

Featured Data

- County-level
  - Default rates from Trend Data (TransUnion)
    - National coverage
    - Default rates based on quarterly samples, 1993-2010
  - Economic characteristics (Census)
  - Spatial locations
  - Small: 3,000 counties x 80 quarters = 240,000
- Multi-level inference
  - Individual -> Tract -> County -> State -> Nation
- Gaps in data...
  - Lender proprietary data (eg, vintage)
  - Individual loan characteristics
  - Housing data is incomplete

Featured Data

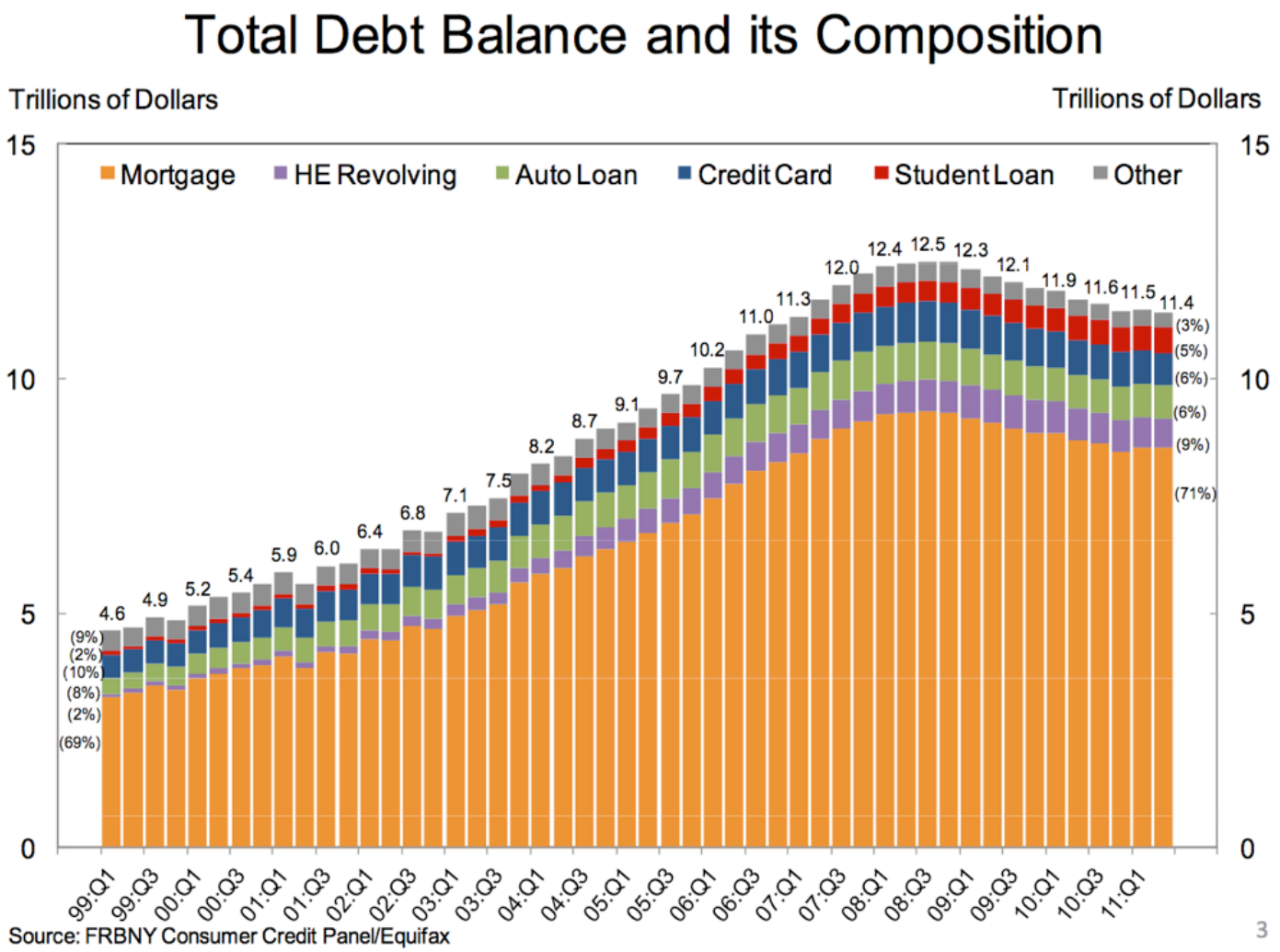
- County-level data
  - County = political subdivision of state in US
  - 3,000 counties within continental US



National Trends

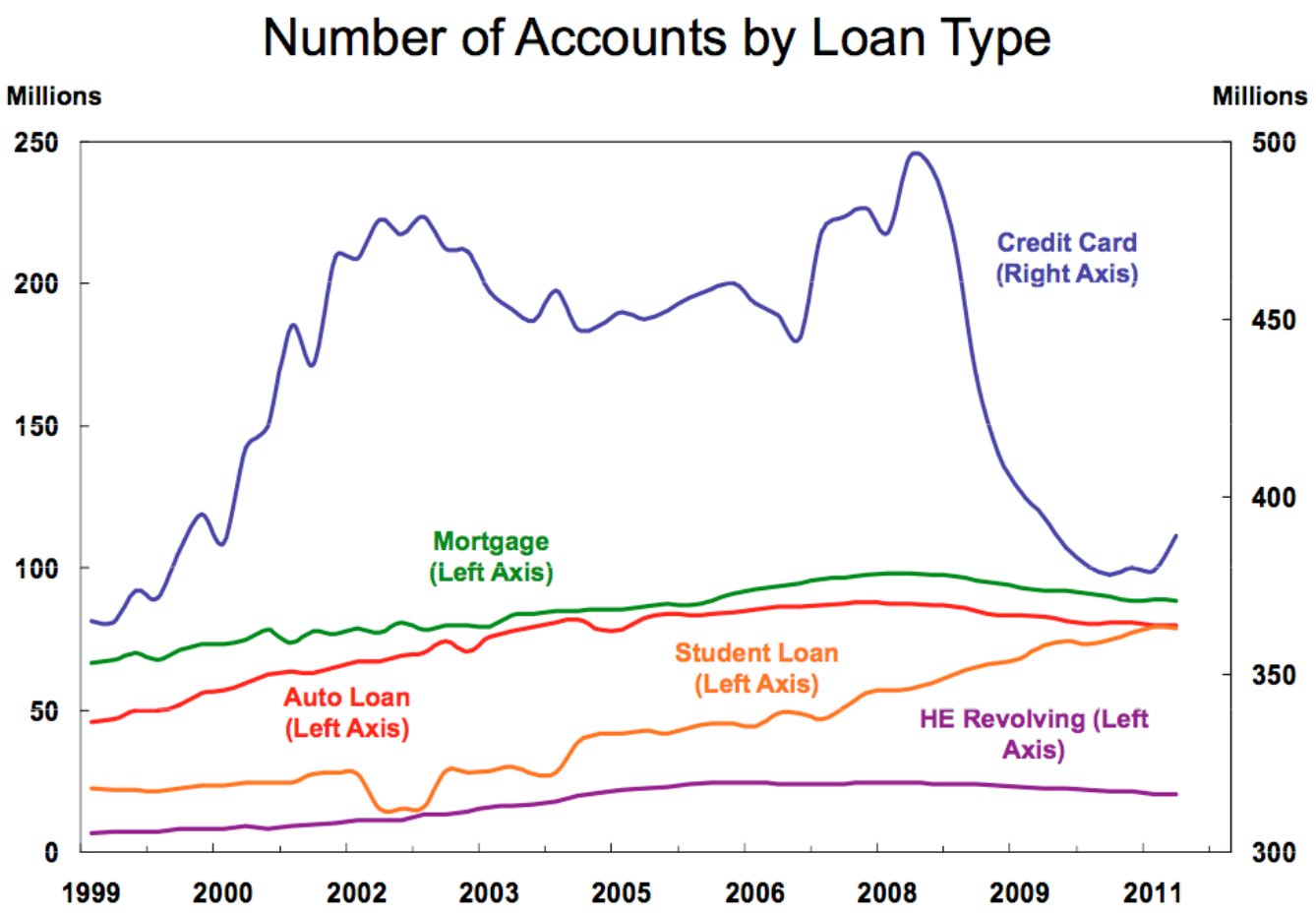
Trends: Consumer Debt

- August 2011 report from US Federal Reserve



Trends: Loan Volumes

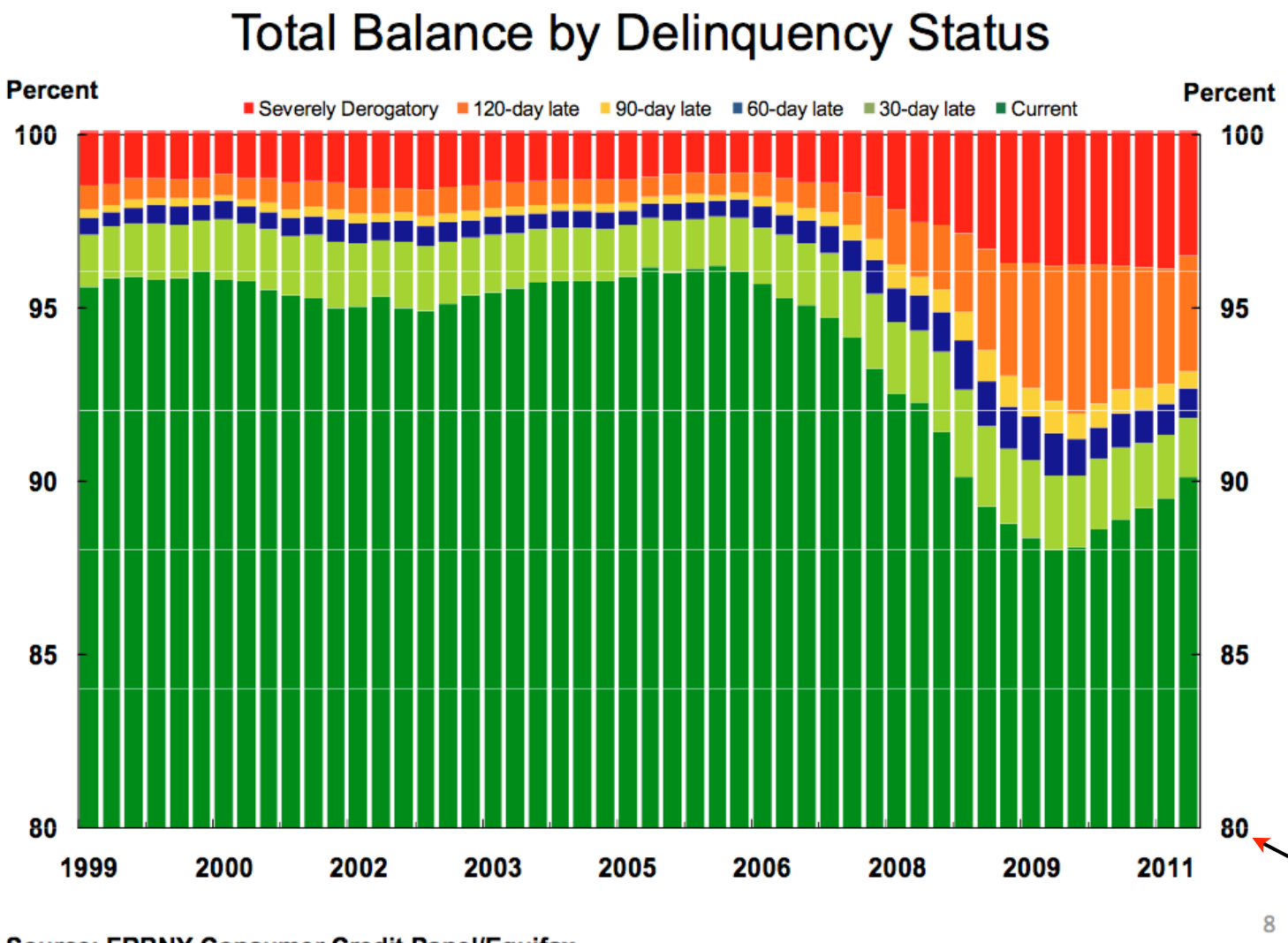
- "Flight to quality"



Source: FRBNY Consumer Credit Panel/Equifax

Trends: Default Balances

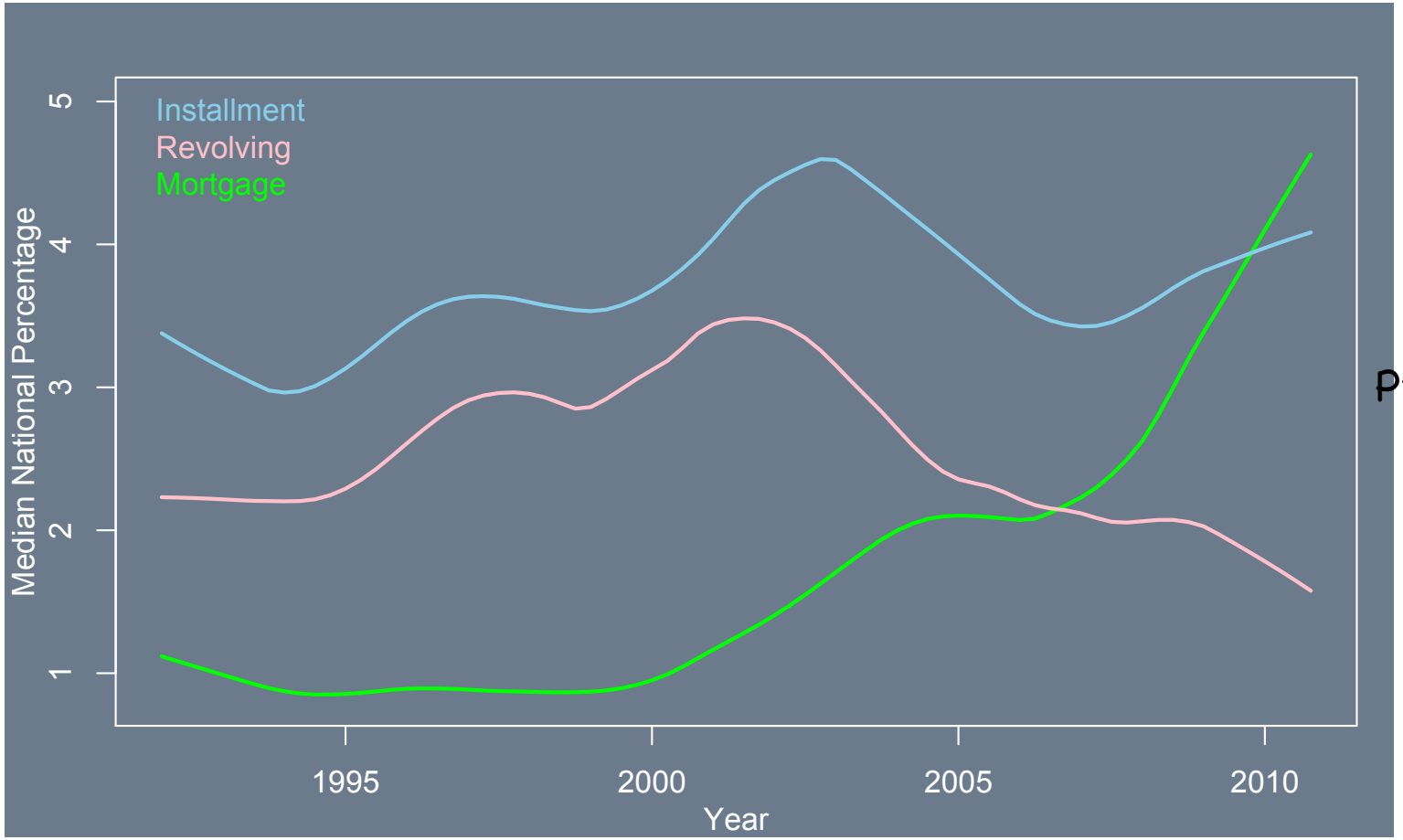
- Balance primarily composed of mortgages.



Source: FRBNY Consumer Credit Panel/Equifax

Default Rates

- Median county-level quarterly default rates, 60 days past due and slightly smoothed
- Changing association among rates

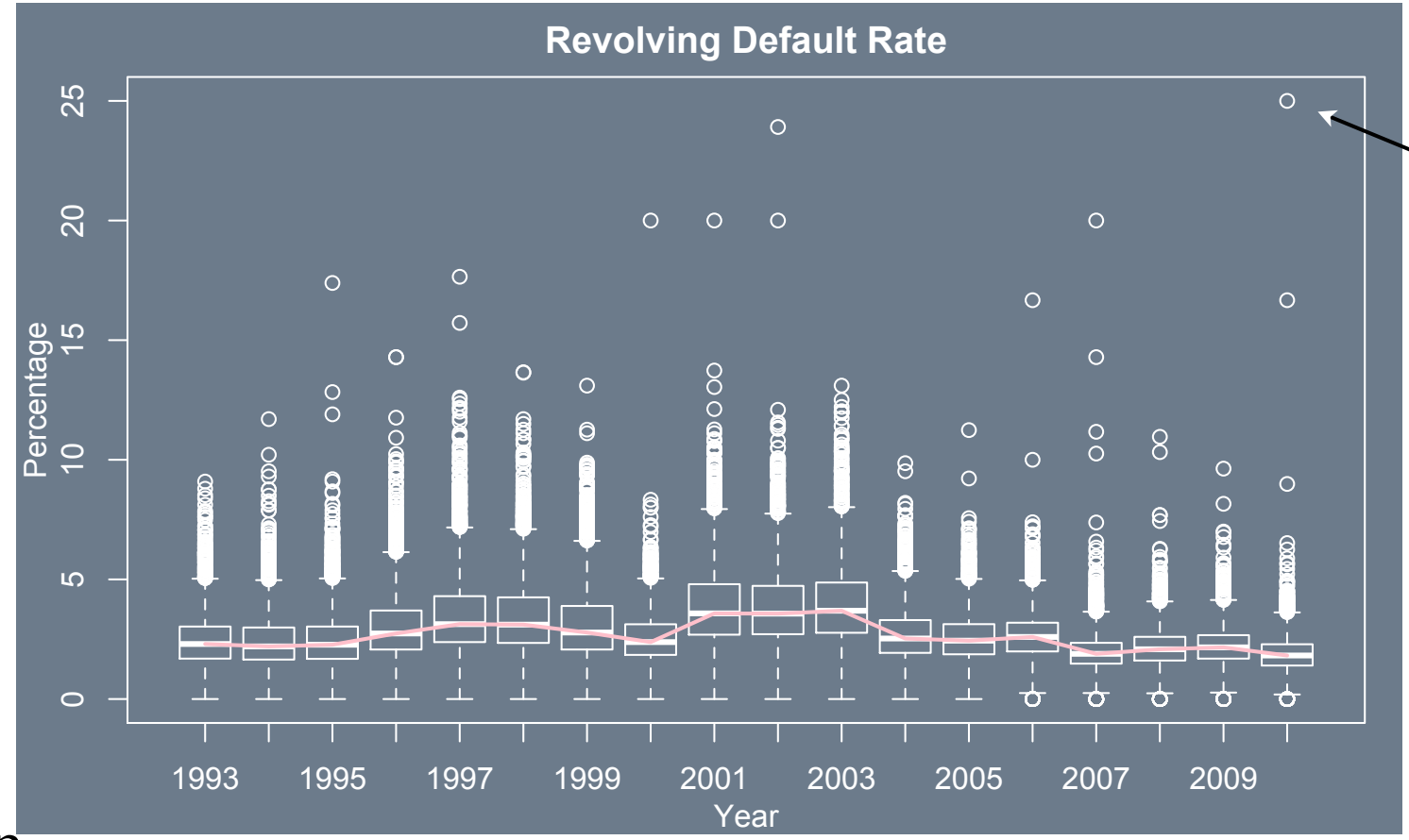


Predict?

Maps

Enormous Heterogeneity

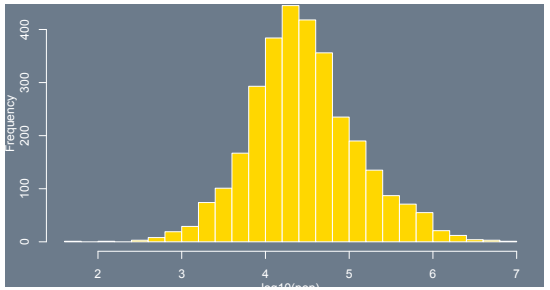
- Revolving default rates
  - Smooth national series
  - Huge regional variation in US:
    Near zero in some counties, 25% in others.



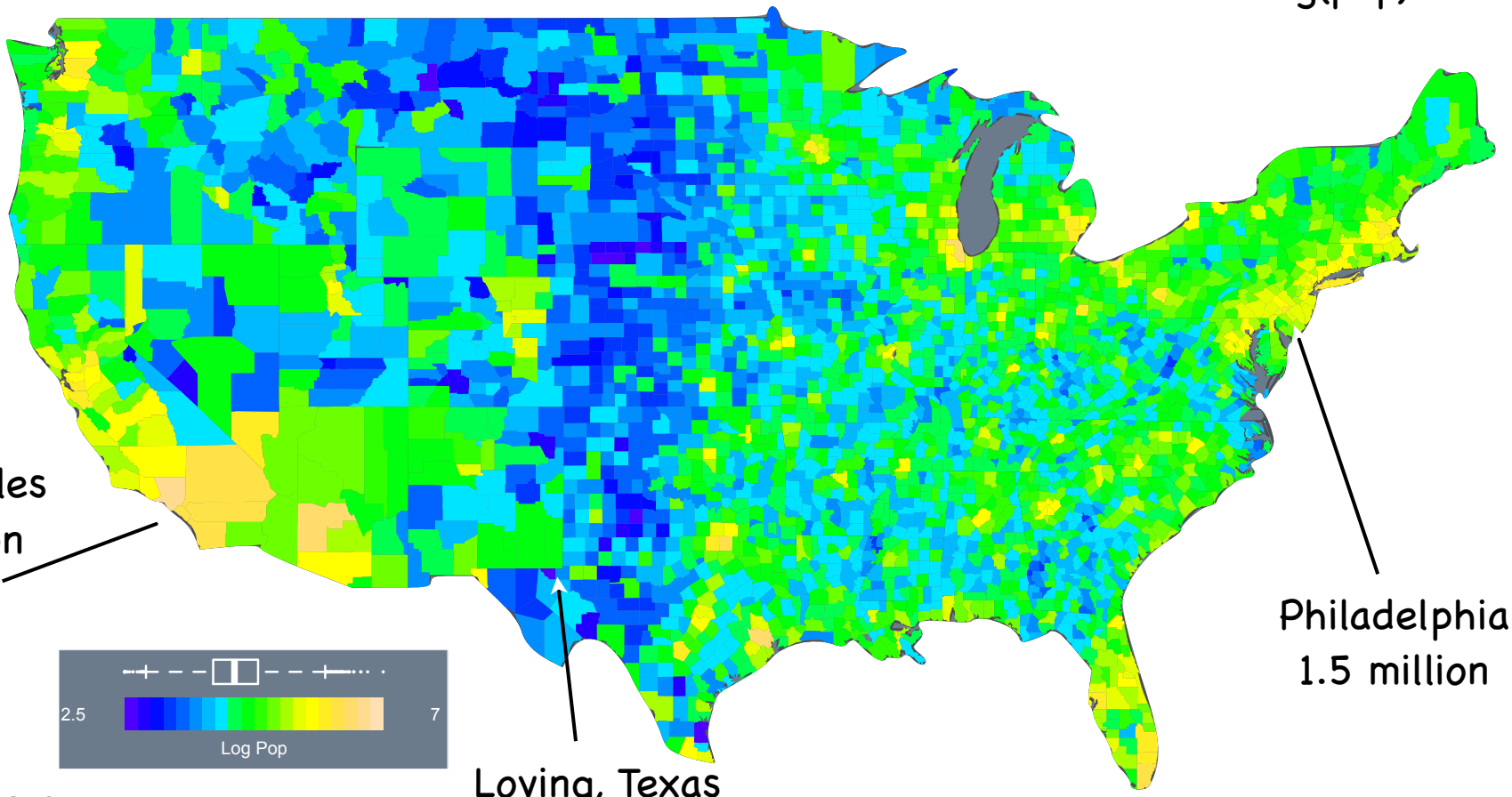
Small populations

Variation in Population

- Some counties have a hundreds, others have millions (lognormal)



log(pop)



Los Angeles
10 million

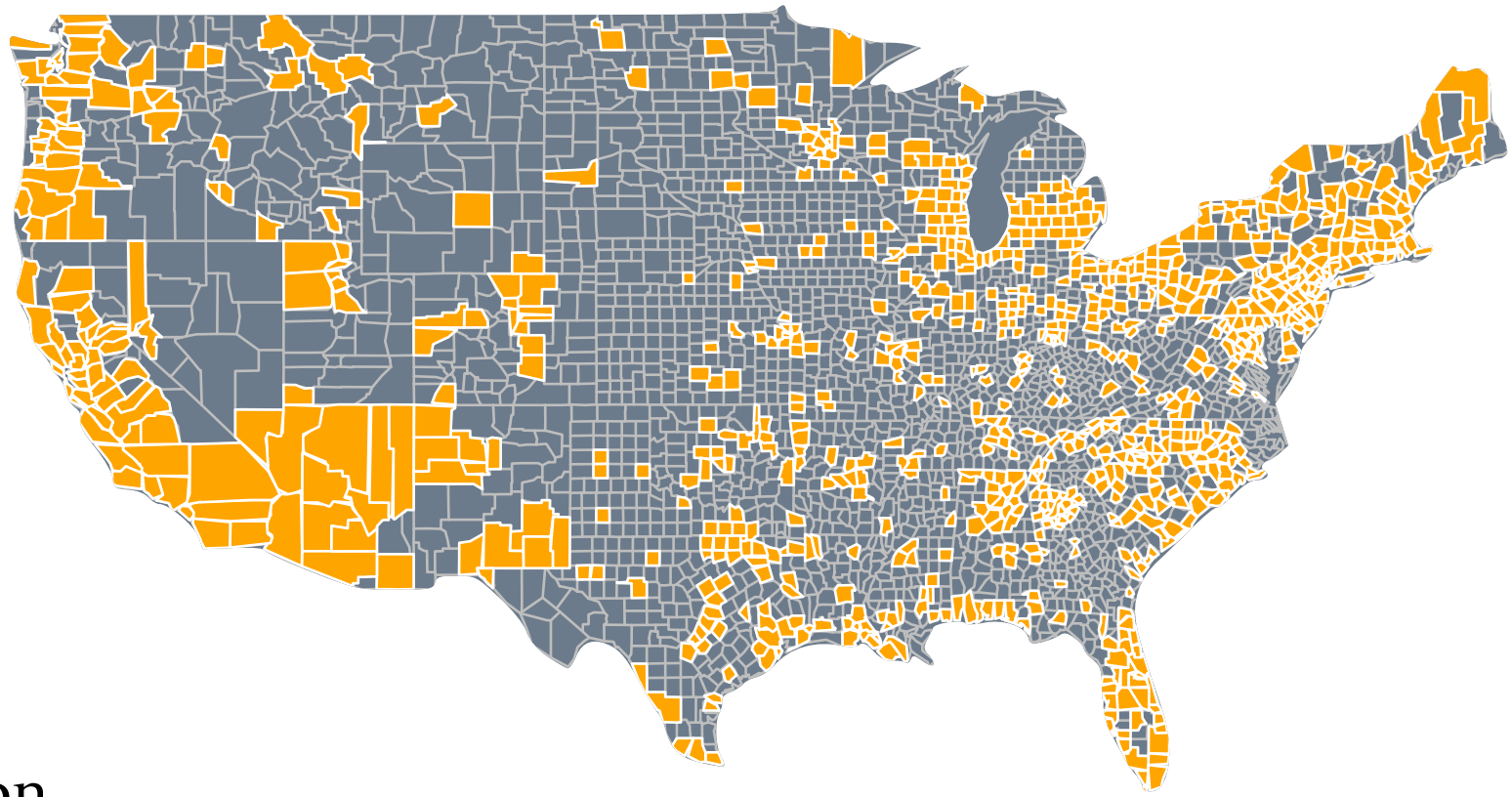
Loving, Texas
60

Philadelphia
1.5 million

Wharton
Department of Statistics

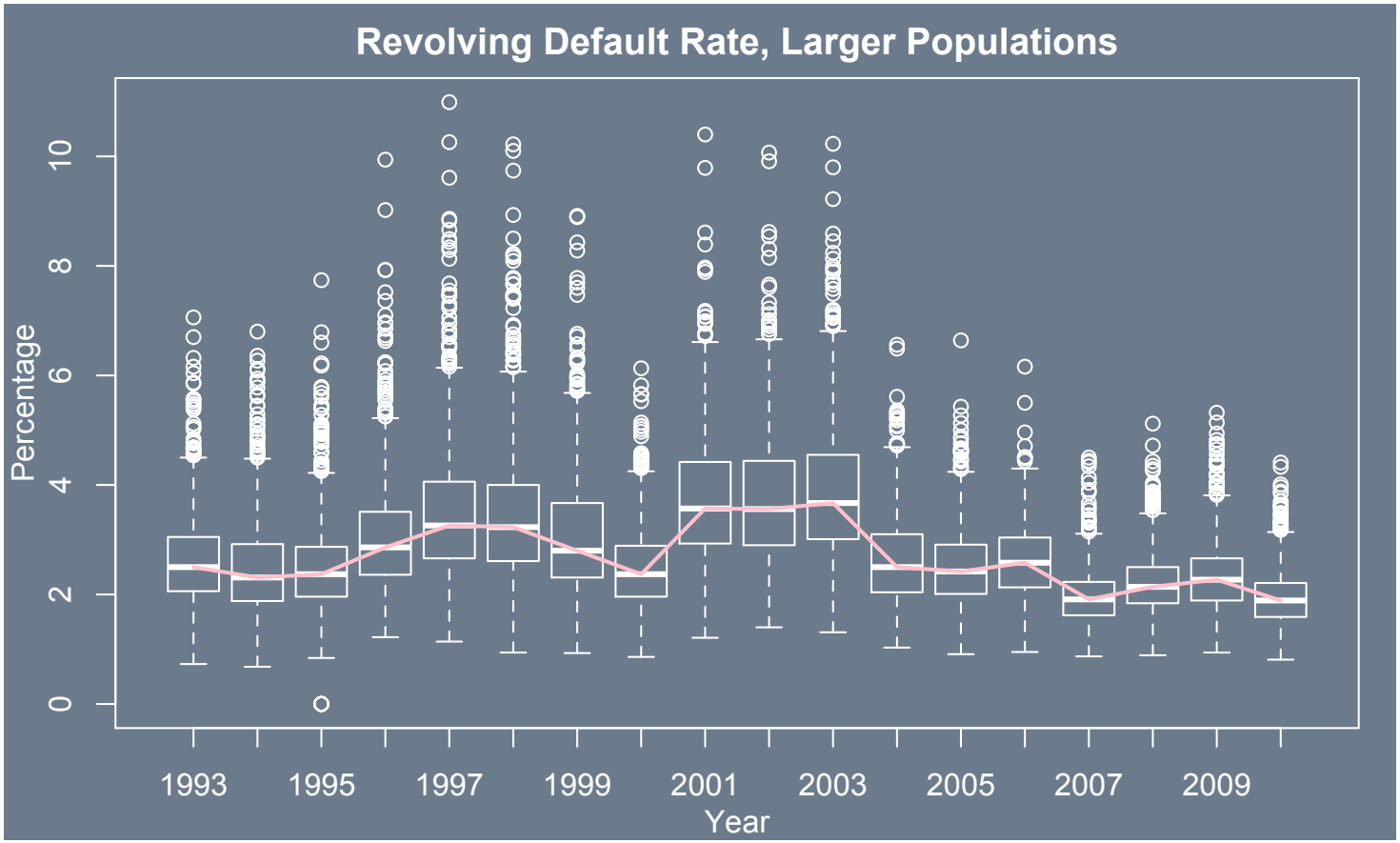
Analysis Subset

- Default rates and demographics are unreliable in sparsely populated areas.
- Limit analysis to counties with 50,000 people
  - Covers 85% of population, 900+ counties



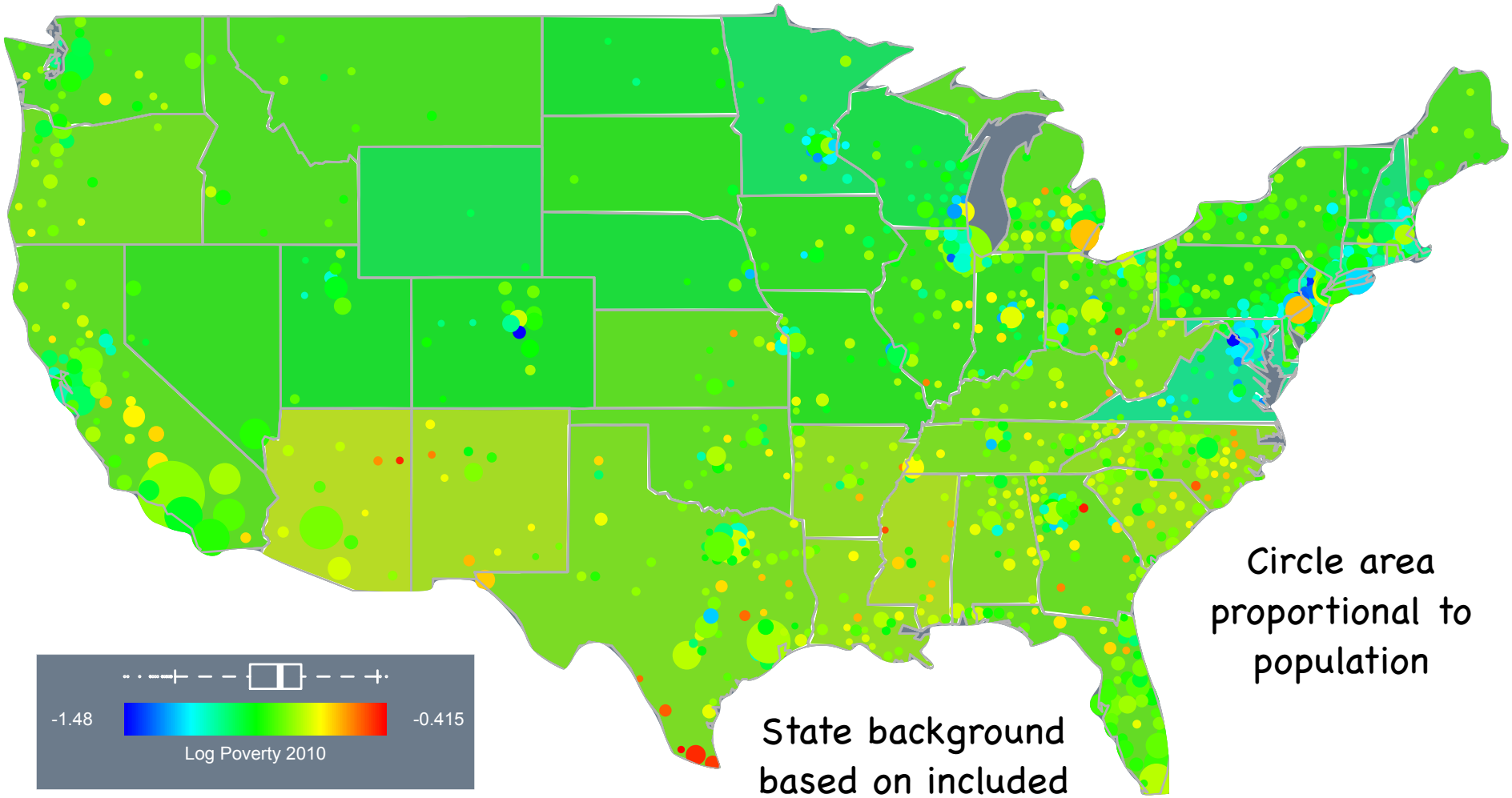
Heterogeneity Persists

- Revolving default rates
  - Rates skewed, close to log normal
  - More reliable, fewer missing



Spatial Patterns

- Poverty rates
  - Wealth concentrates around urban cores

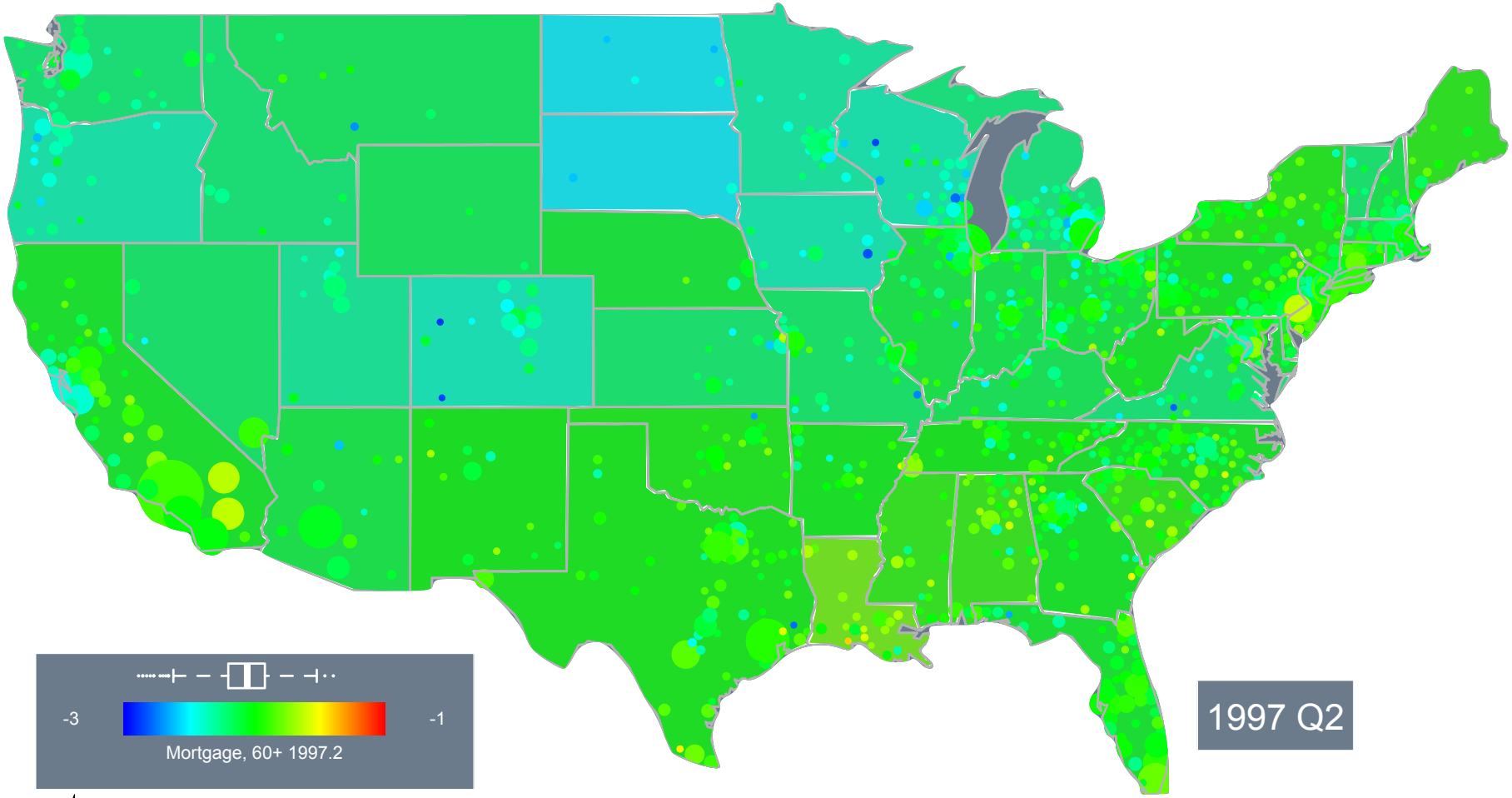


Circle area proportional to population

State background based on included counties.

Evolution of Defaults

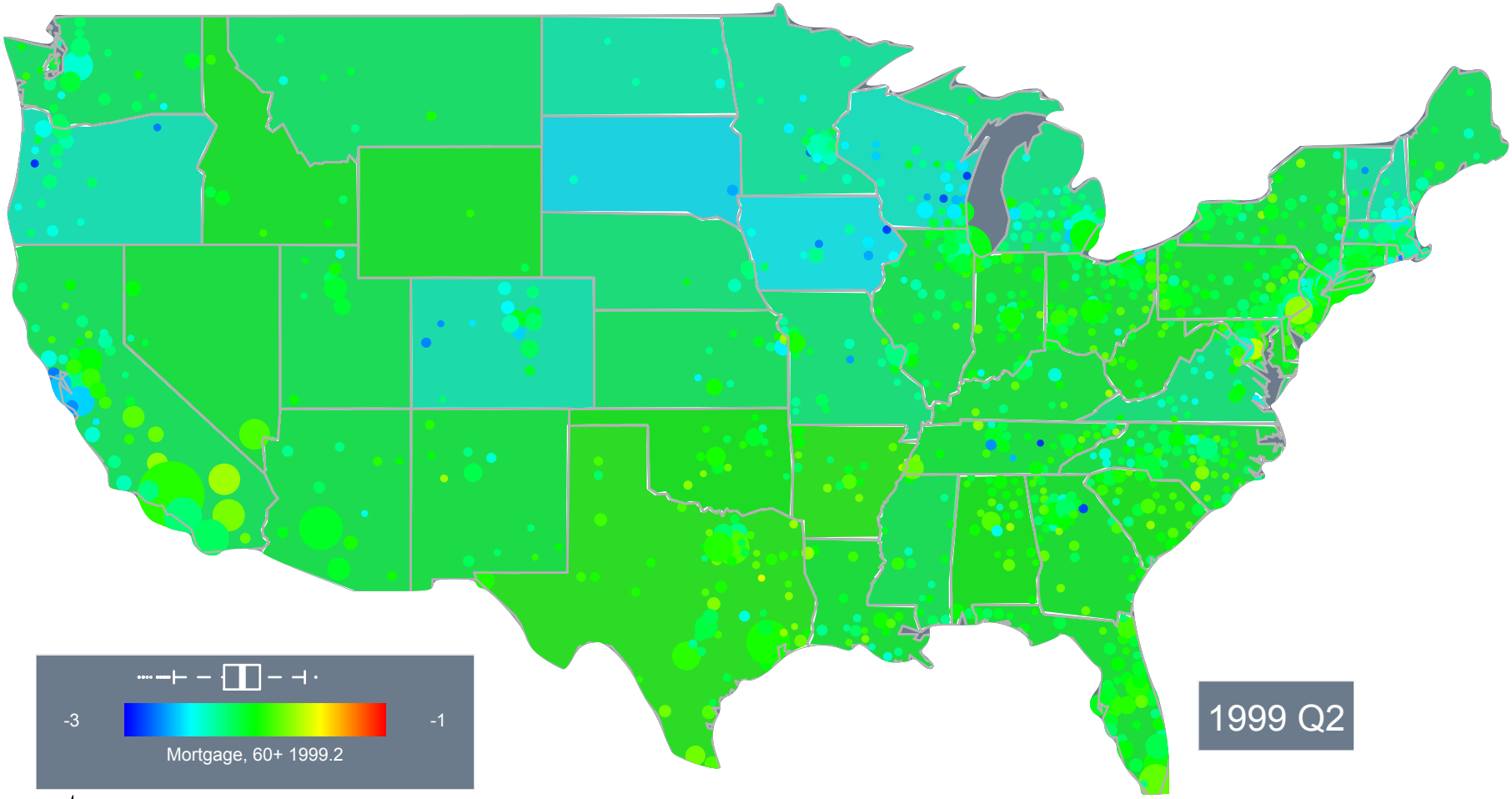
- Mortgage rates
  - Rates on log scale



Wharton
Department of Statistics

Evolution of Defaults

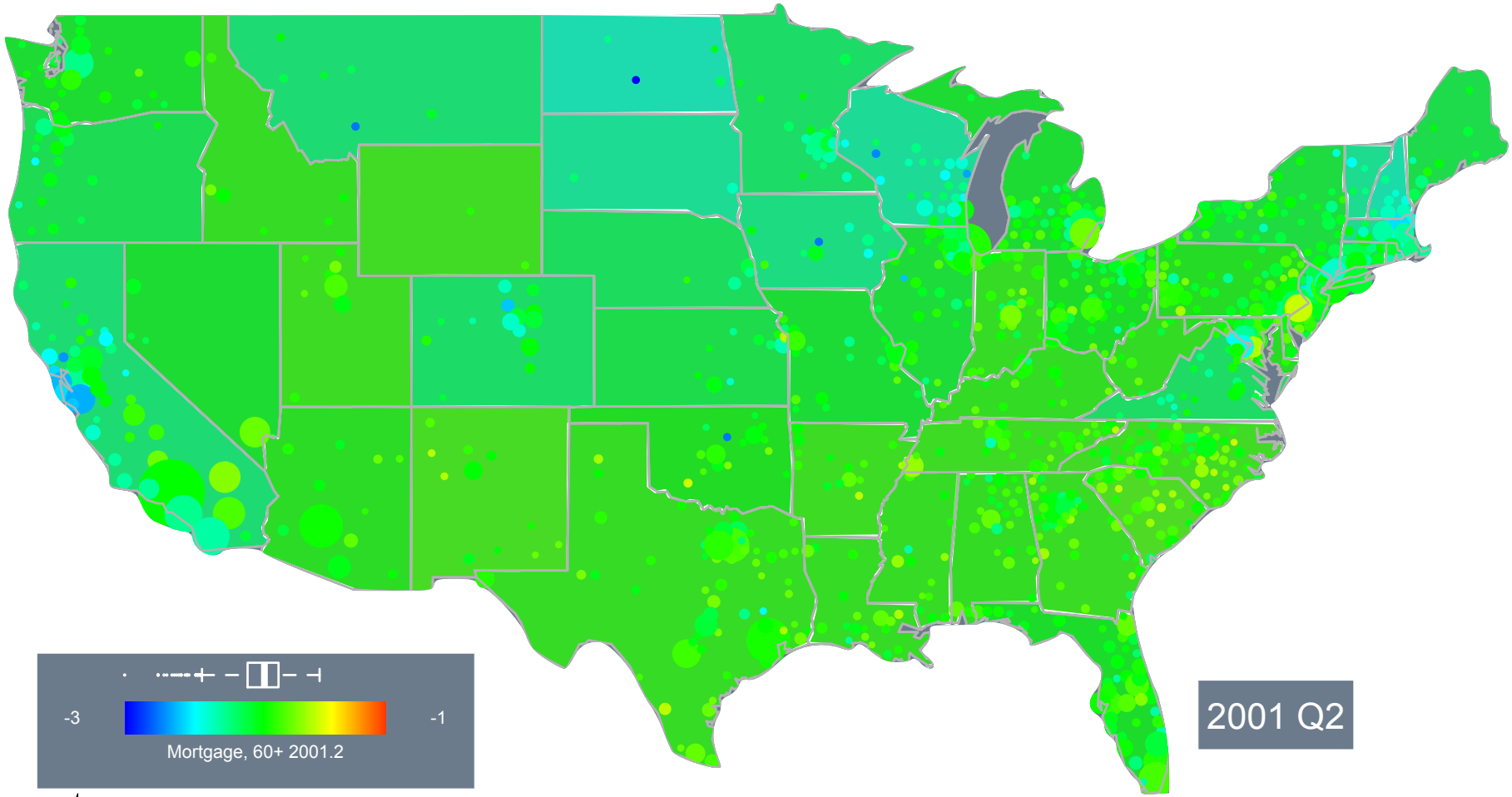
- Mortgage rates
  - Rates on log scale



Wharton
Department of Statistics

Evolution of Defaults

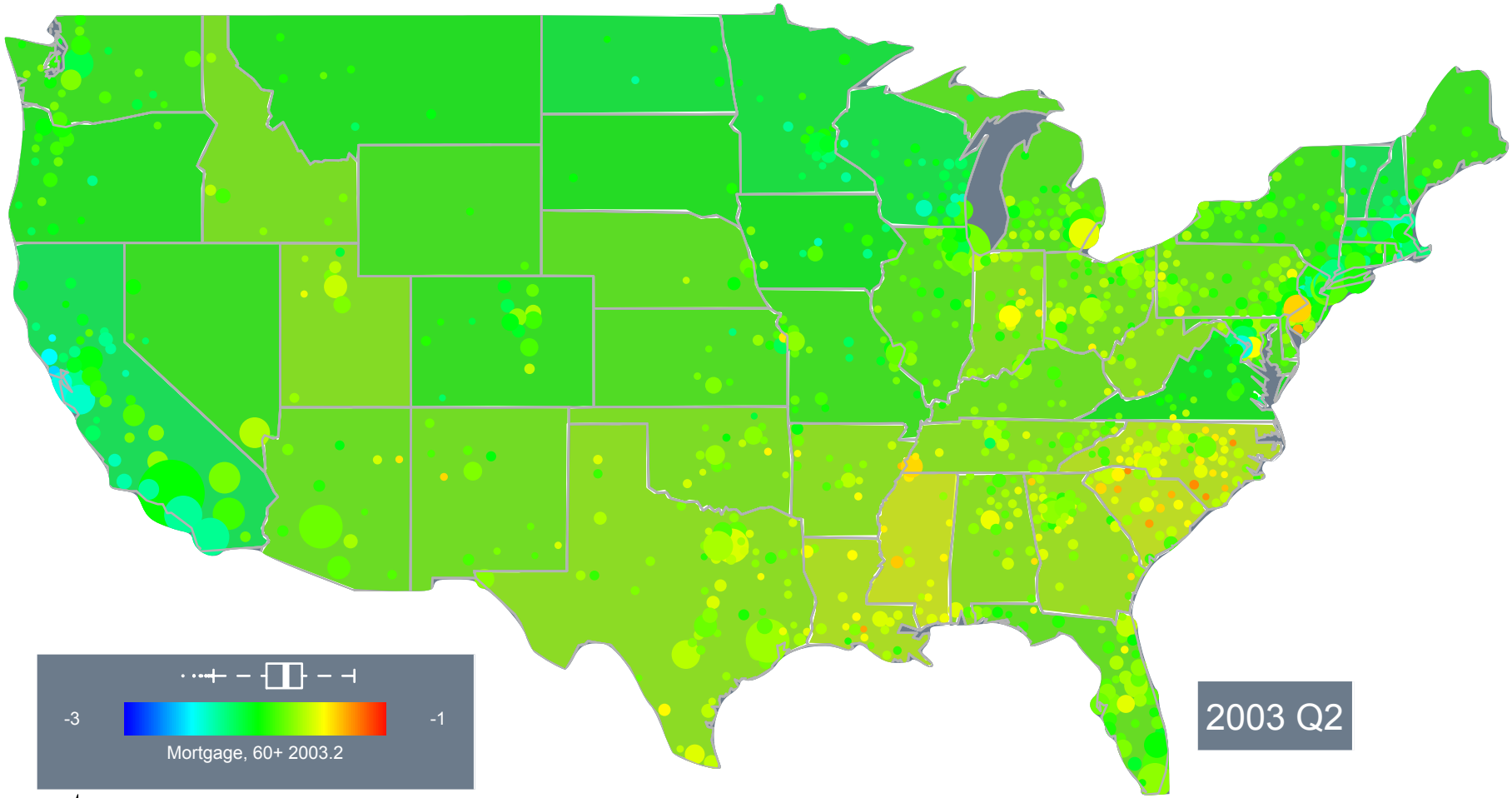
- Mortgage rates
  - Rates on log scale



Wharton
Department of Statistics

Evolution of Defaults

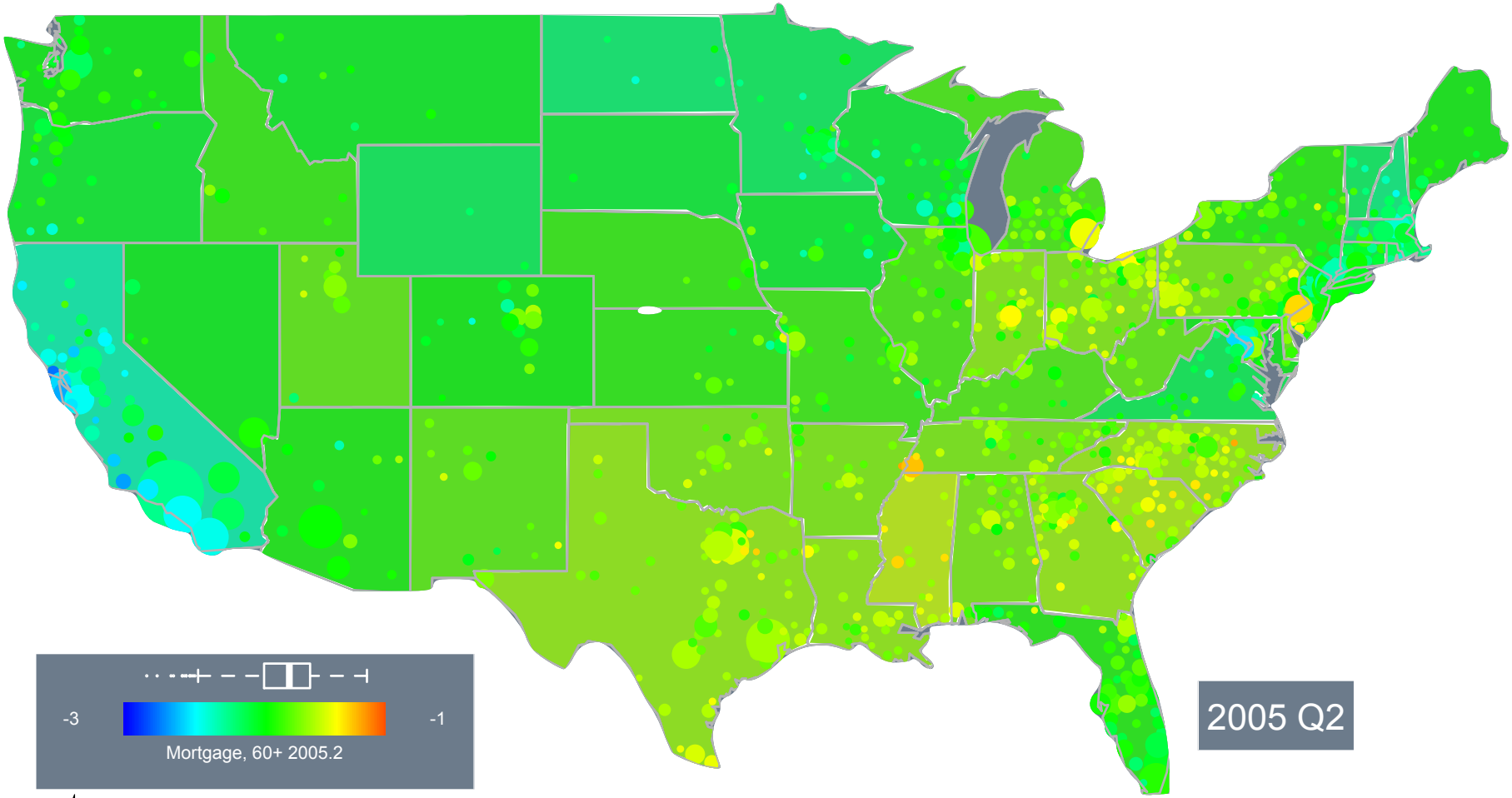
- Mortgage rates
  - Rates on log scale



Wharton
Department of Statistics

Evolution of Defaults

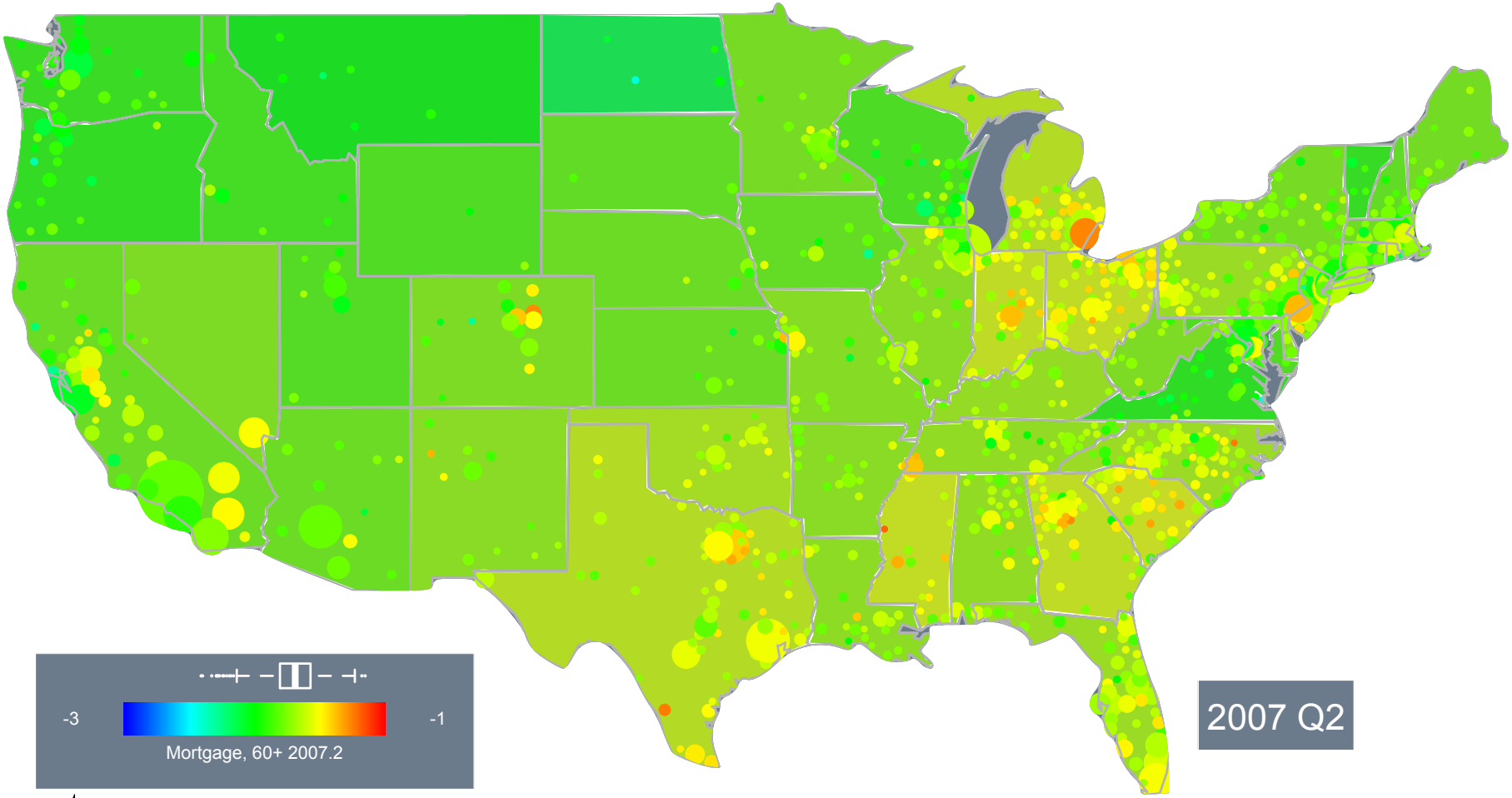
- Mortgage rates
  - Rates on log scale



Wharton
Department of Statistics

Evolution of Defaults

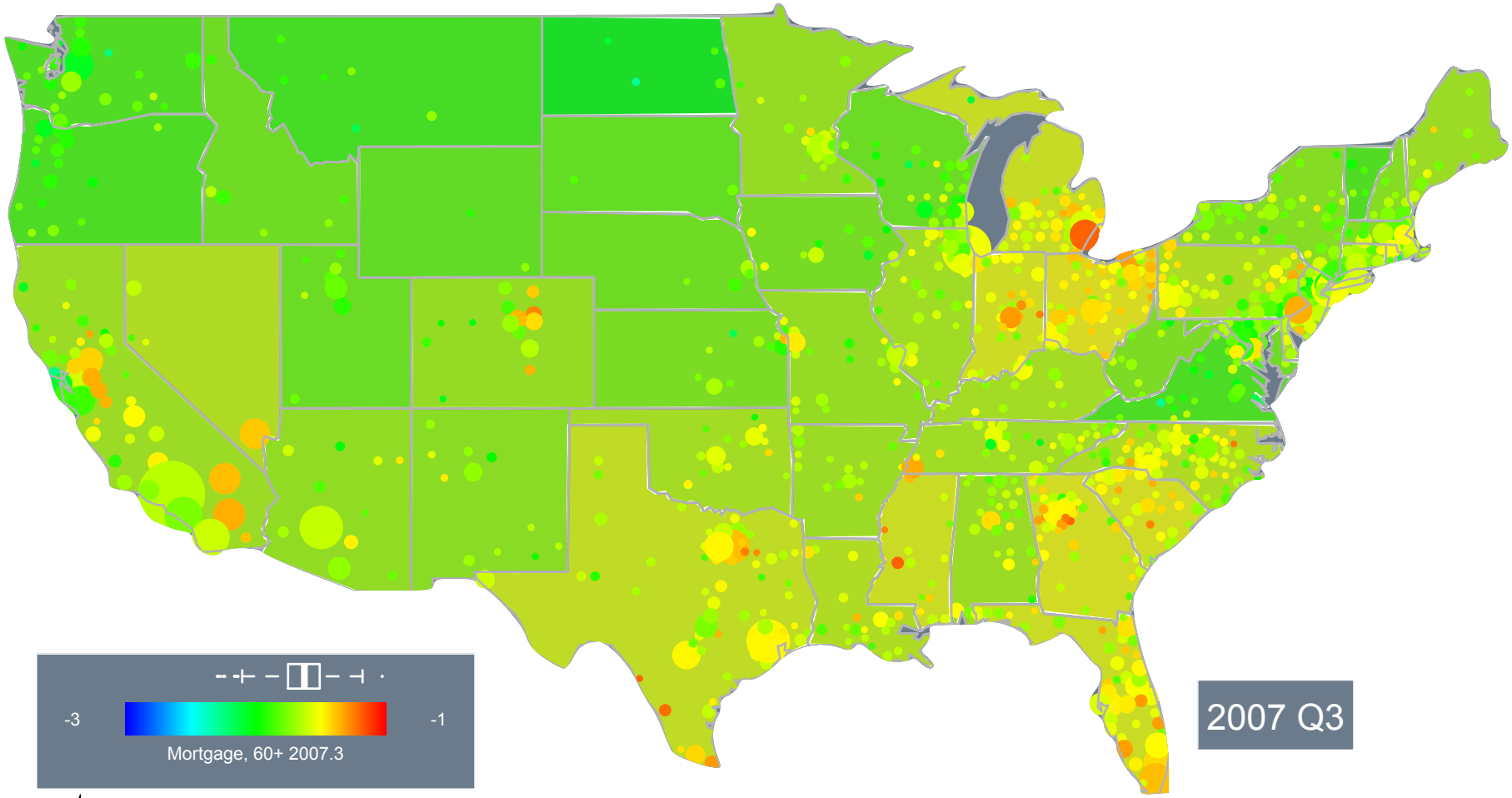
- Mortgage rates
  - Rates on log scale



Wharton
Department of Statistics

Evolution of Defaults

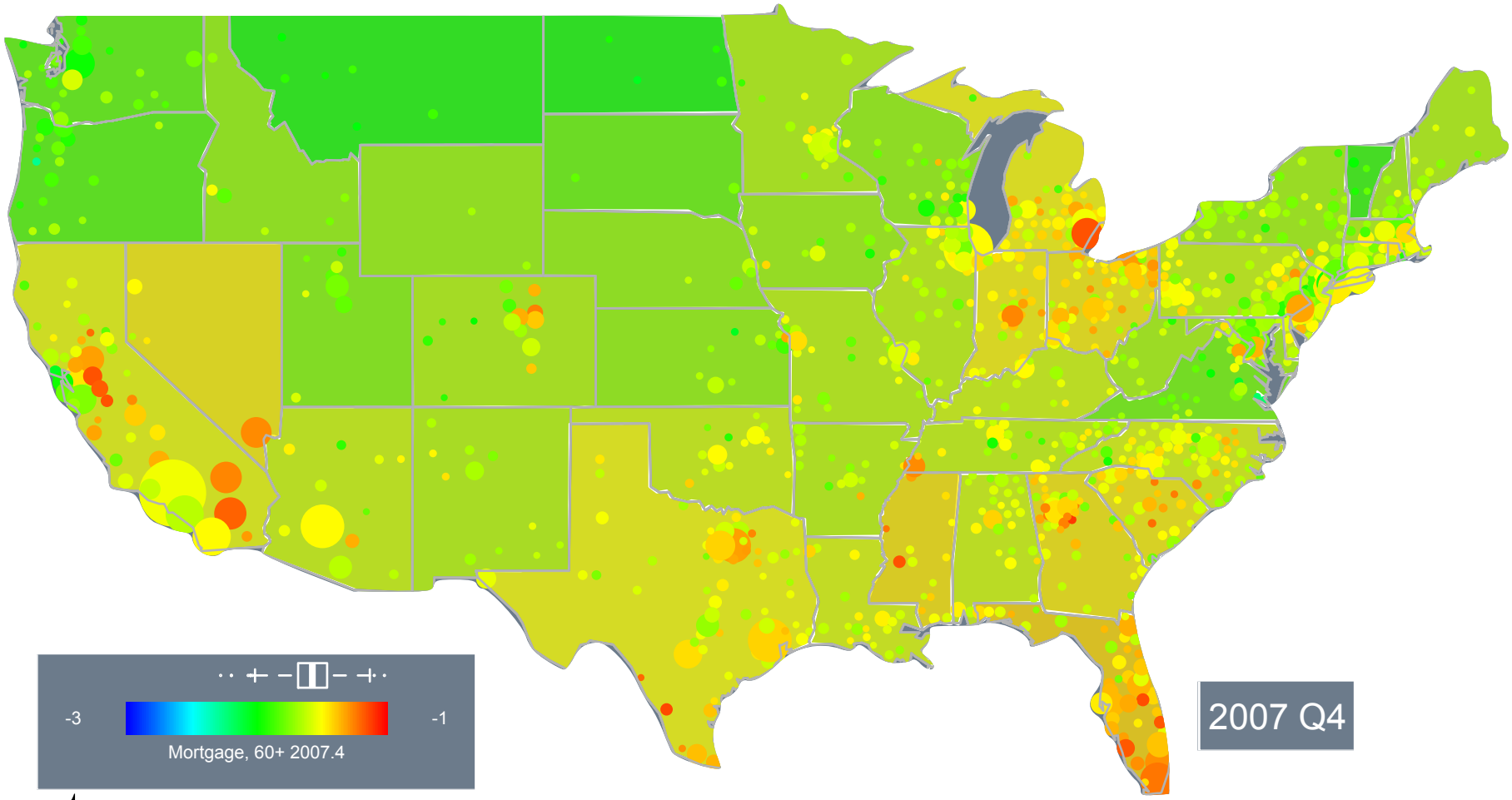
- Mortgage rates
  - Rates on log scale



Wharton
Department of Statistics

Evolution of Defaults

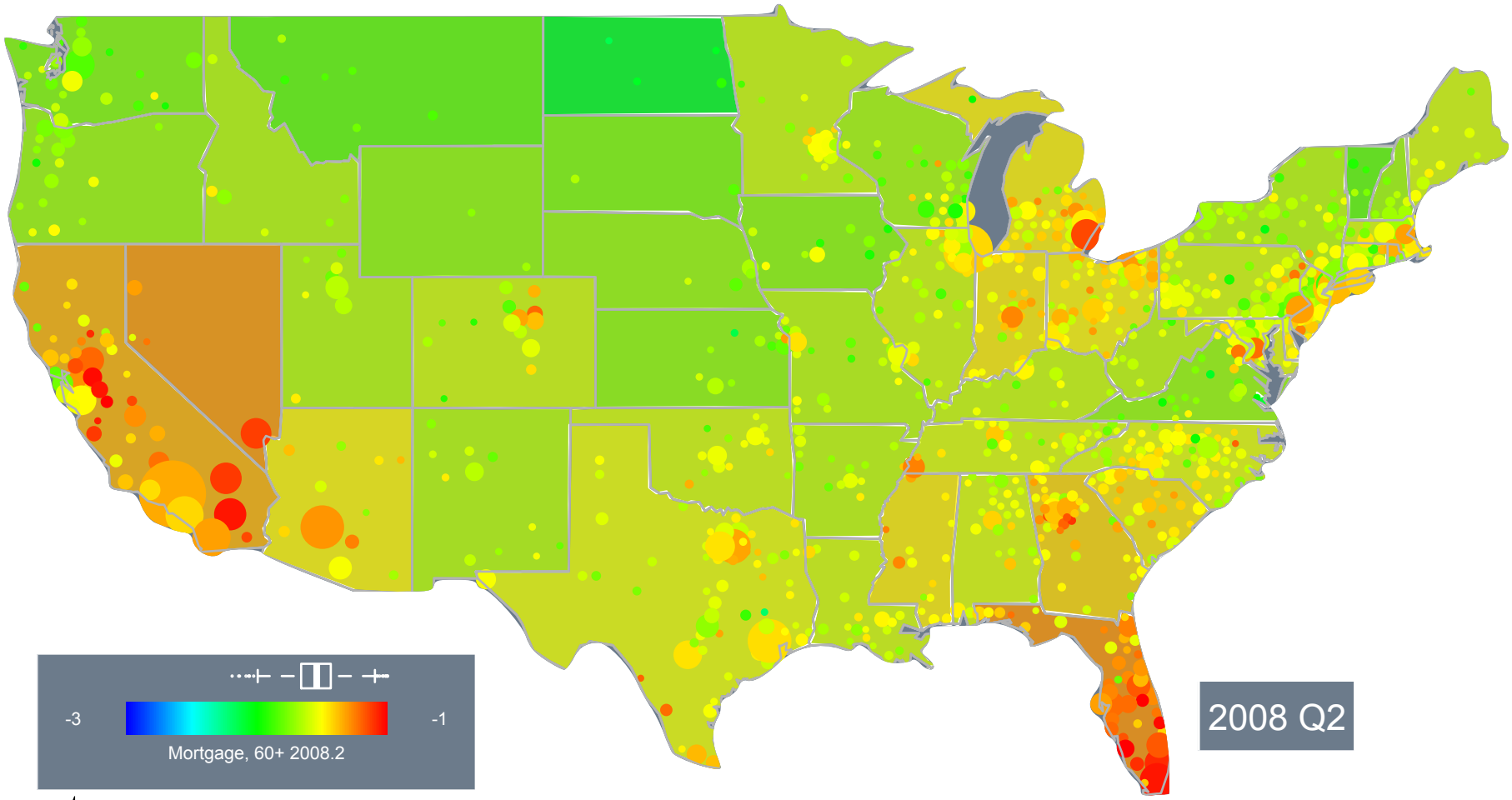
- Mortgage rates
  - Rates on log scale



Wharton
Department of Statistics

Evolution of Defaults

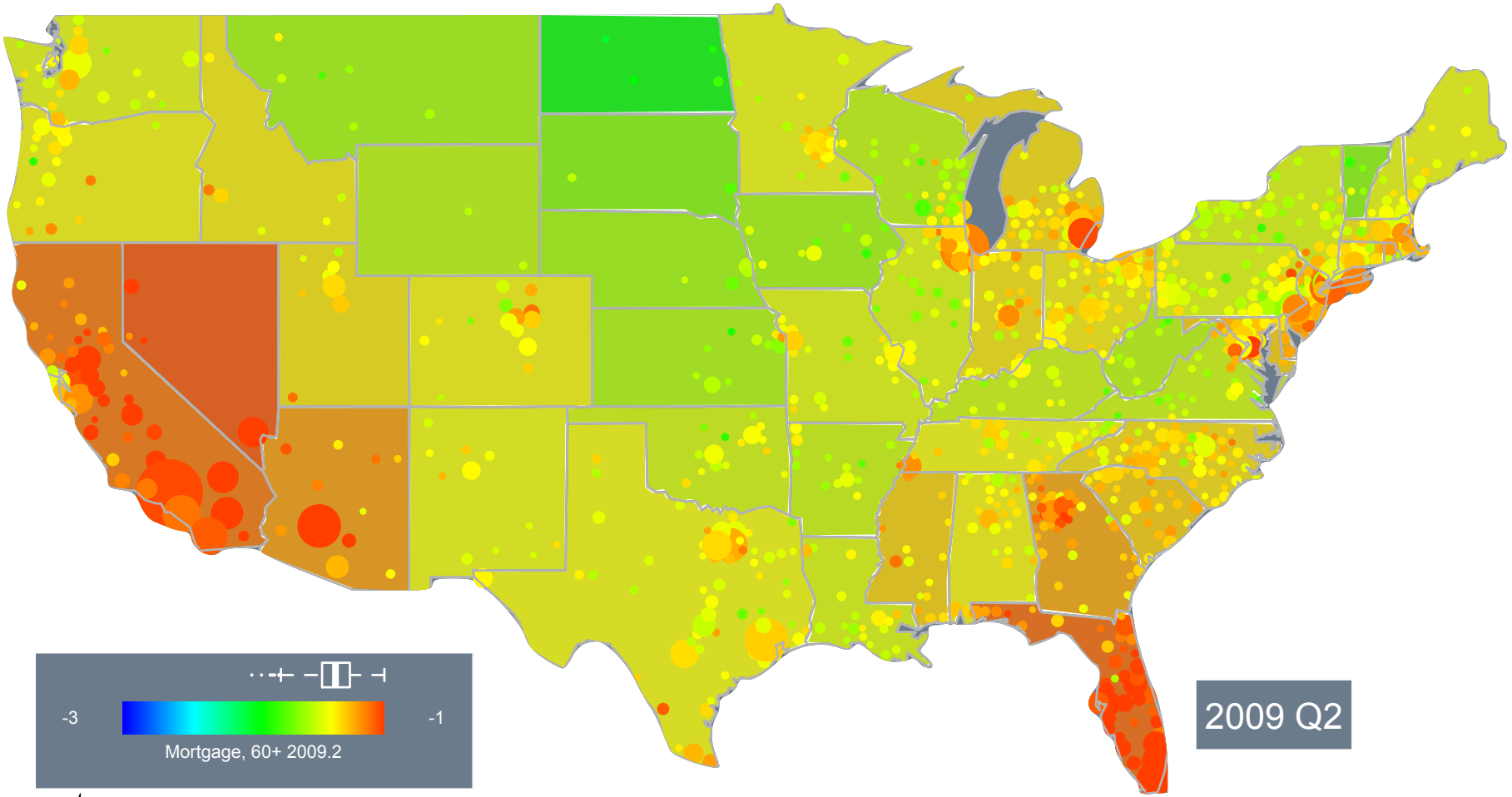
- Mortgage rates
  - Rates on log scale



Wharton
Department of Statistics

Evolution of Defaults

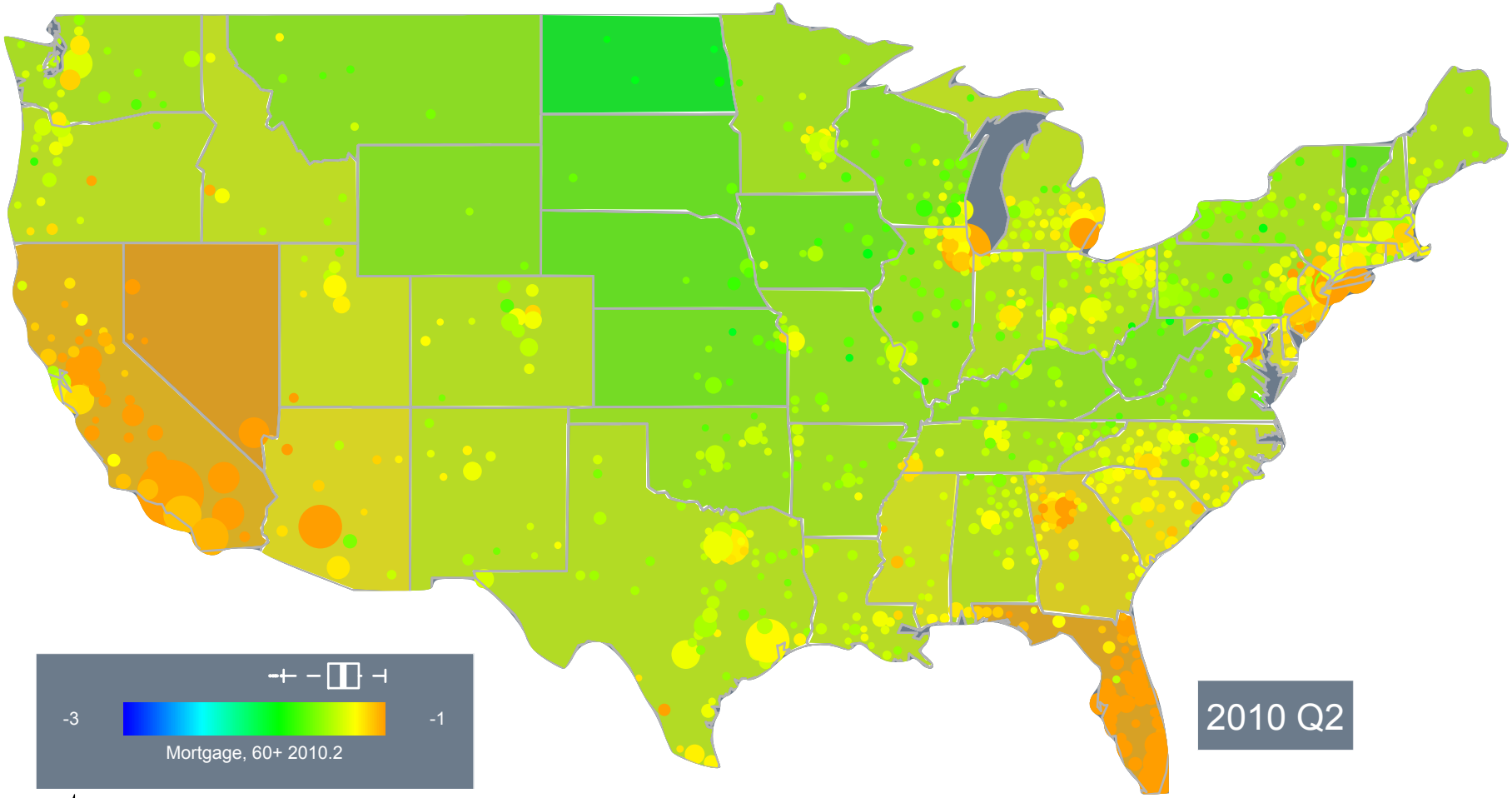
- Mortgage rates
  - Rates on log scale



Wharton
Department of Statistics

Evolution of Defaults

- Mortgage rates
  - Rates on log scale



Wharton
Department of Statistics

Spatial Correlations

- Standard measure of spatial correlation

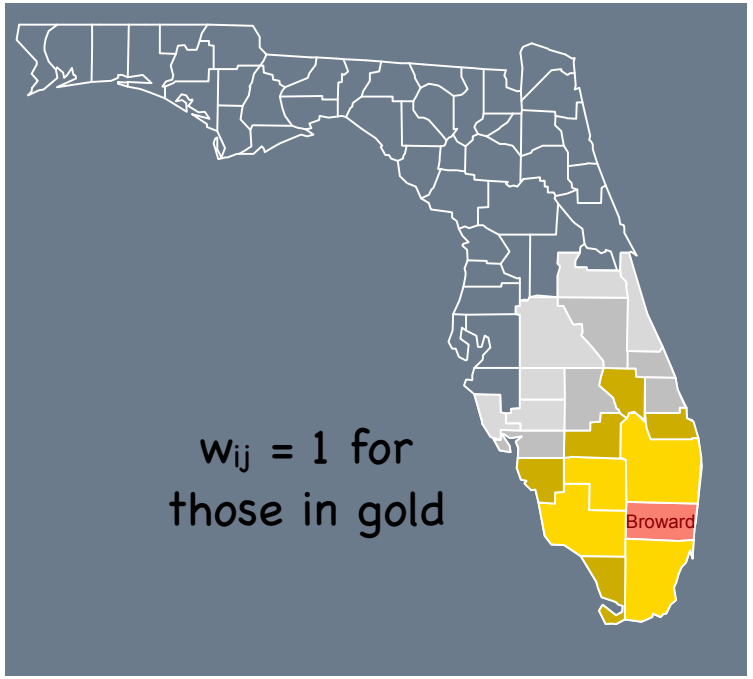
$$\text{Moran's } I = \frac{\sum w_{ij} (X_i - \overline{X})(X_j - \overline{X})}{\sum w_{ij} \quad s_x^2}$$

where w_{ij} identify 'neighbors'.

- Example

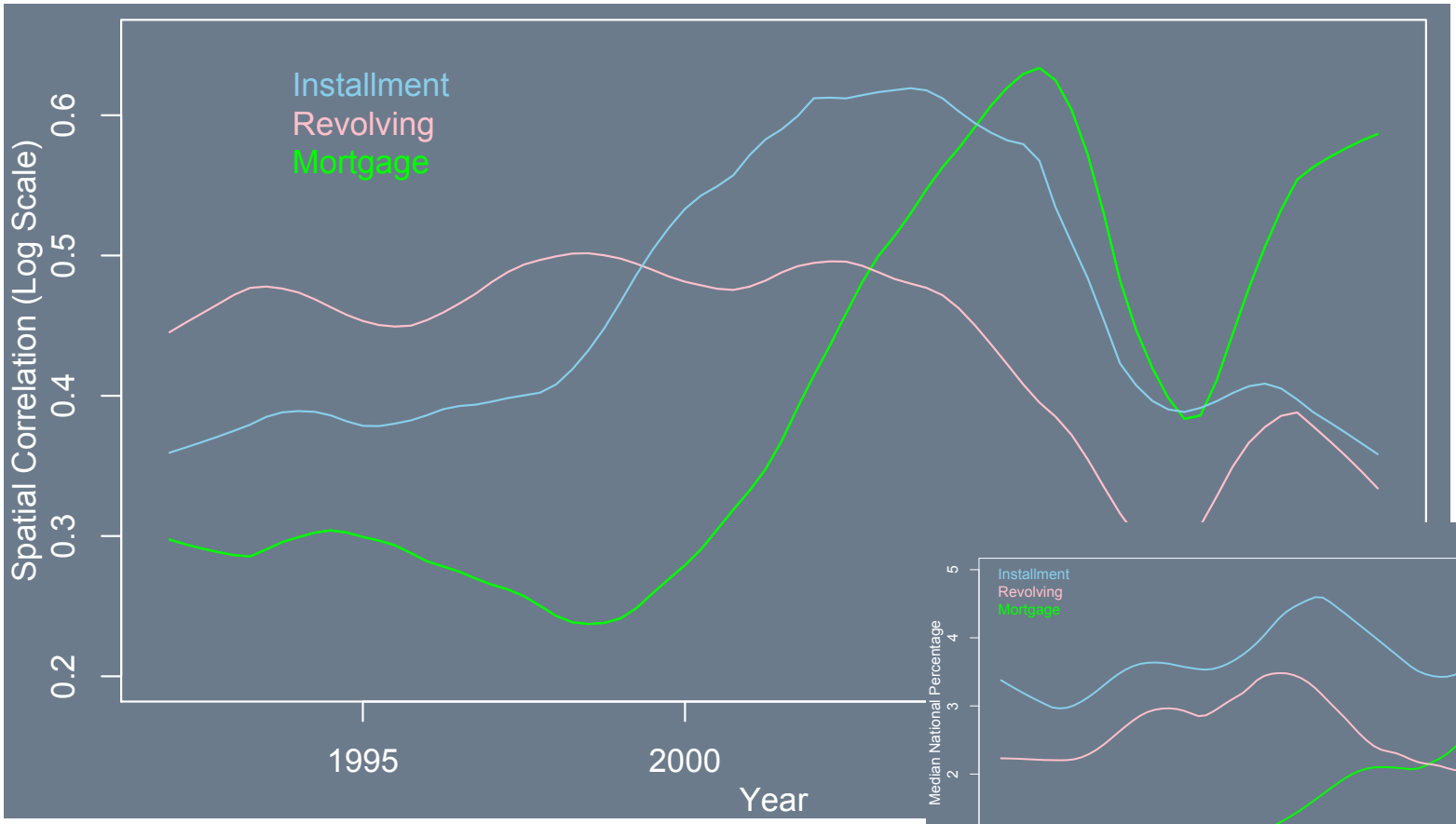
w_{ij} = 1 if within two layers of the target county.

w_{ij} = 0 otherwise.



Spatial Correlations

- Moran's I shows surprising correlation for various types of default.



Spatial Patterns

Correlation Risk

- Spatial correlations suggest correlation risk.
- Question
  - Pick a county c with neighbors N(c)
  - How much of the variation in default rates among neighbors of c can be described using a common trend?

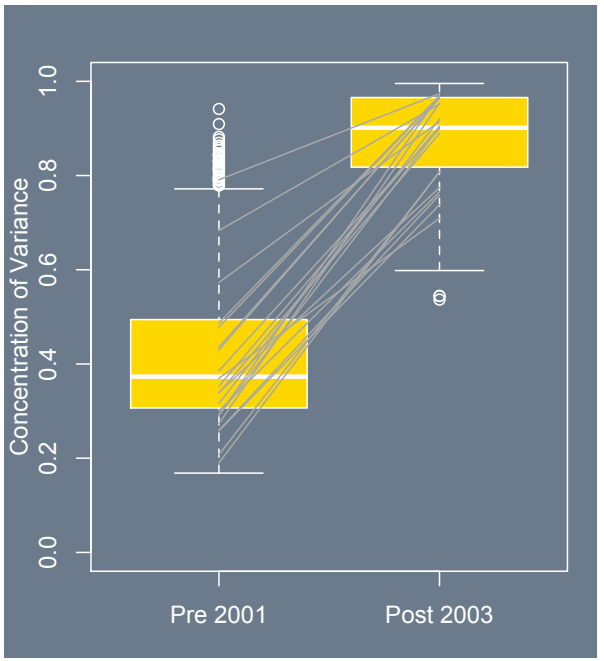
$$D_{N(c),t} \approx u_c \, v_t$$

- Principal components
  - First principal component of the covariance matrix S among the neighbors of a county
  - Largest eigenvalue indicates amount of variation represented by common trend

Correlation Risk

- Neighborhoods
  - Consider all neighborhoods among the 900+ counties in the analysis
  - Compare the percentage of variation in the first component using quarters before 2001 to the percentage in quarters after 2003

- Results
  - Mortgage default rates
  - Percentage of variation rises basically everywhere
  - Median increases from 0.4 up to 0.9.



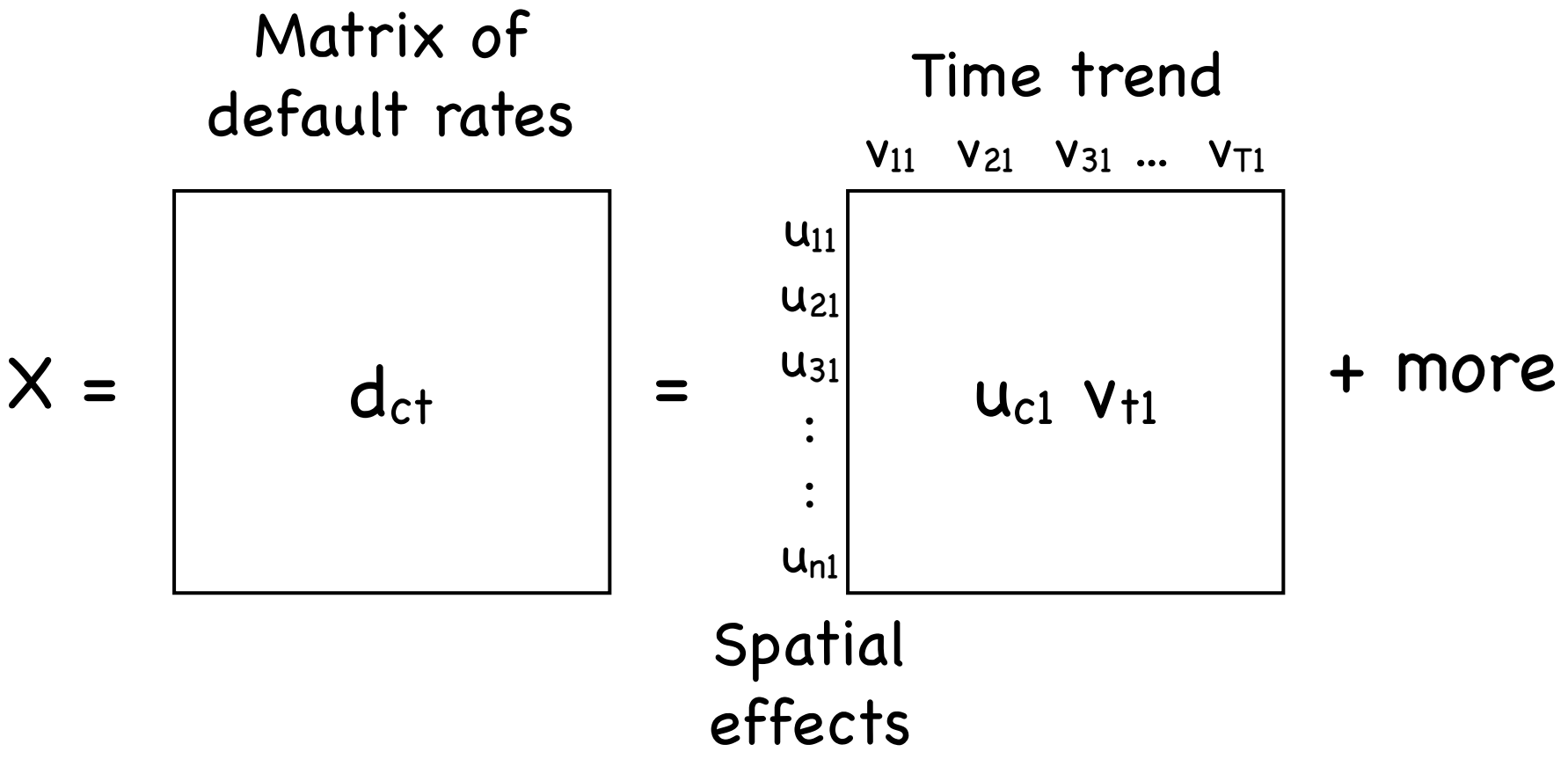
Patterns in Variation

- Borrow technique from climatology
  - Empirical orthogonal functions
  - Segmentation: Find locations that covary in time
- Singular value decomposition
  - Extend principal components
  - X holds default rates at 900 locations, 76 times
  - Approximation

    $$X = UDV', \text{ or } X = \Sigma\, u_i\, (d_i\, v_i')$$

  - U captures spatial patterns, V holds time
- Orthogonal rotation
  - Rotate the orthogonal factors to clarify geographic clustering

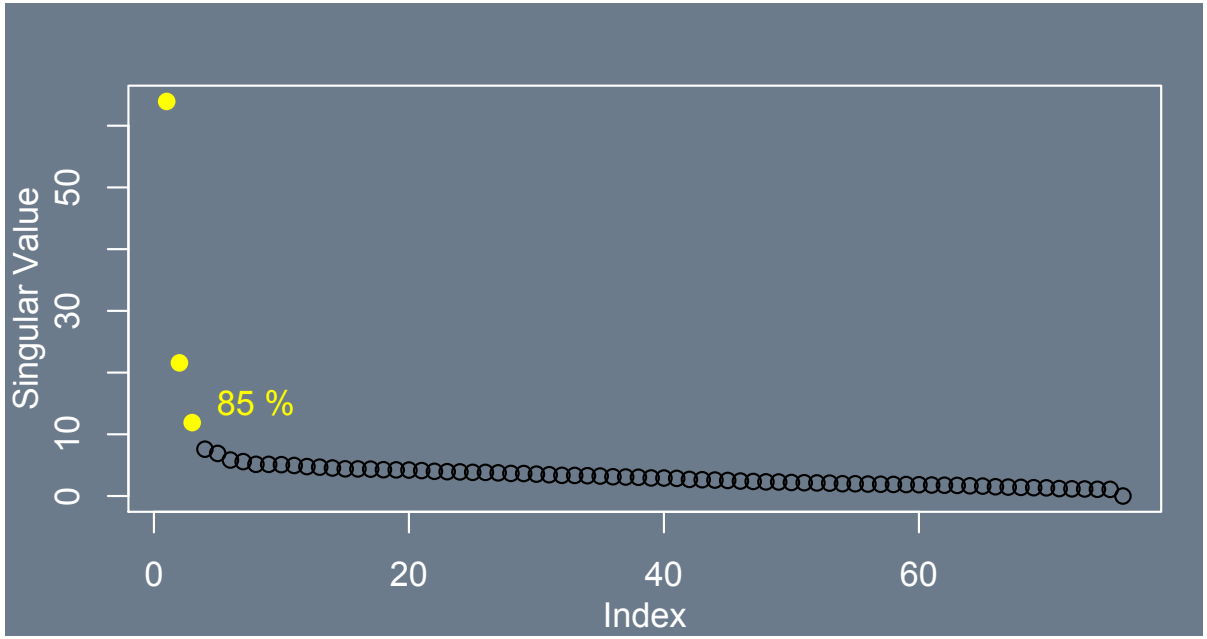
Low-Rank Approximation



Decomposition 'knows' nothing of time or space...
Are counties with common trends adjacent?

Singular Value Decomposition

- How many terms
  - Singular values suggest need three terms to represent variation in mortgage defaults.

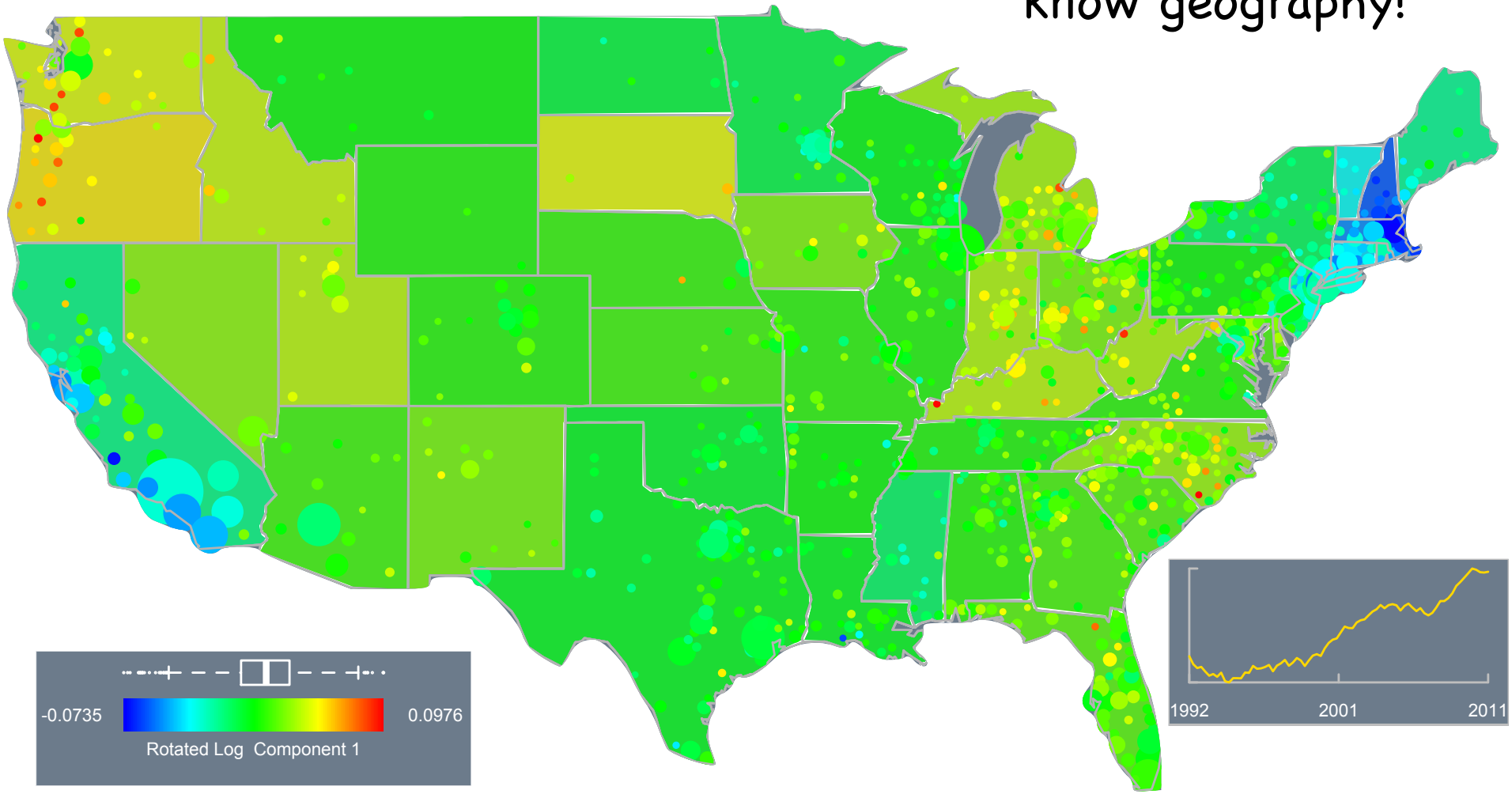


- Rotated components
  - Sacrifice orthogonality to improve interpretation
  - Each rotated component has ≈⅓ of variance

First Component

- Long term problems...

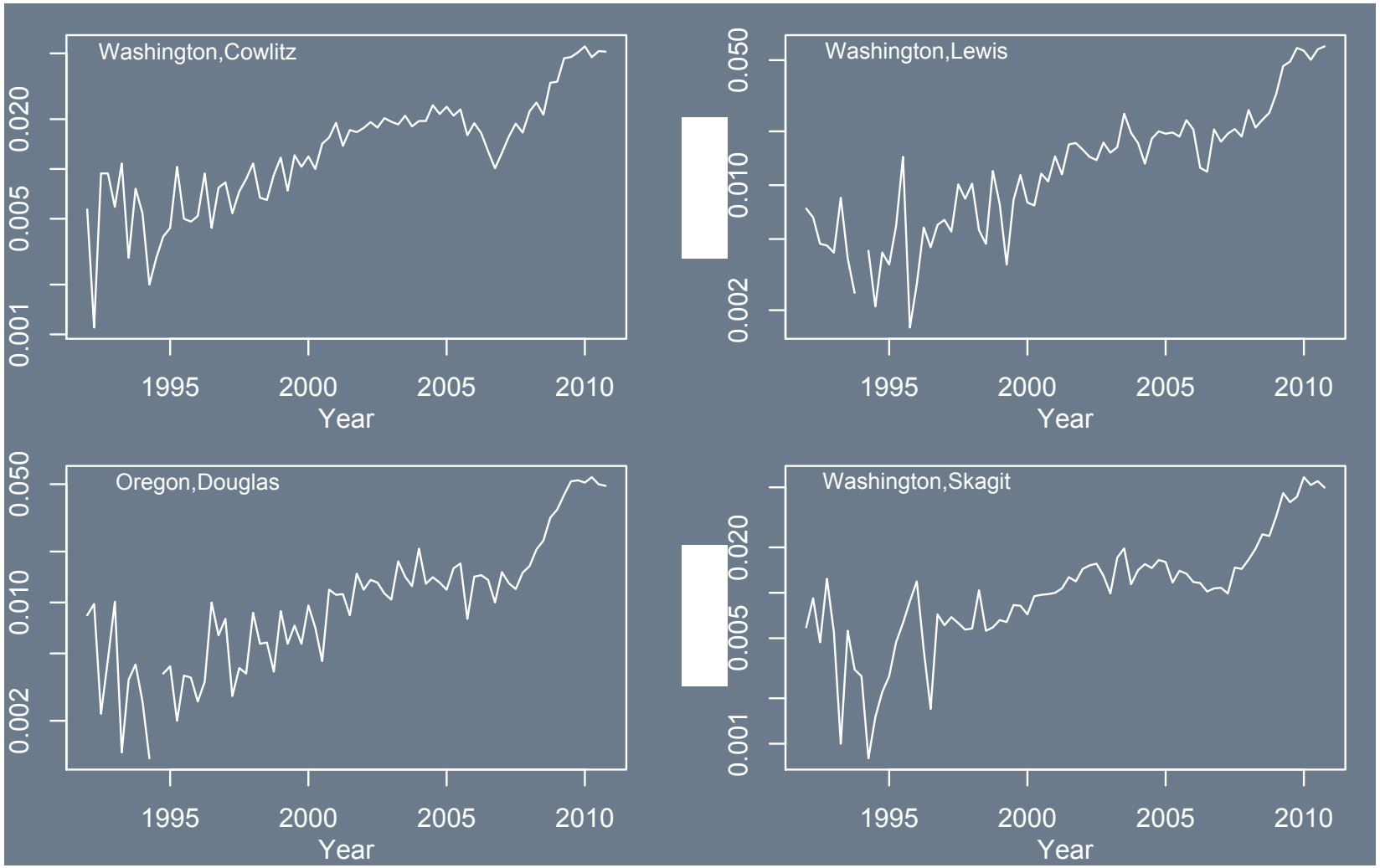
SVD does not know geography!



Wharton
Department of Statistics

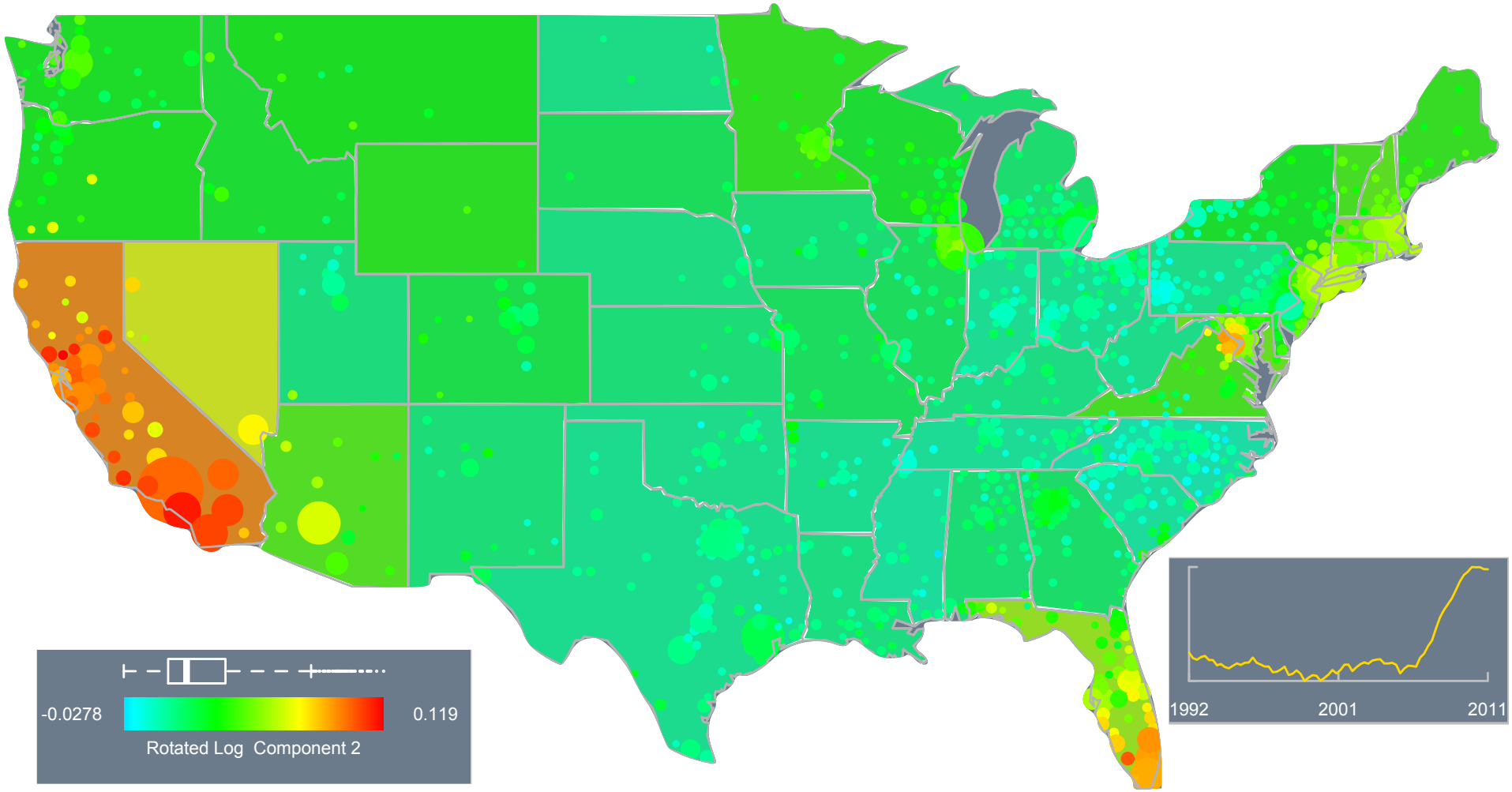
Eg: Long-term Problems

- Defaults in the Northwest US



Second Component

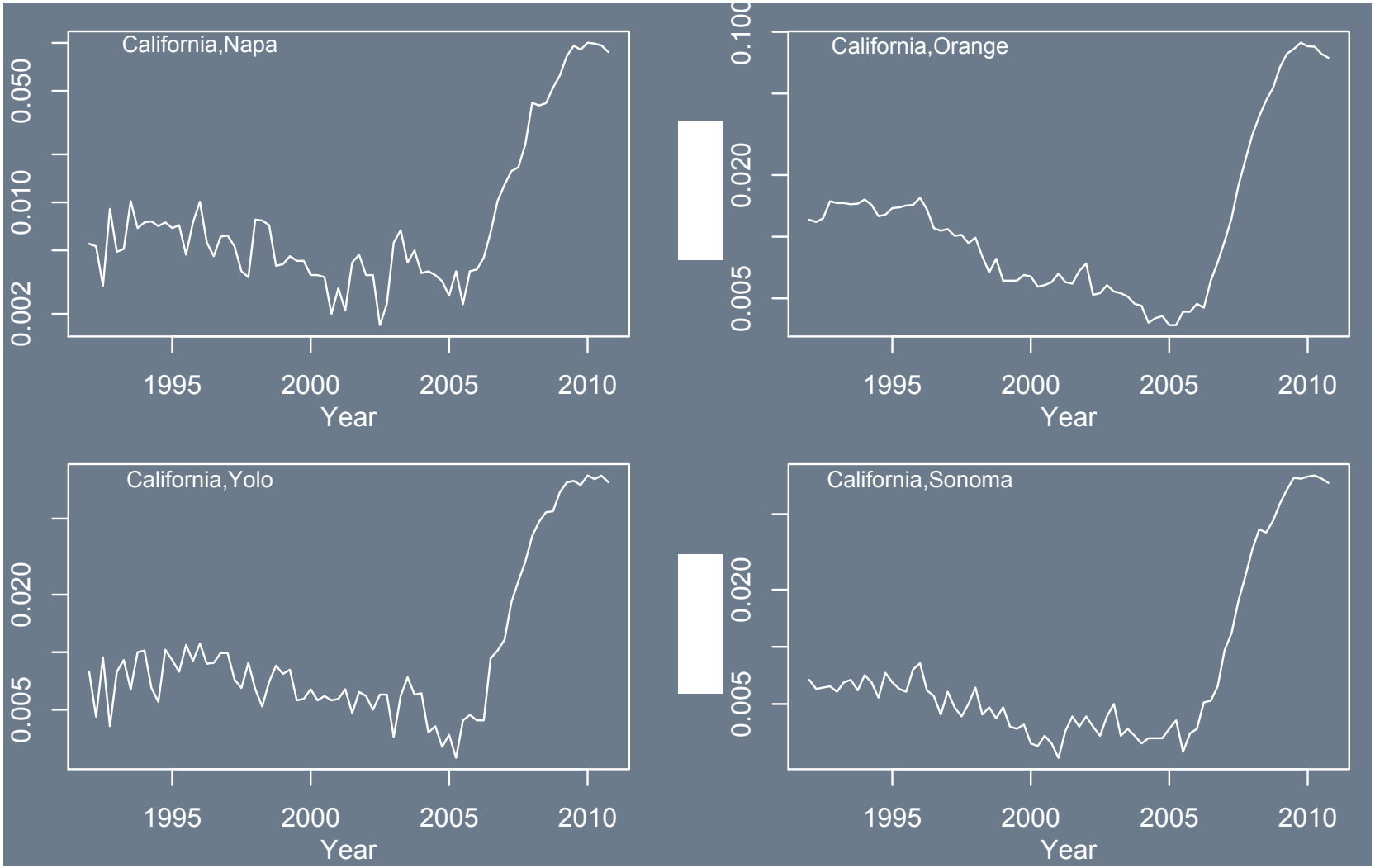
- Recent surge in defaults



Wharton
Department of Statistics

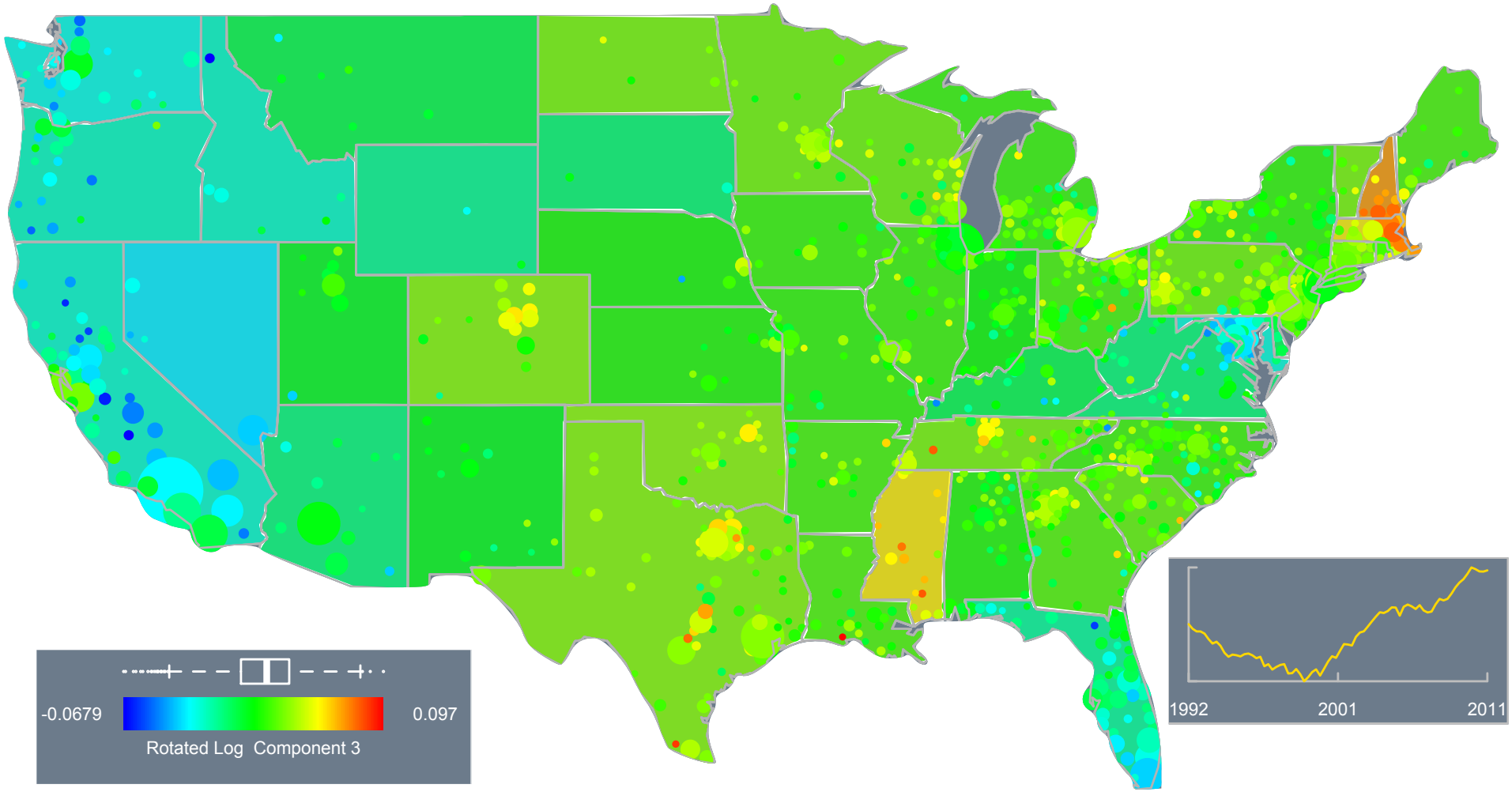
Eg: Recent Surge

- Coastal problems: California, southern Florida



Third Component

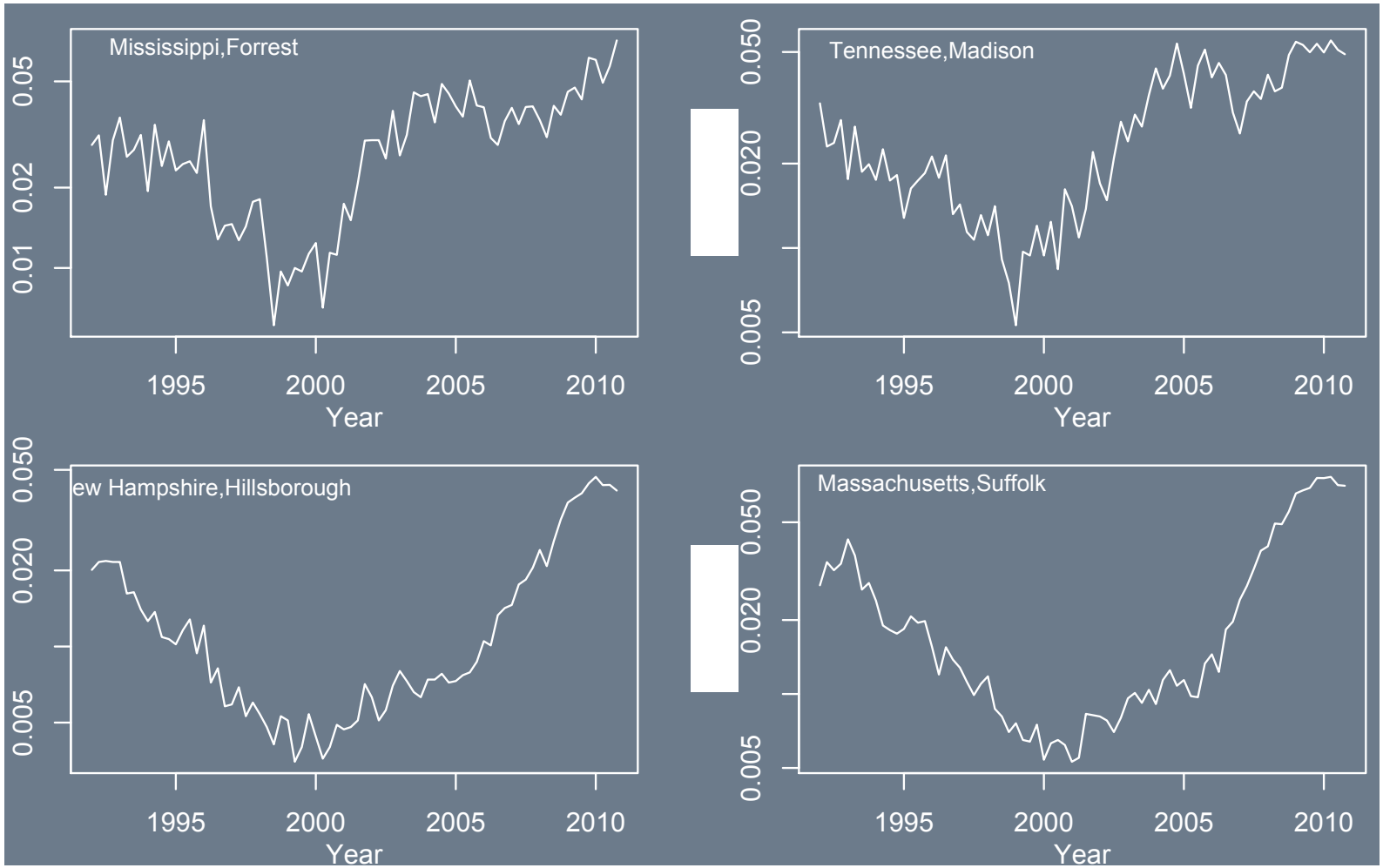
Counties that had been doing well.



Wharton
Department of Statistics

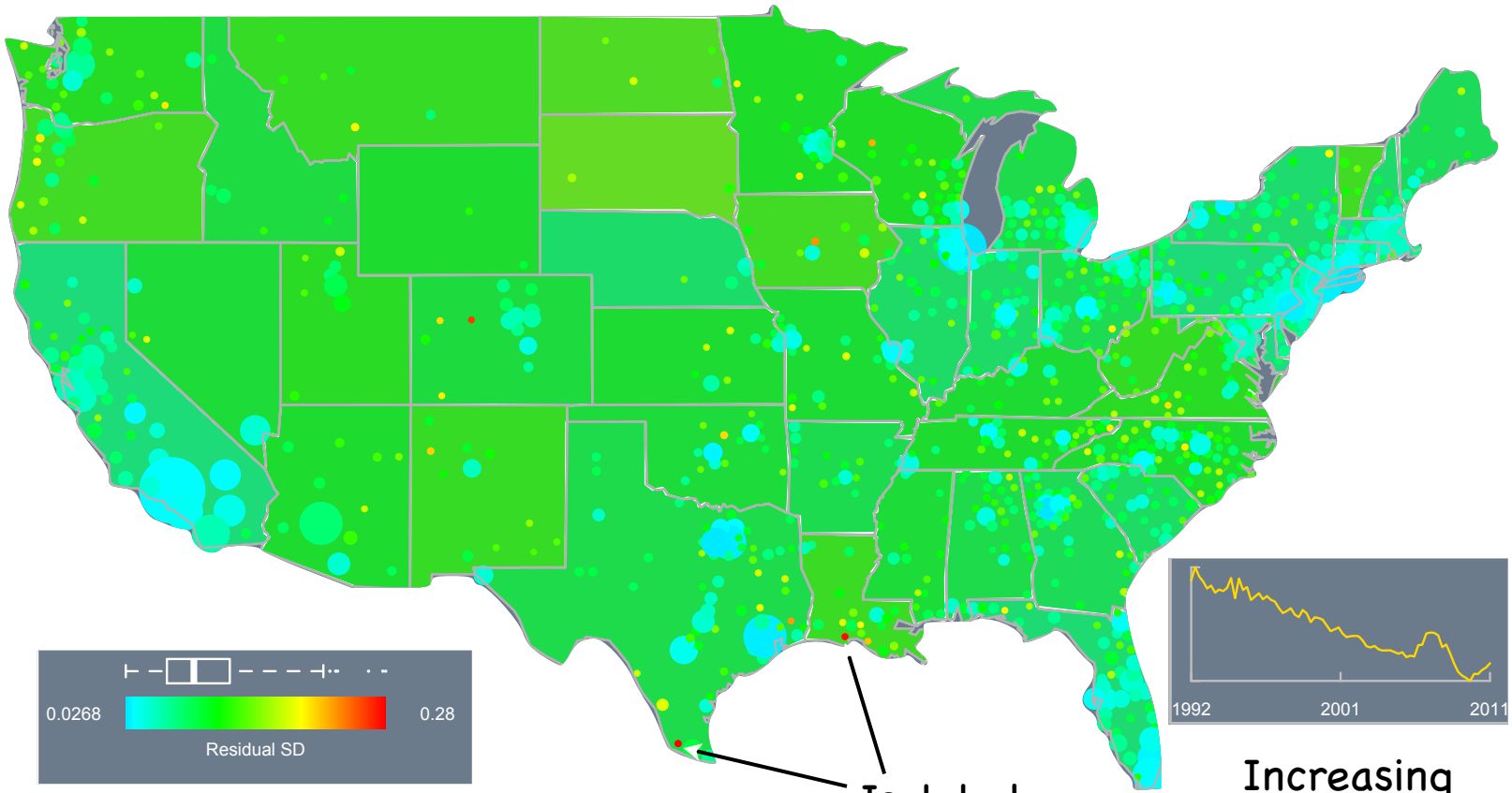
Eg: Were going well

- Some in the South, some in New England



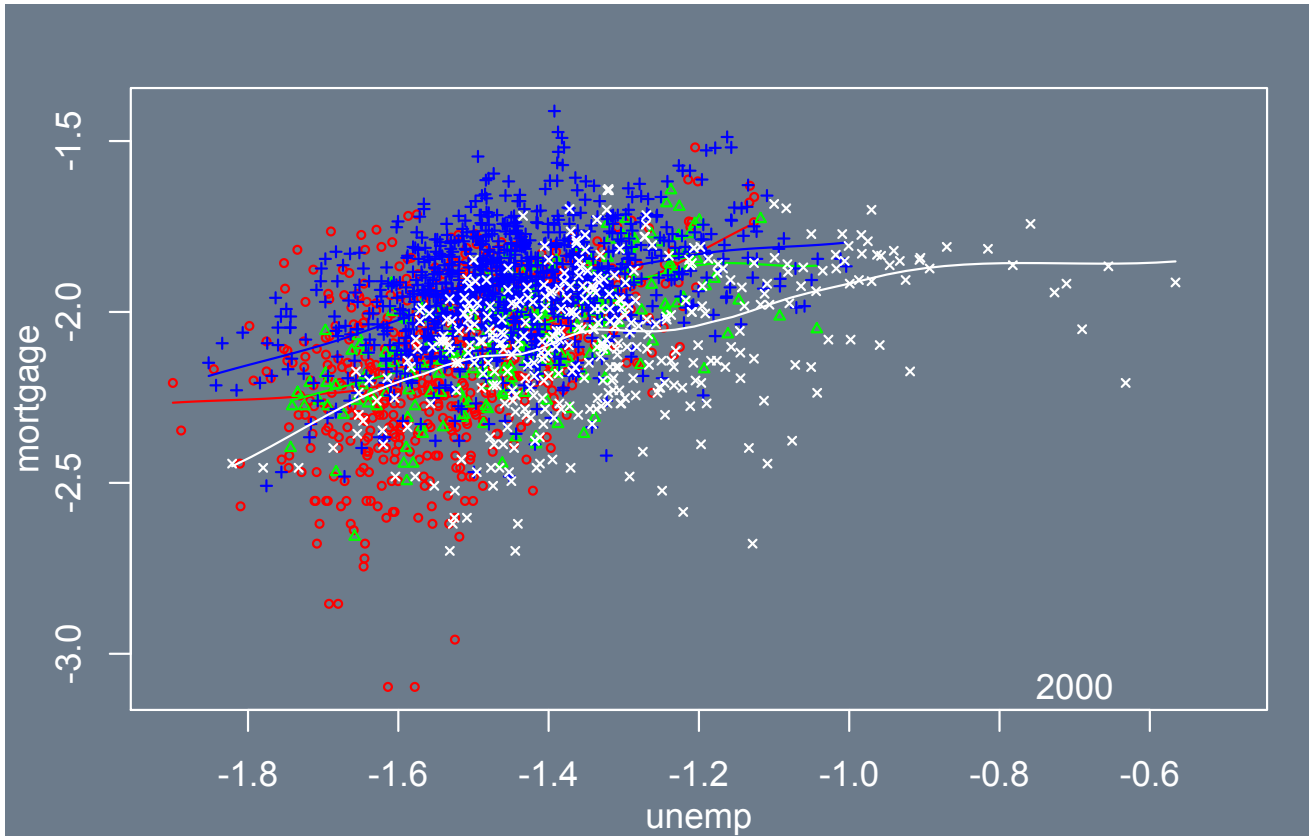
Residual Analysis

- Subtract retained components from data
  - Map shows SD for locations
  - Trend line shows SD over time



Covariates

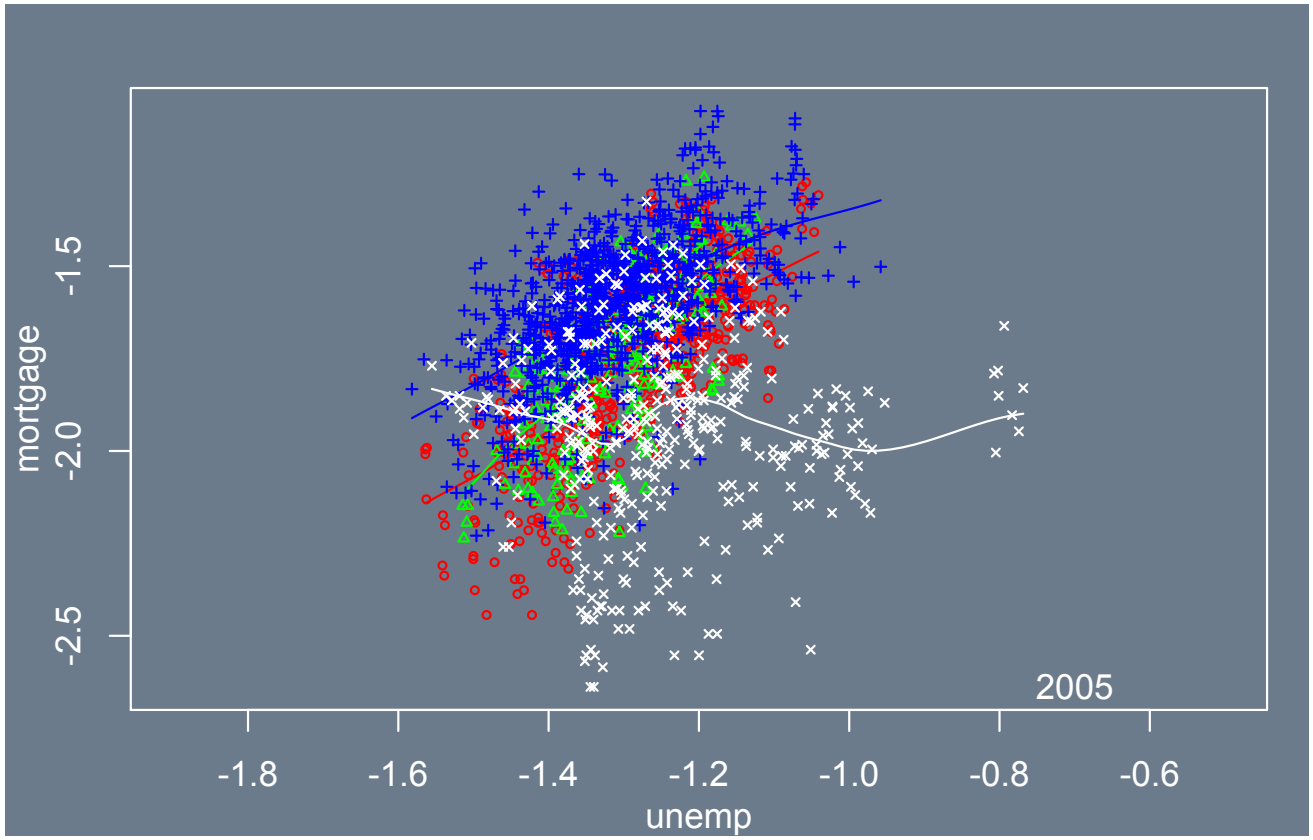
- Covariate effects depend on region
- Regional unemployment
  - Variability changes over time
  - Association with mortgage default changes



N East
South
Midwest
West

Covariates

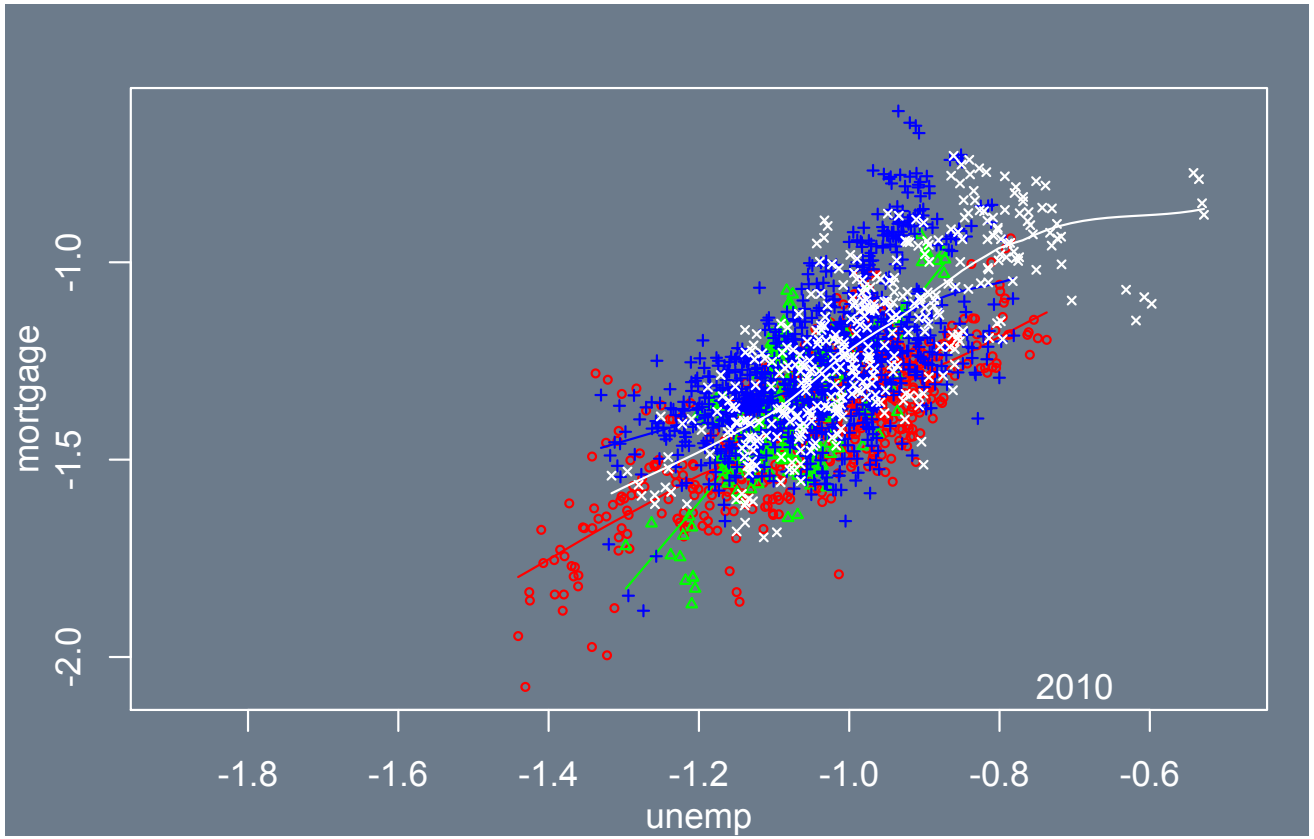
- Covariate effects depend on region
- Regional unemployment
  - Variability changes over time
  - Association with mortgage default changes



N East
South
Midwest
West

Covariates

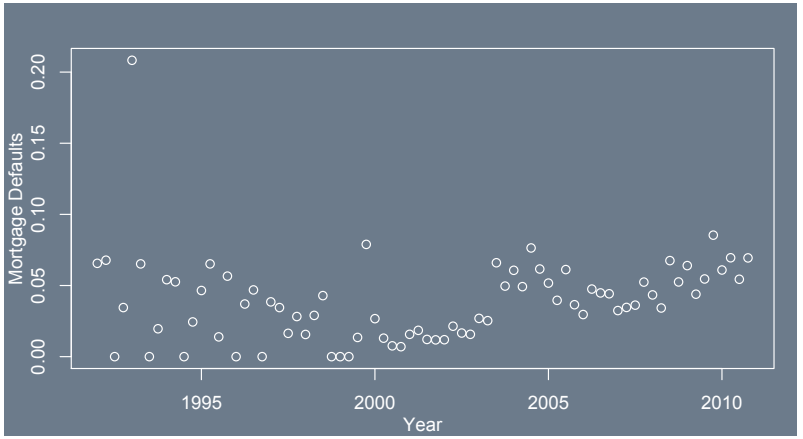
- Covariate effects depend on region
- Regional unemployment
  - Variability changes over time
  - Association with mortgage default changes



N East
South
Midwest
West

Discussion: Spatial Patterns

- General trends
  - Rising defaults
  - Increased spatial concentration
- Timing of mortgage defaults
  - Some have struggled for a long time.
  - Bubble exploded in California, Florida.
  - Less discussed…
    Surge around 2000 in less talked-about locations: Deep South, New England
- Aside
  - SVDs are great for finding outliers!



Exploratory Models

Transition

- Switch type of debt
  from mortgage to cards

- Revolving default rates
  - Data cover most of the US
  - Less political upheaval
- But similar problems remain:
  - Substantial flight to quality in later years
  - Demographic shifts remain relevant
  - Heterogenity in size and characteristics

Local Models

- Consider a reduced-form, economic model
  - Response Y_{cq} = log(default rate)
  - Lags of default rate
  - Economics (unemployment, income)
  - Credit data (utilization, other debt)
- Issues
  - What variables to use in the models?
  - How to obtain an honest standard error?
  - Where's the independence?
- Fit within "slice" of time or space
  - Time: 3,000 counties during a specific quarter
  - Space: Subset of counties over many years

Slices in Time

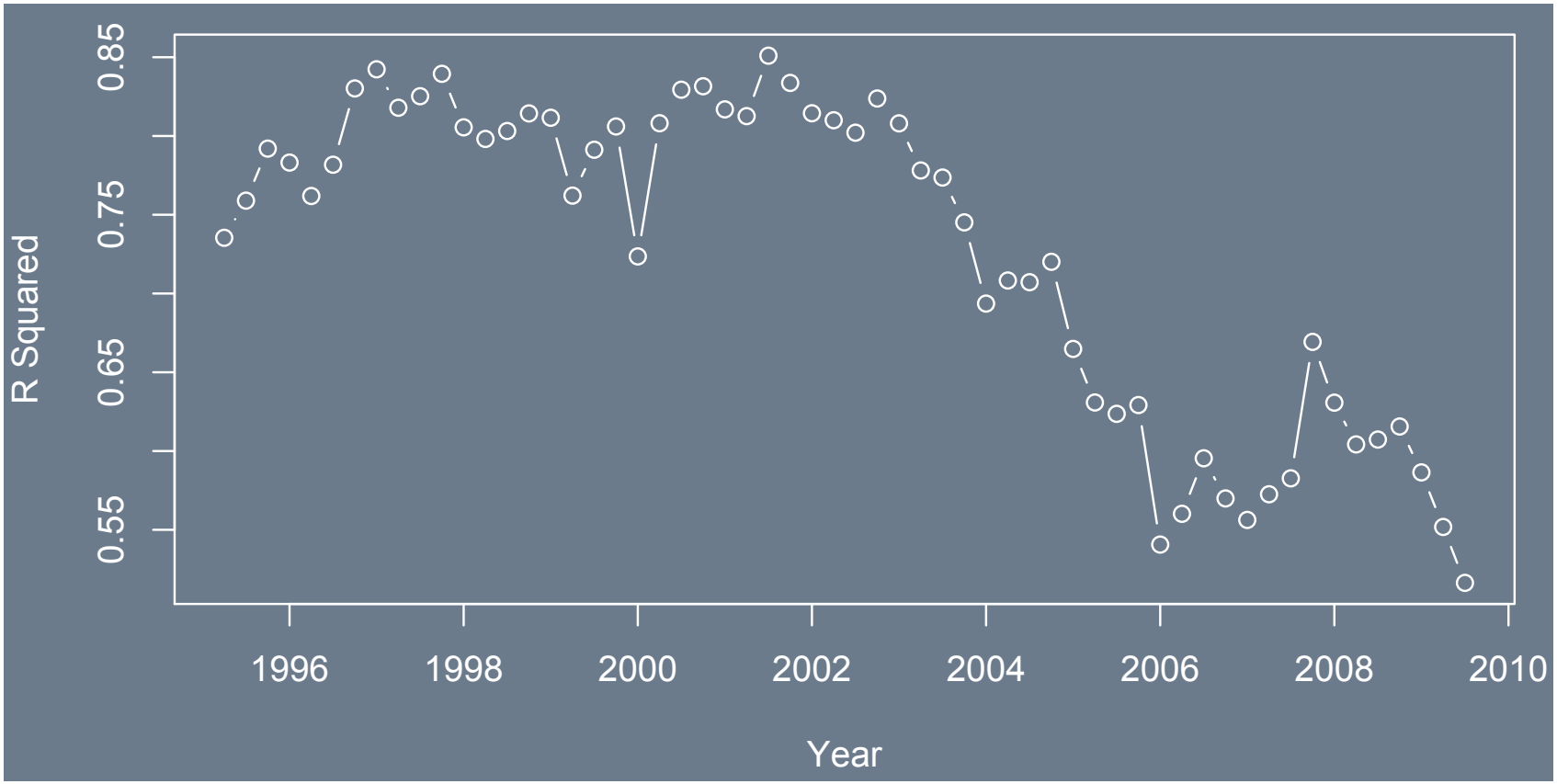
Y_{cq} = log(default rate) in county c, quarter q

Y_{11}		Y_{1q}		Y_{1T}
		...		
Y_{c1}		Y_{cq}		Y_{cT}
		...		
Y_{n1}		Y_{nq}		Y_{nT}

Time

Local-time Models

- Procedure
  - Fit over counties within a given quarter
  - Plot over time, "population drift"
- Goodness-of-fit deteriorates in later years



Local-time Models

- Procedure
  - Fit over counties within a quarter
  - Plot coefficients over time, "population drift"
- Nominal t-statistics identify only lag

t-stat of lag Y_t

Slices in Space

Y_{cq} = log(default rate) in county c, quarter q

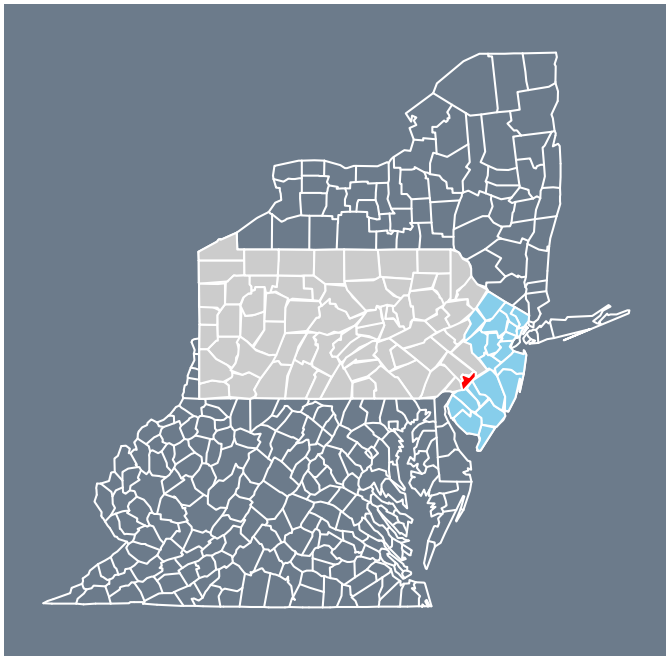
Y_{11}		Y_{1q}		Y_{1T}
		...		
Y_{c1}		Y_{cq}		Y_{cT}
		...		
Y_{n1}		Y_{nq}		Y_{nT}

Time

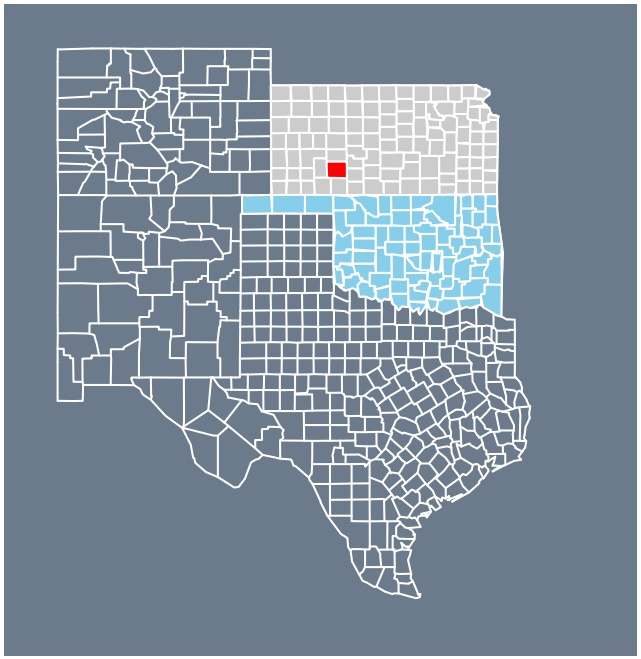
Local-space Models

- Procedure
  - Fit regression model in cluster of counties
  - Measure residual dependence

Urban, densely populated



Philadelphia, PA

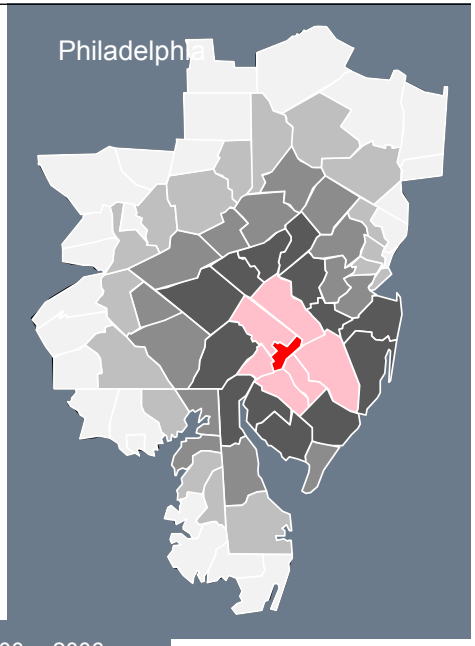


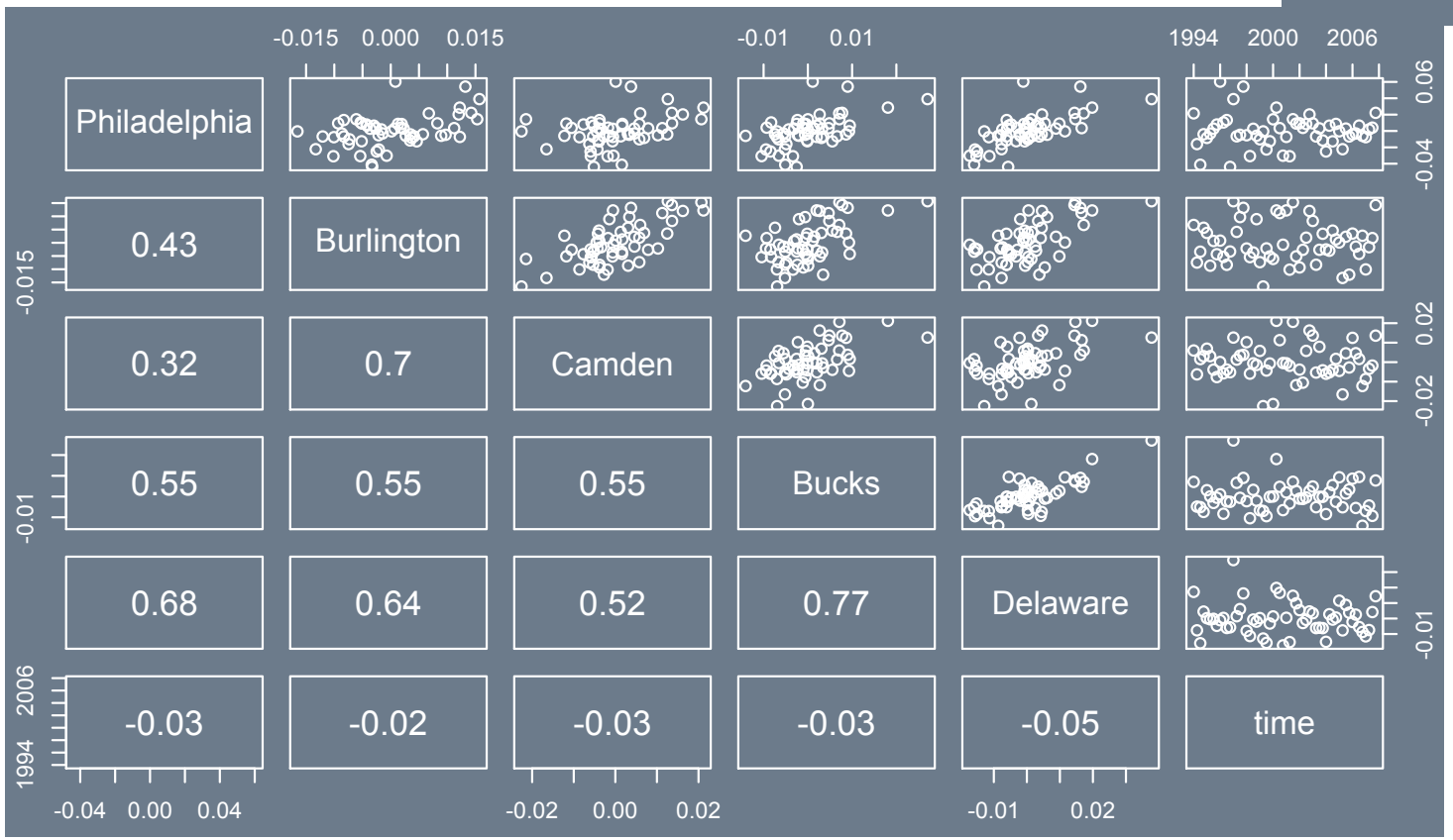
Ford County, KS

Rural, sparsely populated

Urban Models

- Models fit well, $R^2 \approx 80\%$ or more
- Spatial correlations depend on proximity, political boundaries
- No residual autocorrelation

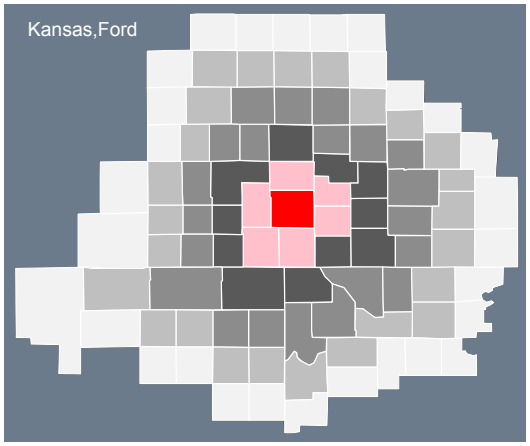


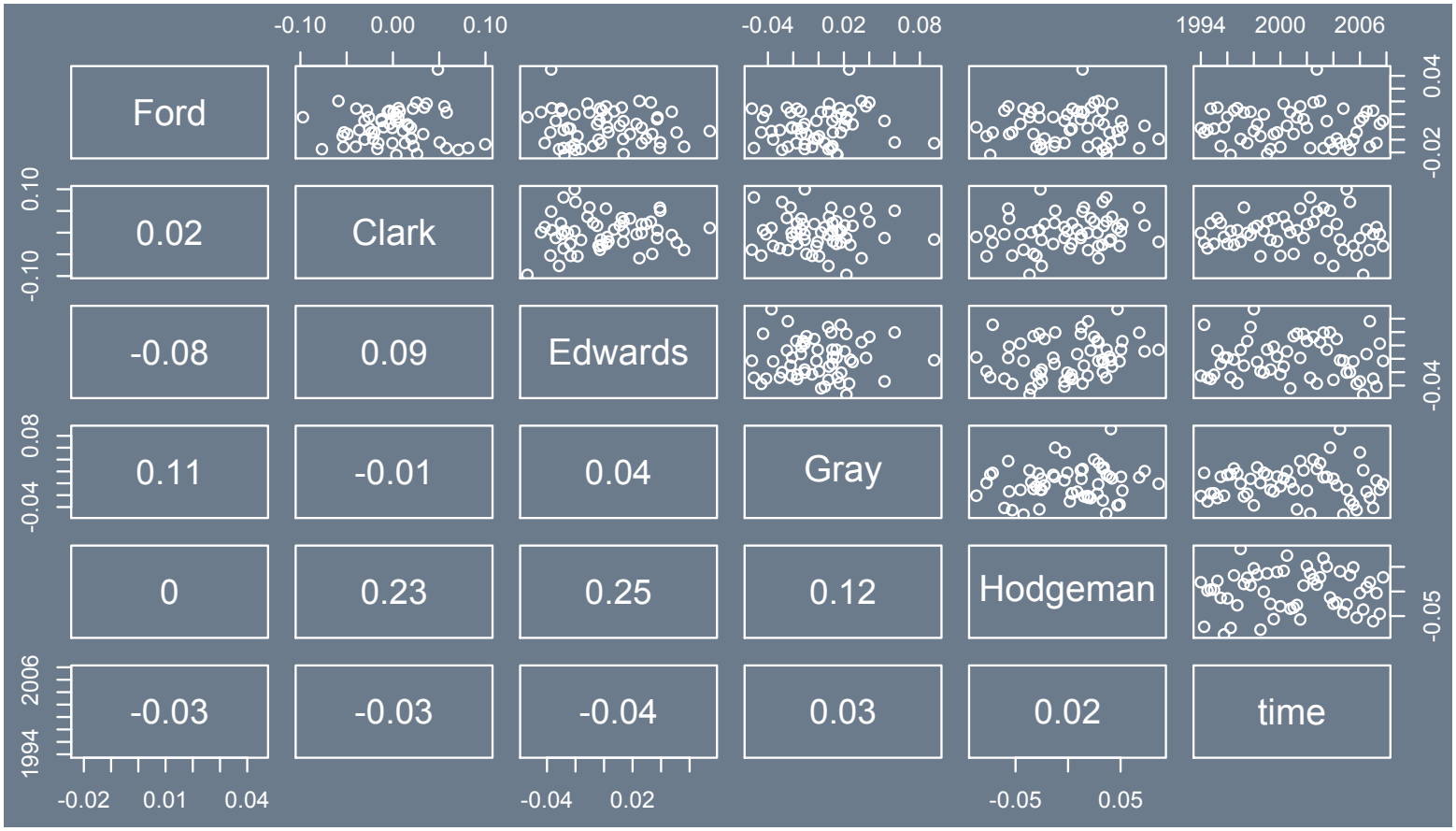


Wharton
Department of Statistics

Rural Models

- Models fit weakly
- Small spatial correlation
- No residual autocorrelation





Lessons from Exploration

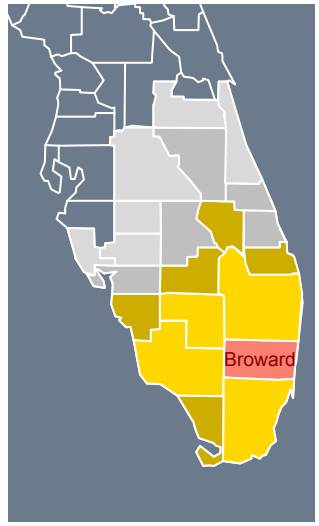
- Over time...
  - Evolving, simple models describe much of the variation in default rates, leaving...
  - Errors appear uncorrelated over time
- Over space...
  - Complex spatial dependence
- Explanatory variables
  - Subtle contribution from local explanatory variables such as income
  - Adjustments for spatial dependence needed to avoid over-fitting

Confirmatory Models

Markov Random Fields

- Idea

  Describe spatial distribution by the collection of conditional distributions



- Conditional independence

  Default rate Y_k in location k depends only on its neighbors N(k),

  $$\{ Y_k \mid Y_m, m \neq k \} = \{ Y_k \mid Y_{N(k)} \}$$

- Gaussian MRF… CAR model

  Covariates model broad structure, with spatial correlations for errors

  $$\{Y_k \mid Y_{N(k)}\} = N(\mu_k + \Sigma W_{km} (Y_m - \mu_m), \sigma_k^2)$$

  where $\mu_k = X_k \beta$

  $W_{kk} = 0$ and $W_{kj} = 0$, j not in N(k)

Conditions for MRF

- Not every set of conditional distributions specifies a valid joint distribution.
- Gaussian MRF (Besag 1974)

  $\{Y_k \mid Y_{N(k)}\} = N(\mu_k + \Sigma\, W_{km}\,(Y_m - \mu_m),\ \sigma_k^2)$

  implies that joint distribution is

  $\{\, Y \,\} = N(\mu,\ (I-W)^{-1}S^2)$

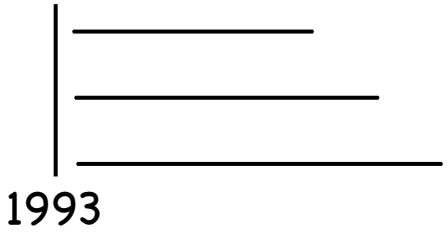
  for $S = \mathrm{diag}(\sigma_k)$.
- Implications
  - (I-W) must be positive definite
  - $(I-W)^{-1}S^2$ must be symmetric
- Spatial pattern matrix (Cressie et al 2005)
  - Obtain spatial correlation parameter γ by setting $\sigma_k^2 = \tau^2/n_k$ and $W_{km} = (n_m/n_k)^{1/2}$

Is CAR right?

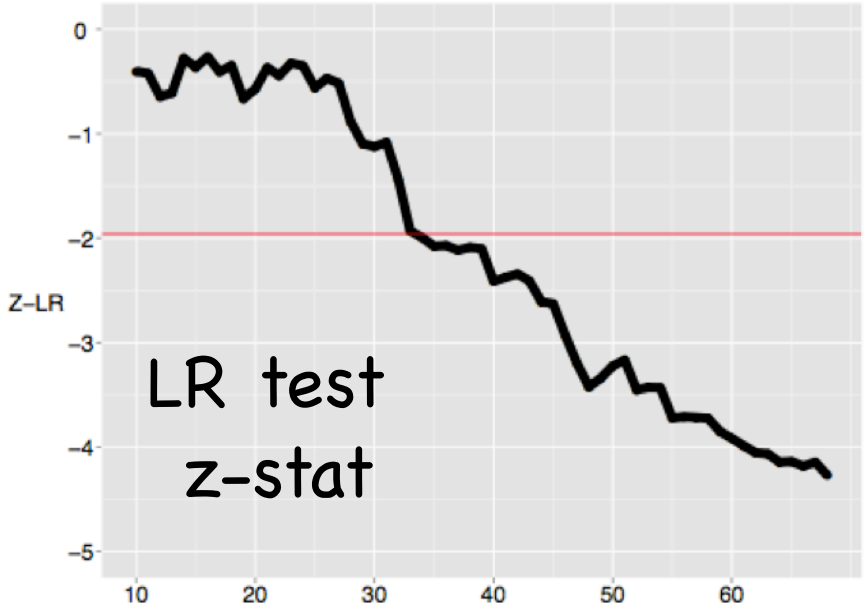
- What is the residual structure?
  - Likelihood ratio test between nested models
  - Equal correlation model is a CAR model with lots of neighbors, N(k) = all indices but k.
- Use all 3,000 counties
  - Logit response p/(1-p)
  - Var(logit) ≈ 1/(np(1-p)) determines σ_k
  - Covariates include local unemployment, poverty...
- Compare error specifications
  - CAR with single layer neighborhood
  - Equal-correlation model

Results of CAR Test

- Recursive estimation
  - Use history from 1993 forward
  - Evaluate model at 'current' time



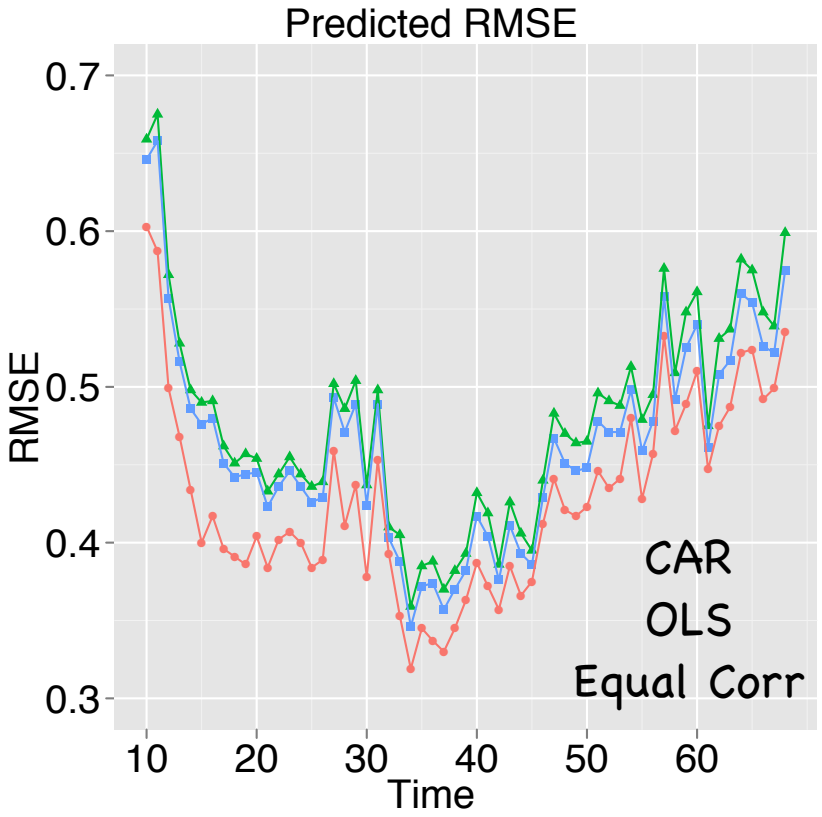
- Test CAR vs equal corr
  - Equal correlation with smaller, broader corr dominates in later years
  - Hints at national latent variable



Prediction Results

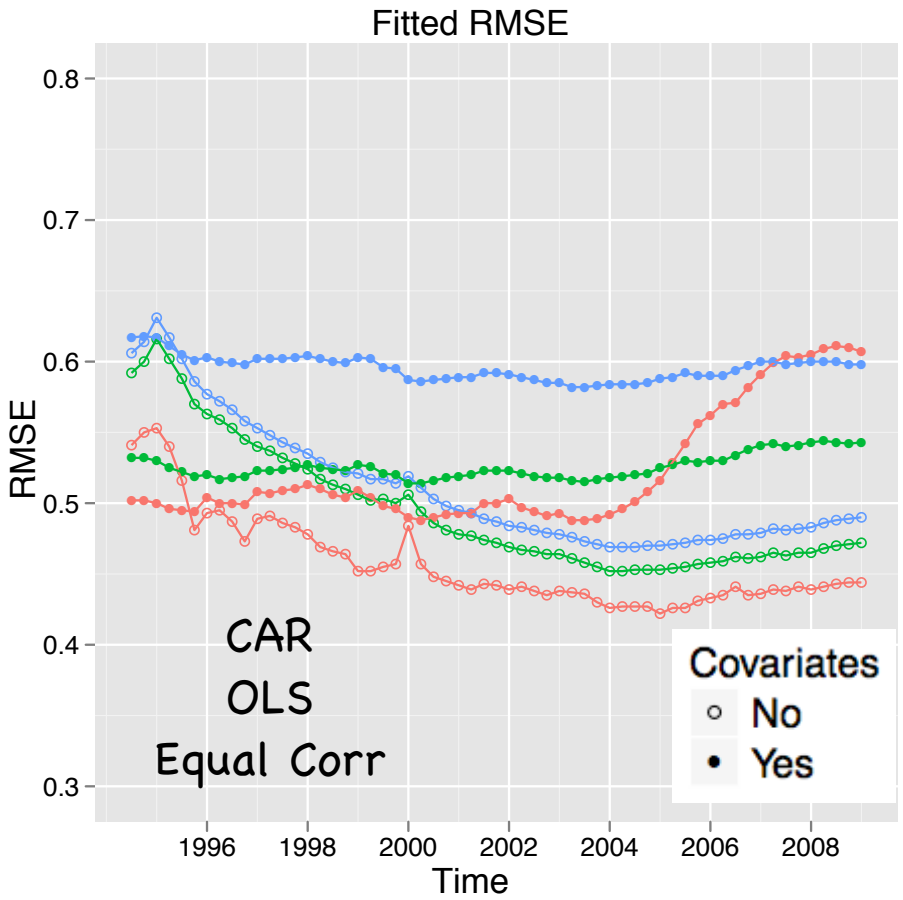
- Comparison of Prediction MSE
  - OLS
  - CAR (with small neighborhood)
  - Equal correlations (large neighborhood)

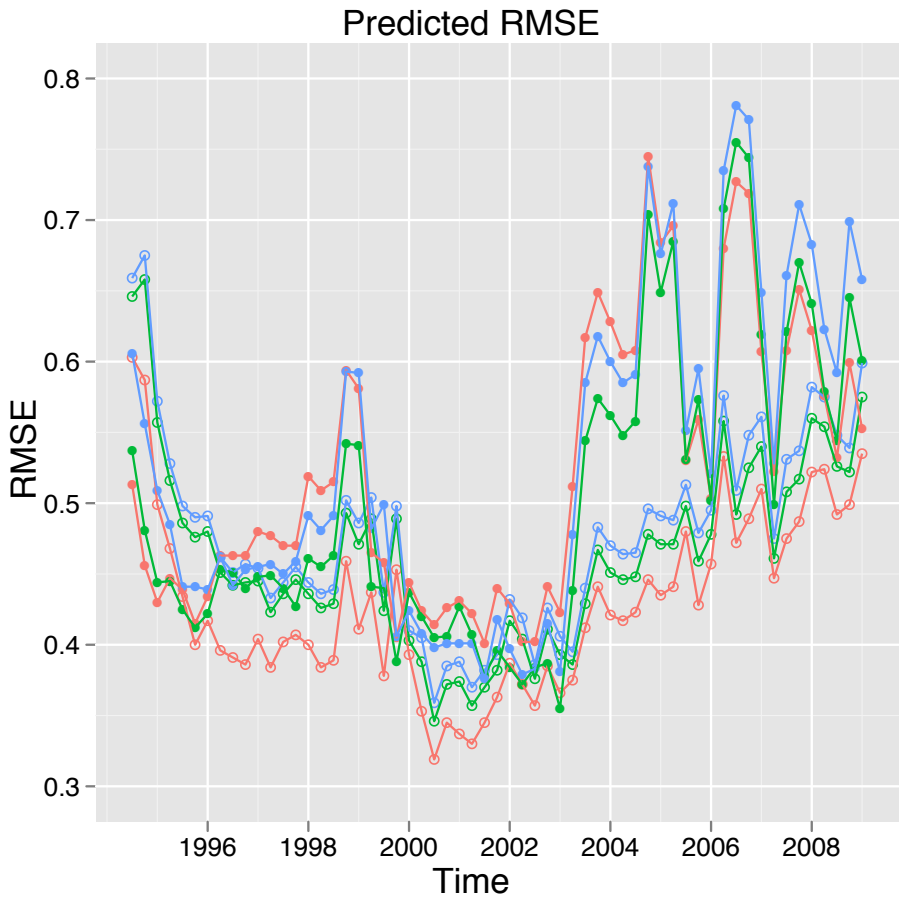
- Results
  - Only lags of default as predictors
  - Equal correlation model has smallest MSE
  - Model performance gets worse as data accumulate



Covariates

- Less accurate with explanatory variables





Summary & Discussion

Key Points

- Substantial spatial correlations
  - Don't have 3,000 independent observations
    Cannot claim 3,000 x 80 = 240,000 d.f. in models
  - Over-stated claims of significant inference
- Time-specific, location-specific patterns
  - Population drift over sub-models
  - Complex models most likely overfit

- Possible remedies?
  - Better economic modeling at consumer level
  - Portfolio view of individual consumer debt
  - Expensive to develop and maintain

Directions in Modeling

- Adaptive, data-driven strategies
- Hierarchical Bayesian models
  - Dirichlet process priors via Markov chain MC
  - Scalable? Have not been able to scale to US.
- Large scale data mining using regression
  - Fast selection from 100,000's of variables
  - Predictive, but not "explanatory"
- Latent process models
  - High dimension hidden Markov models
  - SVD of massive matrices (50,000,000 cases)
  - Currently requires stable training set

Comments

- Epidemic models
- Surface diffusion model
- Multi-mode factor analysis (covariates)
- Voxel correlation analysis

Thanks for coming...

Papers will eventually appear at
stat.wharton.upenn.edu/~stine