Models for Information Integration: Case Studies and Emerging Principles

D.H. Judson
(w/ Carole L. Popoff
Dennis Culhane
Fritz Scheuren)
December, 2005

Purpose of This Talk

- Place our work in the context of the history and philosophy of social statistics
- Criticize official statistics and present case studies of sometimes feeble attempts to provide better information
- Define “information integration”
- Describe first steps toward “information integration” theory
- Peek over the next hill

The problem with “Official” Statistics (from the locality point of view)

- Not fast enough
  - Lag from collection to dissemination
- Not local enough
  - Geographic reporting insufficient for local needs
- Not granular enough
  - Insufficient detail for important demographic groups
- Not integrated enough
  - Differing definitions, time reference, etc.

A tale of two paradigms

- The Sample Survey Method
  - Quetelet and Laplace
    - "l'homme moyenne" (the average man)
  - *The Halcyon Days*
    - 20th century successes of the survey method
  - *The Decline and Fall of the Survey Empire*
    - Declining response (ongoing surveys)
    - The "brutal environment" of telephone surveys
    - "Angry refusal" (field reports)
  - *The Empire Strikes Back*
    - American Community Survey
      - Large rolling survey, multi-mode, sampling for NR
    - Bigger hammer (at the cost of "rolling" data)

A tale of two paradigms, cont.

- The "Administrative Records Method"
  - Administrative records: Collections of already-existing data
    - Used for some other purpose
  - Techniques that use AR databases
    - Direct use
    - Modeling frameworks emerging
  - Examples:
    - O.D. Duncan: Voting(!) in Ancient Greece
    - Graunt: Bills of Mortality
    - Cohort-component population estimates,
    - Geographic Information Systems

What is “Information Integration”?

- Research in the UK: Practice is ahead of theory!
- The “consulting” point of view: 90% today better than 97% tomorrow?
- My point of view: Yes, but that 90% had better be “statistically principled”

What is “Information Integration”?

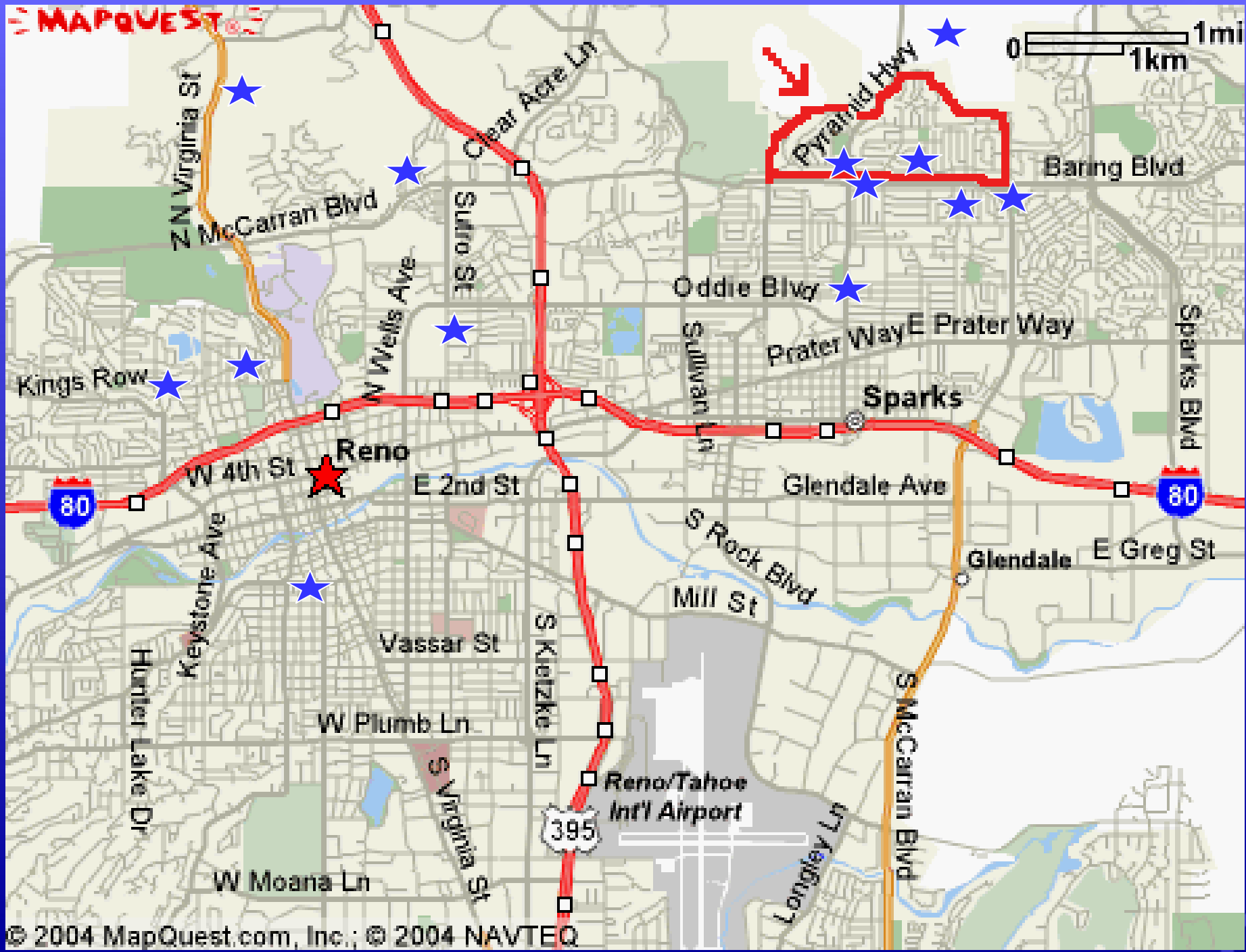
- Information integration is the process of using multiple datasets in concert to construct *statistical estimands* for the purpose of answering questions about those estimands.
- E.g.,
  - What is the ethnicity-specific unemployment rate in Stockport in July, 2003?
  - How many uninsured persons are there in Washoe County, Nevada in 2004?
  - Where is there more daycare demand than supply?

Case studies (Shadowboxing in the dark)

- Locating an airline hub
  - # of machine shops
  - # of employees in SIC 372 (aircraft parts manuf.)
  - Unemployment rate
  - N of departures/arrivals
  - Annual average temperature (?)
- This is not information “integration”

Case studies (Wandering in a fog)

- Locating a daycare center
  - Place a grid over the city
  - Determine:
    - # of children 0-6 with dual income families for each tract in the city
    - Latest bureau of licensure daycare slots and their address, geocoded to census tract
  - For each cell on map:
    - “Demand” = gravity-model weighted sum of children 0-6
    - “Supply” = gravity-model weighted sum of slots
    - Desirability = total “demand”/ “supply”


MAPQUEST
0
1mi
1km
N Virginia St
McCarran Blvd
Clear Acre Ln
Pyramid Hwy
Baring Blvd
Sutro St
Wells Ave
Oddie Blvd
Prater Way
E Prater Way
Sullivan Ln
Sparks Blvd
Kings Row
Sparks
Reno
W 4th St
E 2nd St
Glendale Ave
80
Keystone Ave
S Rock Blvd
Glendale
E Greg St
Mill St
Vassar St
S Kietzke Ln
Hunter Lake Dr
W Plumb Ln
Reno/Tahoe
Int'l Airport
395
S Virginia St
S McCarran Blvd
Longley Ln
W Moana Ln
© 2004 MapQuest.com, Inc.; © 2004 NAVTEQ

Case studies (Stumbling toward the light)

- Evaluating program participants' outcomes with unemployment insurance (UI) wage records versus a 13 week follow-up survey
  - Program completers get a follow-up survey
  - Performance measure = weekly wage
  - Performance standard: Avg. weekly wage of respondents > $xxx
- Problem: Can UI records replace the survey?
  - What "UI weekly wage" ≈ "survey weekly wage"?
- Solution: For linked records:
  - Regress UI weekly wage on survey weekly wage
  - Express performance standards on transformed scale

Case studies (Eschewing obfuscation)

- Determining the number of uninsured children at the county level
  - Problem:
    - Estimates not yet provided by the Fed statistical system (SCHIP expansion: 1997)
  - Solution:
    - Develop own county-level estimates
  - Result:
    - Attempt to integrate state-level survey data (CPS) with county-level cohort-component estimates
  - Combine two separate easier-to-construct quantities: Population estimates and uninsurance model

ARSH (Age, Race, Sex, and Hispanic Origin) synthetic estimation

$$\hat{x}_{a,r,s,h} = P_{a,r,s,h} \cdot \hat{\mu}_{a,r,s,h}$$

$$a \in \{0,...,85+\}$$
$$r \in \{W, B, API, AIAN\}$$
$$s \in \{M, F\}$$
$$h \in \{H, \sim H\}$$

Definition: A "cell" is a specific age, race, sex and Hispanic origin combination; e.g. 15-year old white, male, Hispanic. Within each cell we calculate a proportion uninsured.

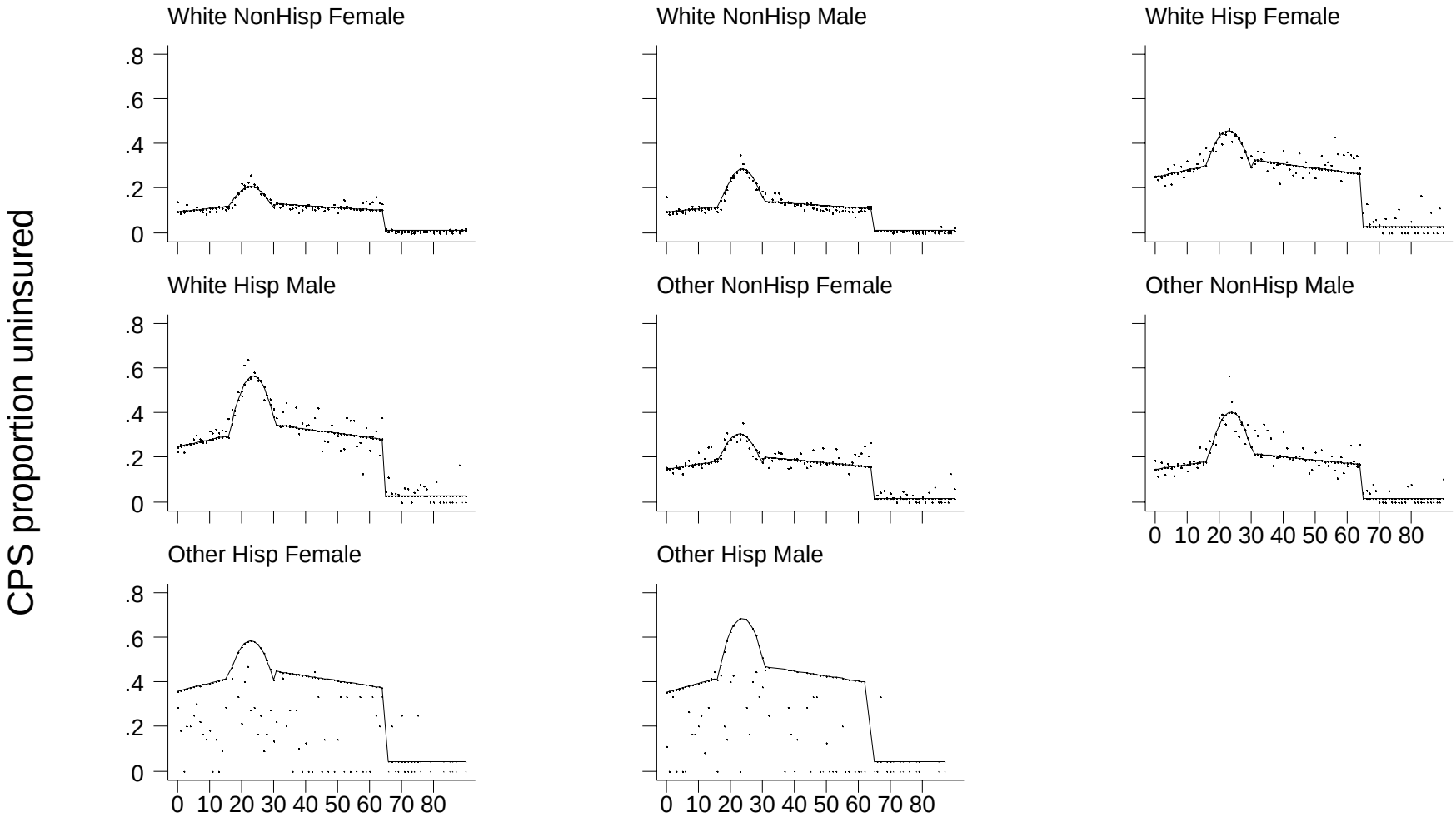
We fit our model by individual record; our estimation is by ***cell***.

$x_{a,r,s,h}$: Number of uninsured in a,r,s,h in county

$P_{a,r,s,h}$: From county-level cohort-component population model

$\mu_{a,,r,s,h}$: From national microdata based uninsurance model (ML logistic regression with a,r,s,h as RHS variables)

CPS prop uninsured by age: Obtained & predicted values (MODEL 2)



Graphs by Wh/Other by Hisp/Non by M/F

Questions to be answered by the information integrator

- Older/better vs. Newer/worse?
  - Data vintage is (surprisingly) important
- Count vs. Model?
  - AR data never seem to match population of interest
- Certainty vs. Uncertainty?
  - As yet unsolved problem
- Statistical Matching vs. Record Linkage?
  - Yesterday: Technology almost exists
  - Today: Technology exists!
- Suppression vs. Detail?

Attacking the problem: Three principles

- Recognize that the estimand *exists*, but is *not* always observed directly
  - (Latent variable principle)
- Recognize that *none* of the bits of data contributing to the estimand are without error or uncertainty
  - (Uncertainty principle)
- *Model* the relationship between estimand and data sources, with weight (inversely) proportional to uncertainty of data source
  - (Modeling principle)

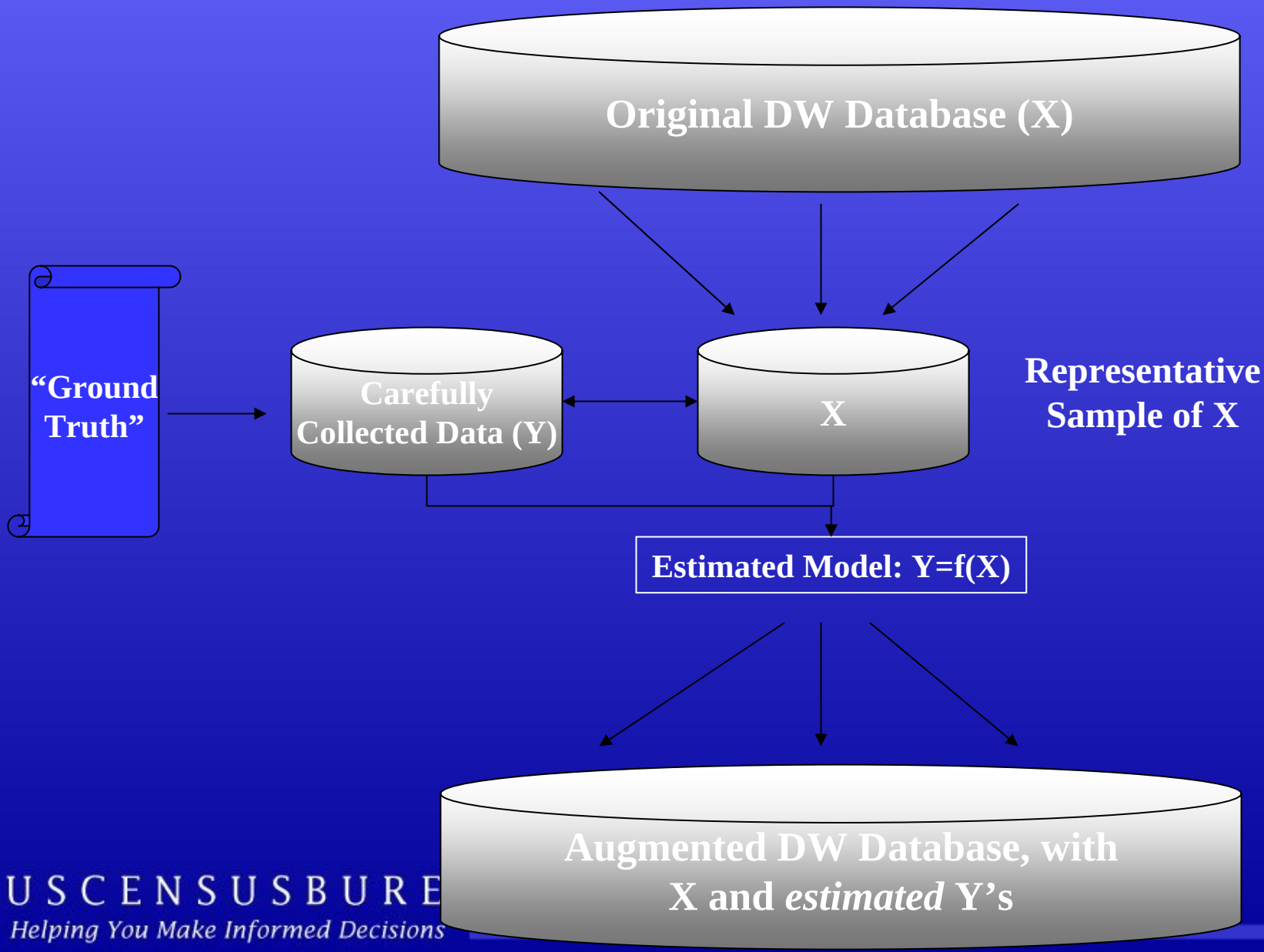
Outstanding challenges (Lacunae in current theory)

- Representing uncertainty
- Adapting to differential temporal/spatial reference
- Sampling and design weights with linked survey/census/administrative data
- Record linkage and statistical matching error
- Covariance between estimation components
- Spatial concentration of change

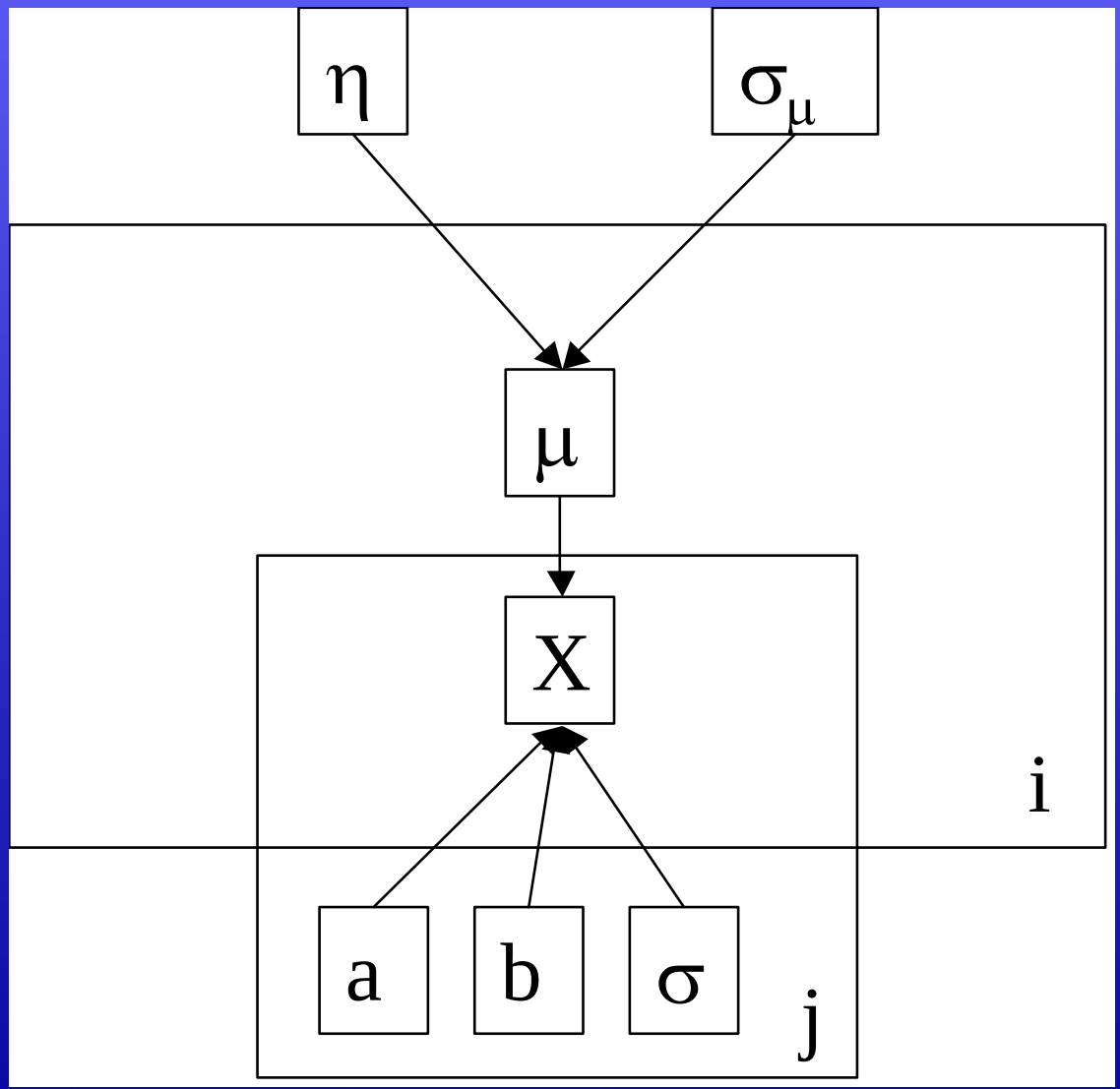
Lacuna #1

- How to represent uncertainty in the "administrative records method"
  - Sampling variance, model variance, and ***procedure variance***?
  - Fisher/Gee (2004) general model:
    - Incorporate sampling variance where it is known;
    - Incorporate model variance based on specification;
    - Procedure variance: Bayesian estimate of uncertainty.

An (Early) Model for "Borrowing Strength" (Judson, Bye)



A More Sophisticated Model (Fisher/Gee)



i=ith area;

j=jth indicator variable;

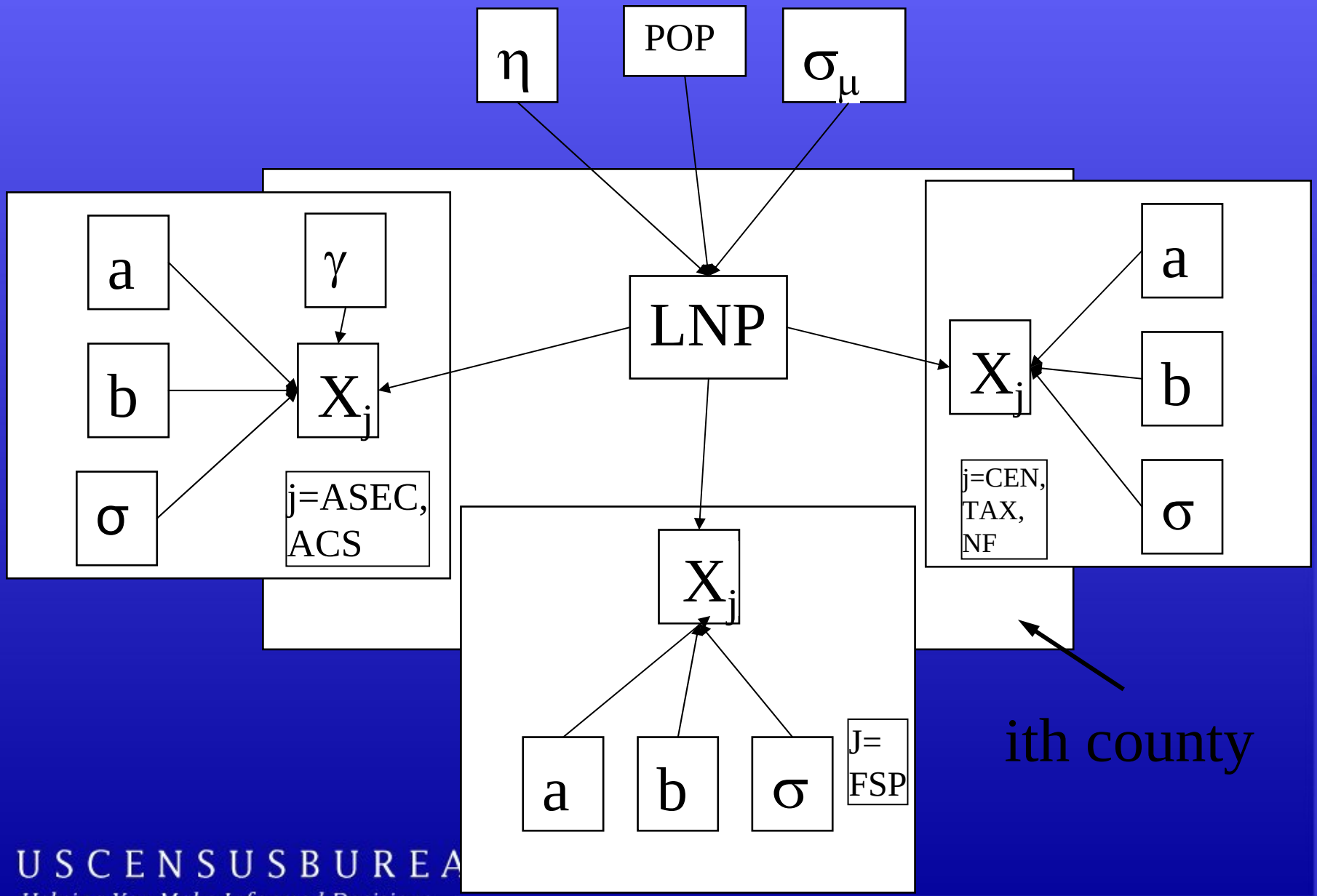
μ=true (latent) value of estimand;

η=mean (latent) value of estimand for all areas;

σ_μ=S.D. of estimand for all areas;

a,b,σ=regression relationship of jth indicator to ith area.

A More Sophisticated Model (Concrete Example)



Lacuna #2

- Adapting to differential temporal/spatial reference (and adapt***ive*** temporal/spatial reference)
  - Information decays over time
  - The population of objects (people, housing units, areas) is changing (at different rates)
  - “Spatial and temporal slippage” (the difference between the reference dates/places of the data and the estimand of interest)
  - “Ontological slippage” (Barr; the difference between one representation of the object of interest and another)

Temporal Information Decay (Stuart)

Level III: Global Parameters:

α: file inclusion probs

λ: Global migration rate (decay)

Level II:Person captured in AR file?

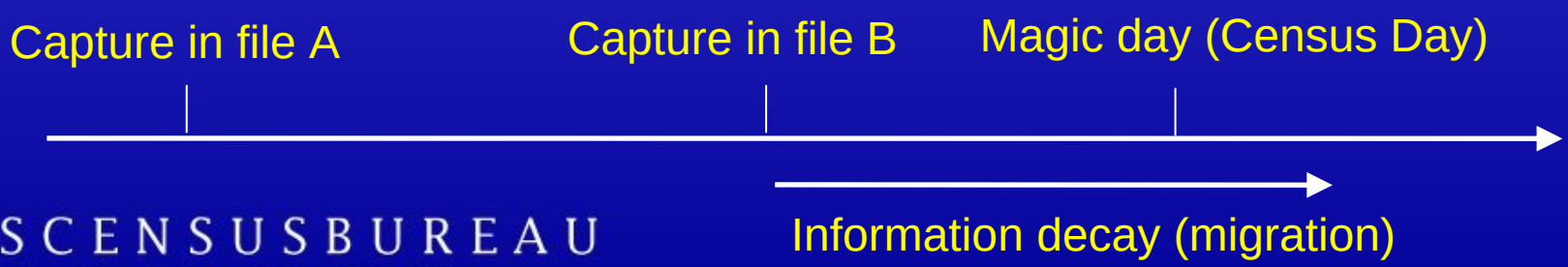
w_i and y_i: capture probs

Level II: Individual-level migration

t_{0i}, t_{1i} : Start and end time

Level I:Individual-level observation

z_i: indicator of capture



Lacuna #3

- Sampling and design weights with linked survey/census/administrative data
  - We know how to analyze survey data with weights; but what about linked data?
  - Proposed weight (Chesher and Nesheim, 2004):

$$\text{Weight for linked pair of records} = \frac{1}{P[\text{Incl } 1]\, P[\text{Incl } 2 \mid \text{Incl } 1]}$$

Lacuna #4

- Record linkage and statistical matching error
  - Known effect – biasing effect on inference
  - False links vs. false nonlinks tradeoff
  - Posterior probabilities:
    - P[records true link | linkage comparison]
    - P[records false nonlink | linkage comparison]
    - Provide factors correcting for linkage error
  - Elaborating on Chesher and Nesheim's weight:

$$\text{Weight for linked pair of records} = \frac{P[\text{True Link} \mid \text{Linkage comparison}]}{P[\text{Incl}\,1]P[\text{Incl}\,2 \mid \text{Incl}\,1]}\left[\frac{1}{P[\text{False Nonlink} \mid \text{Linkage comparison}}\right]$$

Lacuna #5

Covariances between components

- Functional relationships induce covariance
  - Soil example
  - Demographic analysis example
- Spatial autocorrelation induces covariance
- Relationships across levels of geography induce covariance between estimation components

➢ Solutions?

- Error propogation modeling (Heuvelink, 1998)?
- Multilevel modeling?

Lacuna #6

Spatial concentration of change

- Ian Cope, ONS: Address changes tend to be concentrated (Manchester vs. Westminster)
- Today: Small area estimation methods:
  - (demographic) miss change
  - (statistical) smooth out change

➢ Solutions?

Speculations on the future (Like the present, only longer)

- More "data" at finer levels of geographic detail
  - Commensurate increase in re-identification concerns
- More precise legal framework for use
- Emerging "novel" uses
  - AR applications (Imputation, Pop. estimates)
  - Eligibility models (ACS vs. Food Stamps, taxes)
  - Quarterly Workforce Indicators
- *A breakthrough of paradigmatic importance is waiting to happen*

Contact Information

Dean H. Judson

U.S. Census Bureau
Washington, DC 20233
Phone: 301-763-2057
Email: dean.h.judson@census.gov

USCENSUSBUREAU