Statistical Issues Arising from Integrating National Databases

Roger D. Peng, PhD

Department of Biostatistics

Johns Hopkins Bloomberg School of Public Health

simplystatistics.org

@simplystats

Integrating National Databases

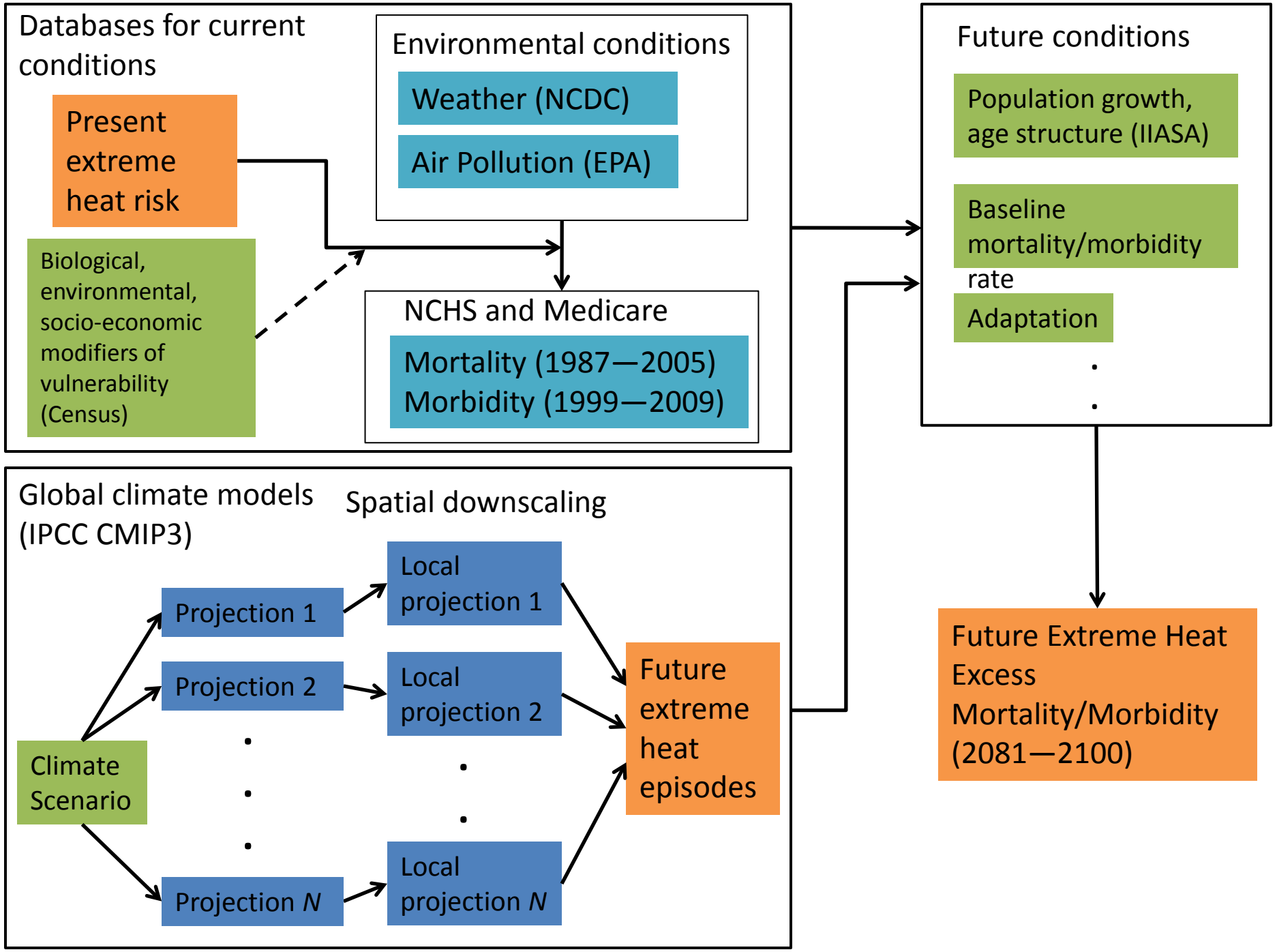
- Many national and sub-national databases exist containing health and environmental exposure information
- Linking/integrating these databases to study health is an efficient use of existing resources
- Allows one to cover very large populations
  - NMMAPS: ~110 million people
  - MCAPS: ~12 million Medicare enrollees
- Allows for the study of heterogeneity across regions, time periods
- Address questions that would be too expensive to address in a single original study

National Databases: Some Examples

- Medicare Part A, B
- CDC/NCHS Mortality files
- CDC/NCHS Nat'l Health Interview Survey
- CDC/NCHS NHANES
- CDC Behavioral Risk Factor Surveillance Survey
- Medicaid
- Census / American Housing Survey

National Databases: Some Examples

- PCMDI: CMIP-3, CMIP-5
  - General circulation models simulating global climate
  - Output from multiple models; downscaled products
- NOAA/NCDC/NCEP: Present-day weather data
- EPA/AQS Air pollution monitoring networks
- EPA National Emissions Inventory


Databases for current conditions
Present extreme heat risk
Biological, environmental, socio-economic modifiers of vulnerability (Census)
Environmental conditions
Weather (NCDC)
Air Pollution (EPA)
NCHS and Medicare
Mortality (1987—2005)
Morbidity (1999—2009)
Future conditions
Population growth, age structure (IIASA)
Baseline mortality/morbidity rate
Adaptation
Global climate models (IPCC CMIP3)
Spatial downscaling
Climate Scenario
Projection 1
Projection 2
Projection N
Local projection 1
Local projection 2
Local projection N
Future extreme heat episodes
Future Extreme Heat Excess Mortality/Morbidity (2081—2100)

General Challenges

- Data are used “off-label”
  - Health data not designed to be linked with environmental exposure data (Medicare is billing information)
  - Environmental data not designed to be used in health studies
  - Scientific questions may be driven by the data available
- Health outcomes tend to be blunt
  - Hospitalizations, mortality
  - More subtle outcomes may not be measured/reliable

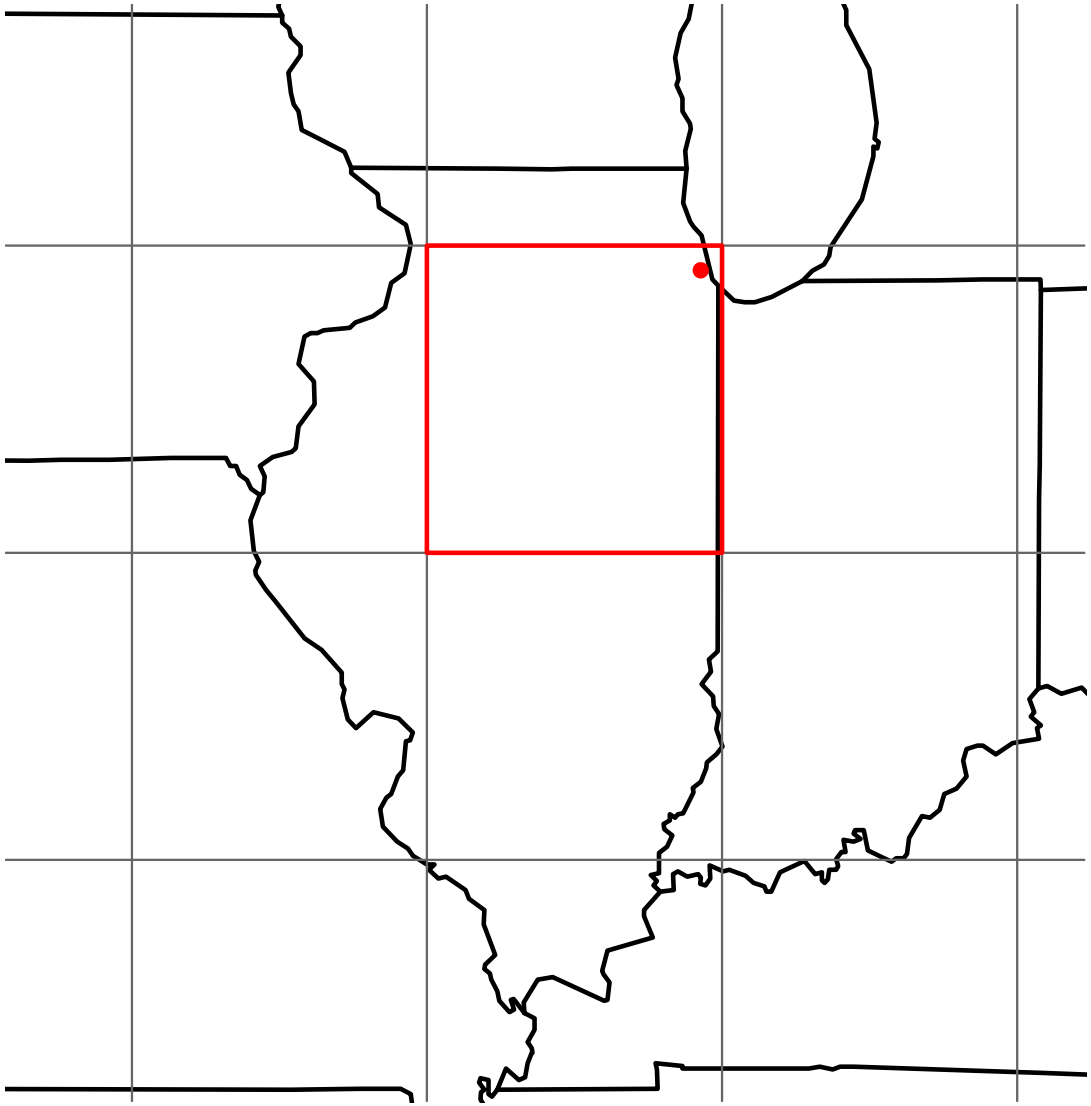
Statistical Challenges

- Temporal misalignment
  - Scales of variation over time can be mismatched
  - e.g. Monthly health outcome vs. hourly temperature
- Spatial misalignment
  - Spatial scales of variation can be mismatched
  - e.g. County-level health outcome vs. point-level air pollution monitor
  - e.g. Zip-code level health outcome vs. grid-cell level temperature data
  - Linkage requires a model, implicit or explicit
- Missing data/Measurement error
  - No opportunity to “go back”; WYSIWYG
  - Health data may have coding errors; lab problems
  - Missingness pattern may be good for one application but bad for another (e.g. daily particulate matter)

Statistical Challenges

- Temporal and spatial misalignment
  - Magnitude of problem depends on the temporal/spatial variability of the process being studied
  - May lead to bias and/or underestimation of uncertainty in regression models of health outcomes
- Measurement error/Missing data
  - Pattern of missingness (informative or not)
  - Nature of measurement error (classical, Berkson)
  - May lead to bias and/or underestimation of uncertainty

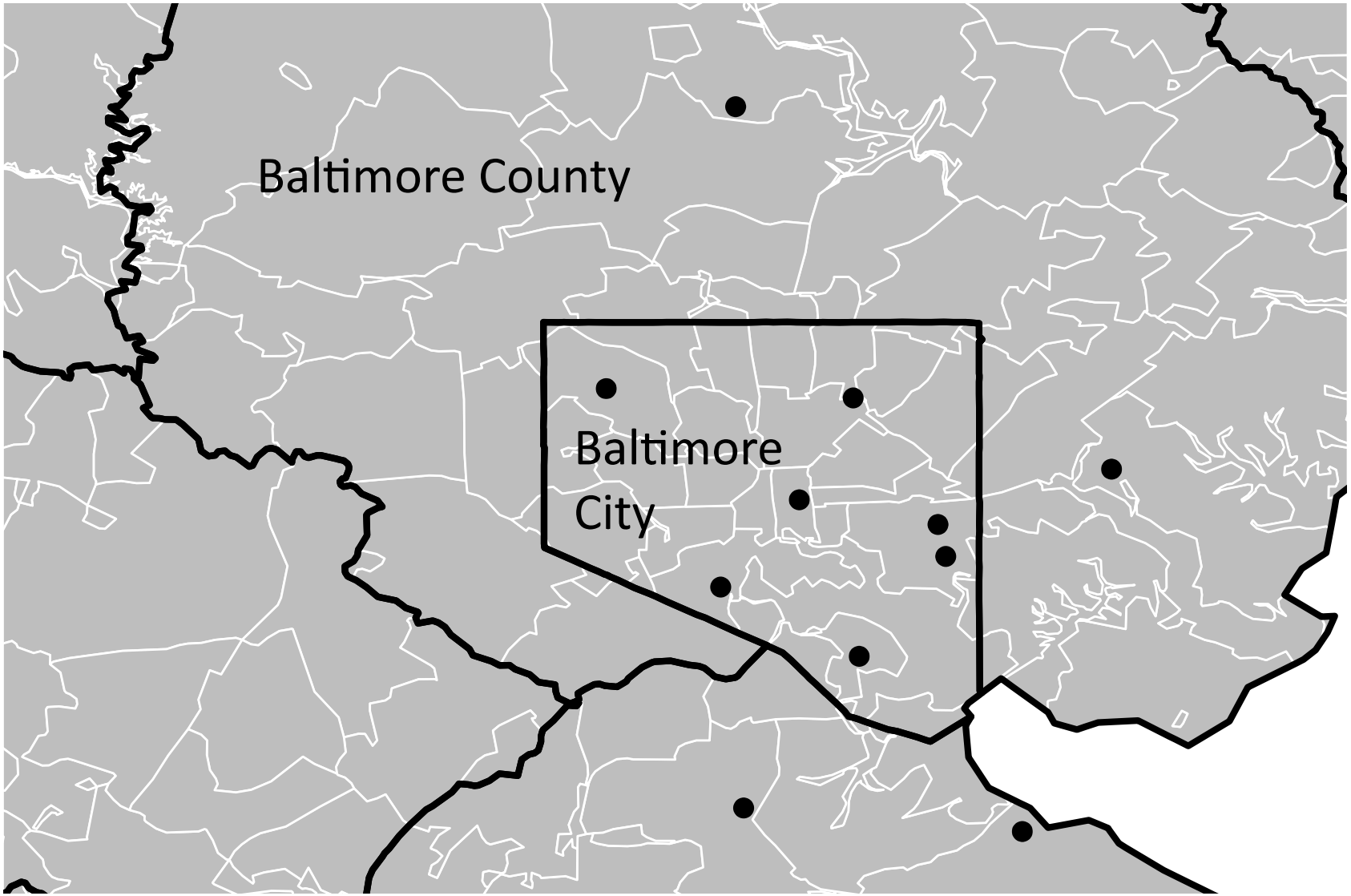
Spatial Misalignment



Peng, *et al*. 2009 *EHP*

GFDL CM 2.0

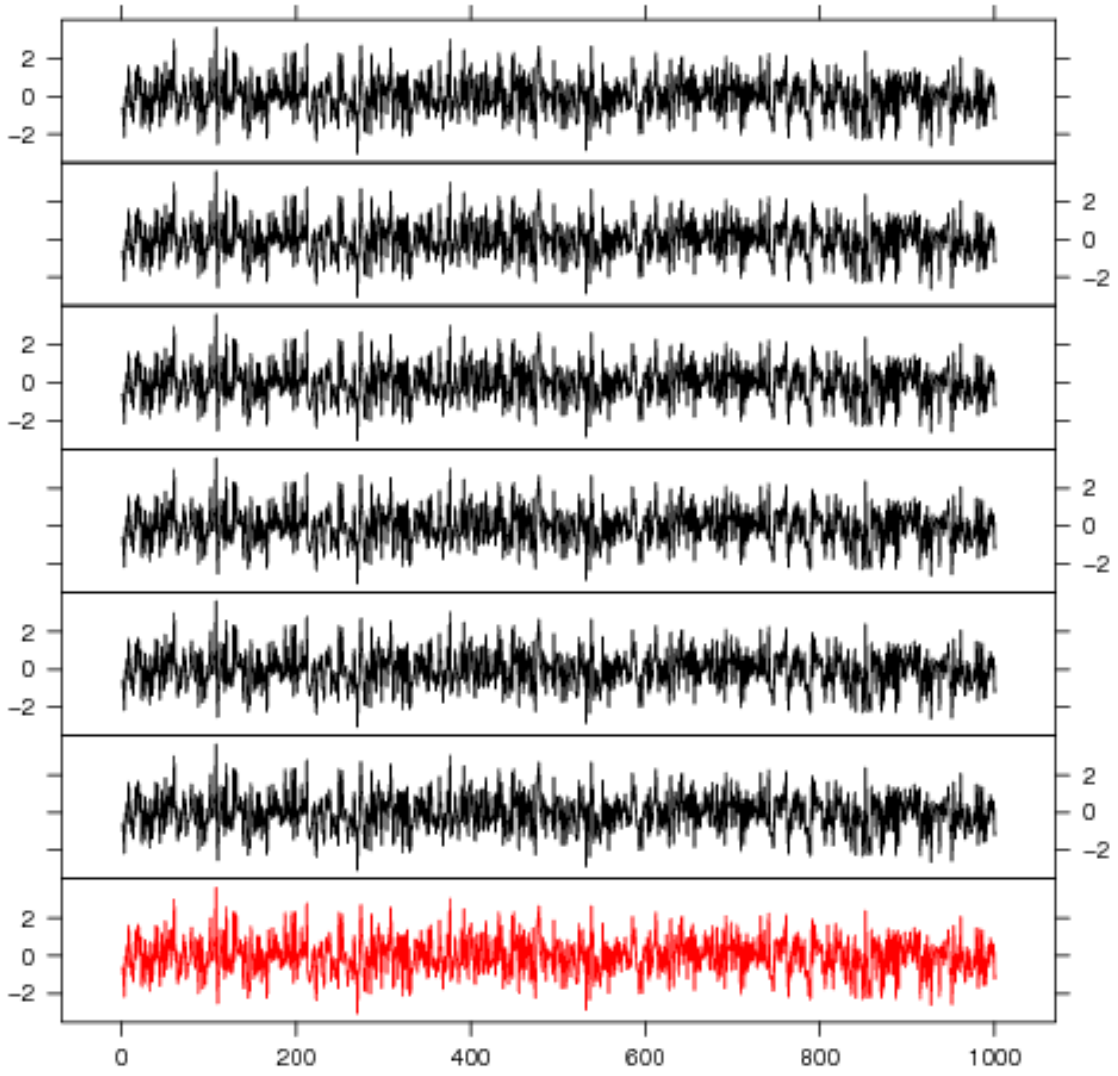
Spatial Misalignment



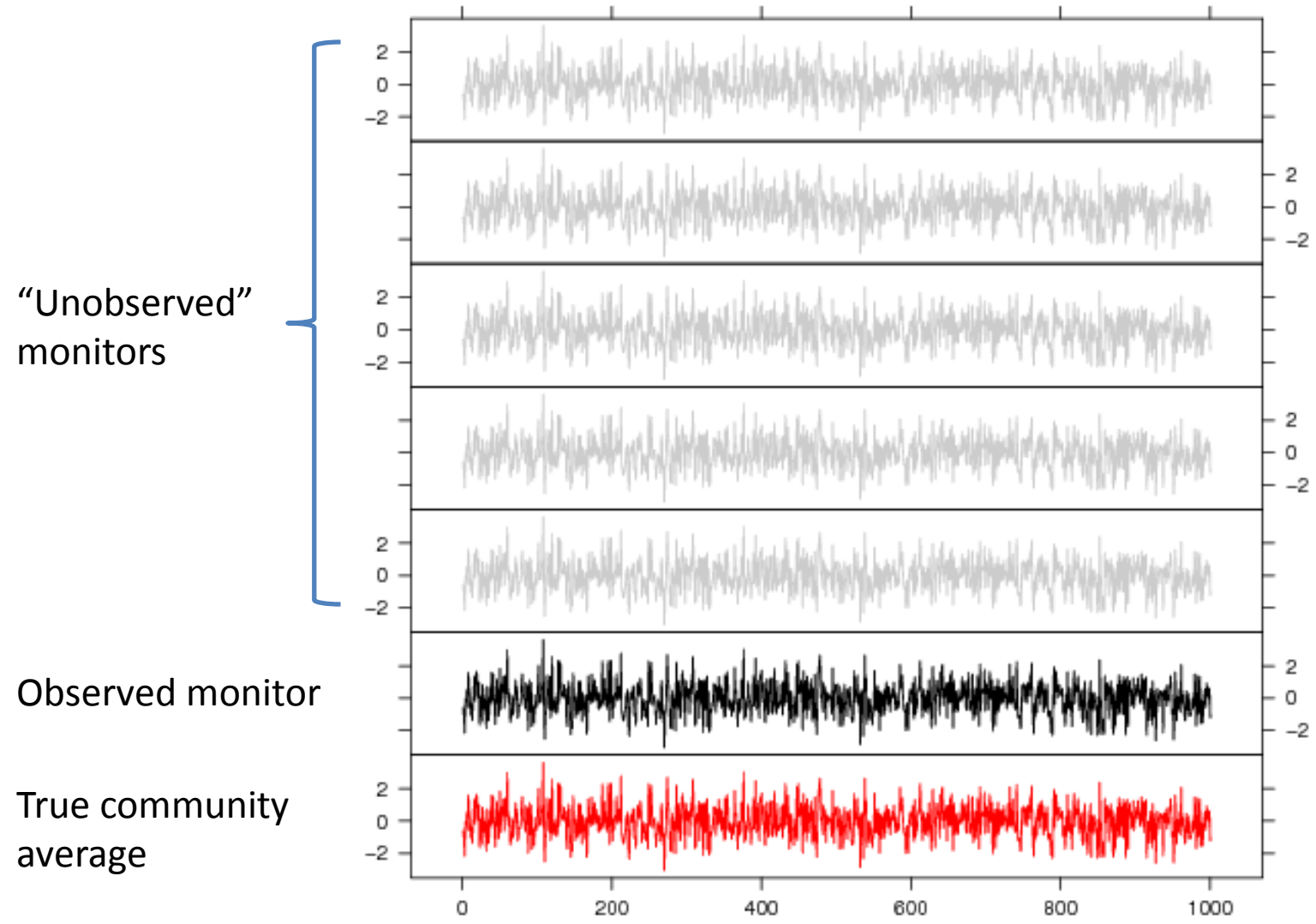
Point vs. Area: Spatially Smooth Process

Monitors

True community average



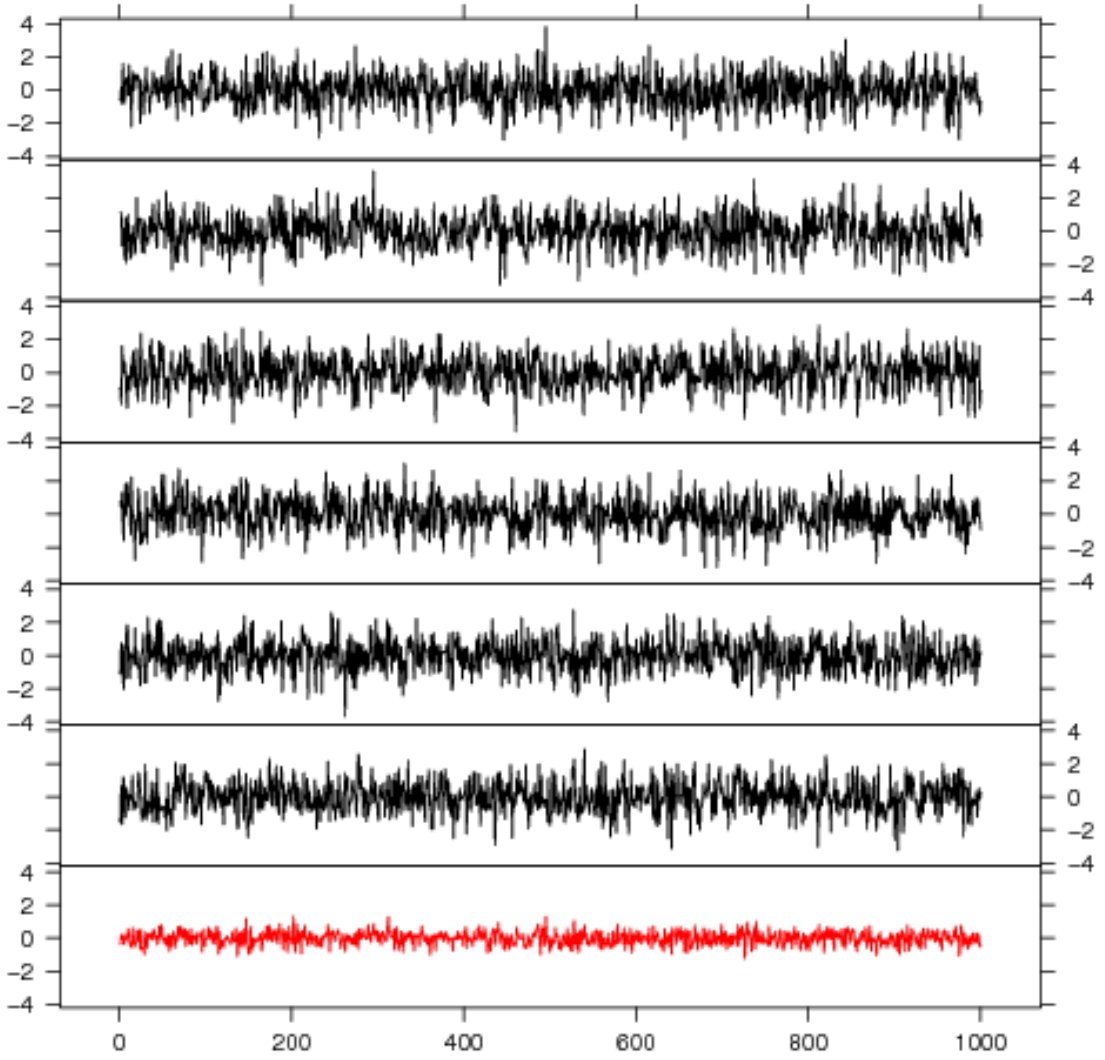
Point vs. Area: Spatially Smooth Process



Point vs. Area: Spatially Rough Process

Monitors

True community average

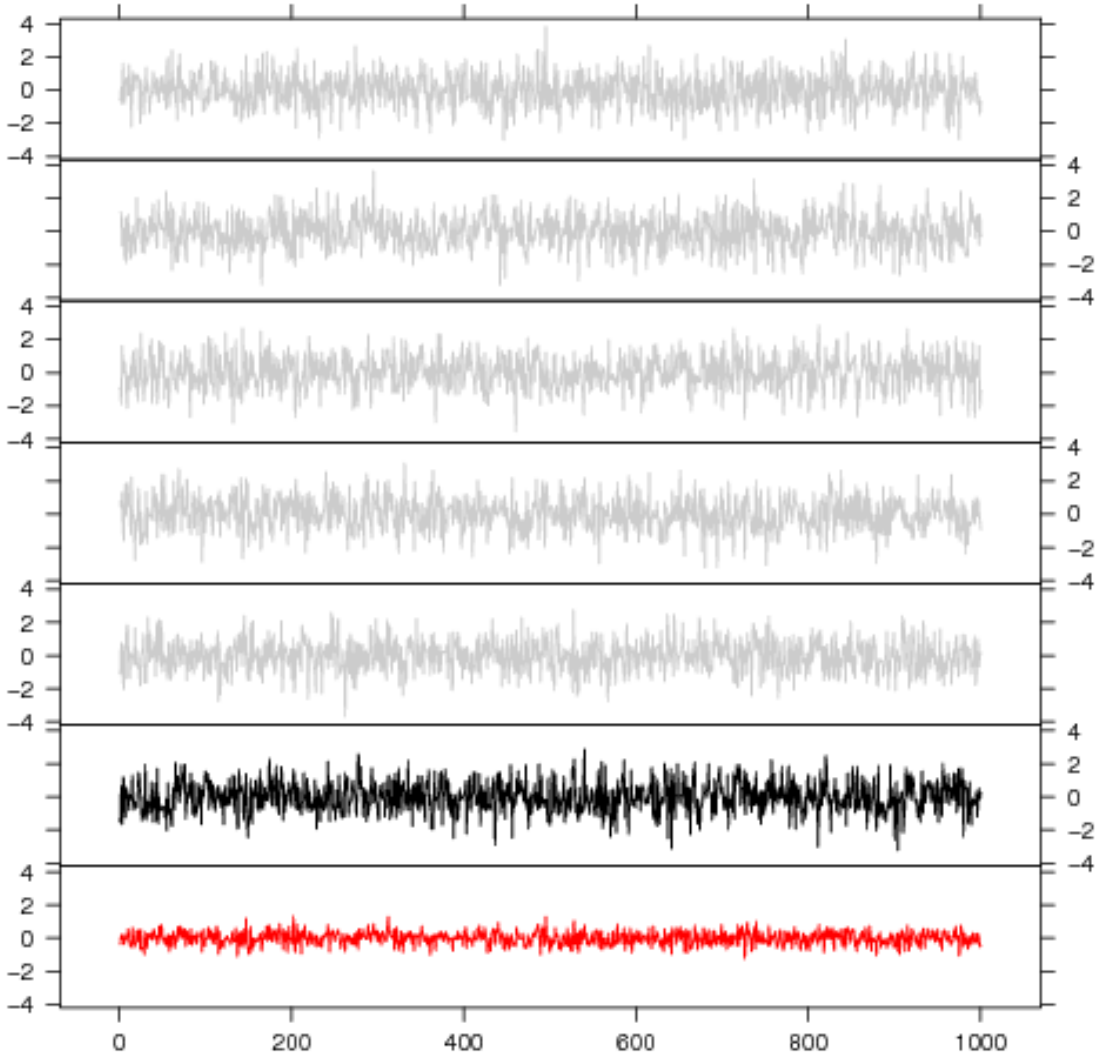


Point vs. Area: Spatially Rough Process

"Unobserved" monitors

Observed monitor

True community average



Statistical Challenges

- In theory, statistical challenges have been "solved" – we do not need more theory
- Spatial/temporal misalignment can be addressed via spatial/temporal modeling of process variability
- Missing data can be addressed via imputation models
- Effects of measurement error can be examined via measurement error modeling

Statistical Challenges

- When working with national/large databases, there is a sizable mismatch between theory and practice
- Existing approaches do not scale well to large spatial or temporal domains
- Computation quickly breaks down/becomes infeasible due to high dimensionality
- Missing data/measurement error can make standard models infeasible requiring more complex modeling/computation

Statistical Challenges

- Need statistical approaches/models that scale easily to large databases
- Simple, approximate, ad hoc, computationally efficient approaches that allow for
  - Exploring sources of uncertainty
  - Examination of “information flow” between databases
- Practical software to allow for intelligent application of these approaches to available databases