Georeferencing in the Social Sciences
- Promise and Peril

Micah Altman
Harvard University

Archival Director, Henry A. Murray Research Archive
Associate Director, Harvard-MIT Data Center
Senior Research Scientist, Institute for Quantitative Social Sciences

E: micah_altman@harvard.edu
W: http://maltman.hmdc.harvard.edu/
The Structural Challenges for Progress in Social Sciences

- Pervasive Measurement Error
- Scattered Data
- Controlled Experiments not Available in Many Fields
- Weak Theory
Georeferencing Can Make Measurements far More Accurate

- E.g. travel, time spent exercising, commutes, time at work, agriculture, distance to voting booth

Correlation between reported and real distance to tax office.
Source: [McKenzie and Sakho, 2007 as quoted in Gibsen and McKenzie, 2007]

LA Voting Precincts Relocated.
Source: [Hui and Brady, 2006]
Georeferencing Can Unify Data

Establishing comparability of most social science measurements is a major undertaking.

Yet... most social science phenomenon are unambiguously located in time and space.

Complete georeferencing would link almost all datasets at a basic conceptual level.

However, most social science data is not yet georeferenced... this is an engineering challenge.

Once done, coincident concepts can be revealed...

Source: [Weeks, et al. 2007]
Can Georeferencing fix Experiments Theory?

Not in general … although visualizations may help

Source: [J. Snow, 1854]

Source: [Calabrese, et al 2007; Real Time Rome Project 2007]
Mountains of Unified, Accurate Data
... What’s not to like?

“The increasing use of linked social-spatial data has created significant uncertainties about the ability to protect the confidentiality promised to research participants... At this time, however, no known technical strategy ... adequately resolves conflicts among the objectives of data linkage, open access, data quality, and confidentiality protection across datasets and data uses” -- [Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data, National Research Council, 2007]
Can Privacy Problems be Fixed?

Maybe not, some challenging findings…

- Large, sparse datasets can “leak” private information when correlated with external data. Even when significantly sub-sampled, perturbed, etc. [Narayan and Shmatikov 2008]
- Repeated release of perturbation-masked geospatial point data leaks increasing amounts of information. Does not help to combine with aggregation masking [Zimmerman and Pavlik 2008]
- Possible to identify other relationships in networks if you can generate seemingly innocuous relationships in same network [Backstrom, et. al 2007]
- Pseudonymous communication can be linked through textual analysis [Tomkins et. al 2004]
- K-anonymized data still vulnerable if homogenous, or attacker has enough background knowledge. L-diversity offered as replacement [Machanavajjhala, et al 2007]

Additional anonymization challenges for geospatial data

- Very fine grained location – versus multi-state aggregation mask required by HIPAA, and large social science surveys
- Background knowledge very likely
  - Easy to integrate with other datasets
  - Some data points may be directly observable
- Sequences of locations even more challenging
  - May cross aggregation units
  - Repetitive, temporally correlated
  - Induce unique networks
Managing Privacy Issues With Digital Libraries

Embedding all sensitive data access in a digital library can greatly improve subject privacy:

- Authentication, vetting, and access control
- Standardized license terms governing analysis (derived from metadata and data characteristics)
- Models can be run on-line without access to raw data
- Monitoring and auditing of data use
- Limit sequence of analyses by a user, in some cases (for promising results, see [Dwork, et al 2006])
Federated and Virtually Hosted Digital Libraries

Federated Data Networks
- Community based
- Federated collection
- Diversified portfolio of institutions
- Distributed management

Archival Virtual Hosting
- Nothing to install, yet rich services
- Depositor retains total control over content, access, presentation
- Archival network supports citation, cataloging, preservation

http://dvn.iq.harvard.edu/
Summary

Georeferencing would (partially) solve big problems for social sciences: measurement error, data integration.

Privacy is likely the fundamental challenge for social scientists using this data.

Privacy problem may never be fully solved mathematically.

Digital libraries can provide leverage for management of data privacy issues with social, legal and technical means.
References

  [also see the Real Time Rome Project [http://senseable.mit.edu/realtimerome/]
- J. Snow, 1855, On the mode of communication of cholera. London
Contact Information

http://maltman.hmdc.harvard.edu/
<Micah_Altman@harvard.edu>