# Total Survey Error and Geographic Information Systems

Ned English[1], Kevin Brown[1], and Chang Zhao[1]
[1] NORC at the University of Chicago, 55 E. Monroe St. Ste 3100, Chicago, IL, 60603

**Abstract**
Geographic Information Systems (GIS) comprises a powerful avenue for researchers to take advantage of spatial data. We have seen widespread expansion of the power and scope of GIS over the past decades due to advances in computing power, cloud-based storage, and the proliferation of mobile devices for data collection. In the social sciences generally and survey research world specifically, GIS may be employed prior to, during, and after data collection in multiple ways. For example, a researcher may geocode address information in advance of a study, plot survey respondents during production, and conduct data linkage and spatial statistics post-hoc.

As GIS is preoccupied with representing geographic information and enabling subsequent analysis, the total survey error (TSE) framework applies at multiple stages. For example, geographic data models themselves contain generalization at all scales, prior to any analysis. We explore aspects of TSE that are specific to GIS and carry implications for researchers in the social sciences.

**Key Words:** GIS, total survey error, geospatial

## 1. Introduction

Statisticians have worked to identify error sources in survey estimates since the 1930s (National Academies, 2017). Groves et al. (2004, 2010) provide a widely-cited framework known as total survey error (TSE), which summarizes errors that arise at different stages of the survey inference process. While such errors arise from numerous sources, one question is how spatial data and the tools we use to process them, or geographic information systems (GIS), contribute to the TSE framework. Specifically, one may ask where spatial data necessarily introduce errors of their own, and how GIS may be used to reduce their impacts.

As such we contend that GIS and geospatial data may both increase and decrease total survey errors in specific aspects of the TSE framework. We do so by using the exemplar CHART or “Chicago Health and Aging in Real Time” project as an illustration. CHART is a social-science research study with an environmental data-linkage component that used geospatial models to associate respondent location with measures of air pollution. In addition, we present fundamental aspects of geographic data that contribute to TSE in ways that aren’t often acknowledged, as well as how GIS can be used to ameliorate error propagation in survey data.

## 2. Background

According to Groves et al. (2004, 2010), TSE contains specific components linked to steps in both *measurement* and *representation*. In the TSE model, the *measurement* process may be distilled to measuring a survey response of a given construct and processing the resulting data (see figure 3 of Groves et al. 2010). *Representation* is concerned with creating a sampling frame for a defined target population, drawing an appropriate sample, and interviewing a sufficient share of such respondents. So, representation is often judged by sample coverage, non-response, and related survey statistics (Groves et al. 2010). We would expect issues endemic to geographic or spatial data to contribute additional measurement error to the survey process for the reasons to be described in this paper. At the same time, we would anticipate coverage and sampling error to be ameliorated by mapping and locating. In addition, GIS may facilitate measurement and estimation through the appending of ancillary data. GIS and geographic data may thus both contribute to and improve TSE in surveys.

### 2.1 Errors Specific to Geographic Information

We can define Geographic Information Systems (GIS) broadly as “tools and techniques for the management of spatial data” (Burrough and McDonnell 1998). Spatial or geographic data may be defined as geographically referenced information, which carry specific complications. For example, the earth and objects on it have infinite or “fractal” detail, and so it is not possible to ever map every feature (Monmonier 1991, Wood 1992). Even with computing technology we are limited by tools of representation, which may be as simple as the width of a line having geographic meaning when on a map at scale. Maps are thus characterized by generalization, as we cannot include all detail, and selection, where we cannot include every feature (Monmonier 1991, Wood 1992).

Beyond fundamental issues of geographic information, there are also errors characteristic of geographic data representation e.g., storing such information in a computer database. One limitation is that the real world is not composed of points, lines, or polygons, the fundamentals of vector GIS data models, or pixels, the same for raster data models. Spatial location and reference, geographic coordinates such as longitude and latitude, are themselves subject to error and distortion. The geocoding process, common in the social sciences, is known to carry errors of non-uniform impact (Eckman and English 2012). Scale itself carries ambiguity, with a line being 100 meters wide on a 1:100,000 map. Moreover, the user or viewer of maps or mappable data may have a very different impression of the same information depending on the scale results are presented at, an issue known as the “modifiable areal unit problem” (Openshow 1979). It is clear that geographic data and the tools we use to represent and present them have fundamental limitations that are not often considered by researchers.

### 2.2 The Array of Things and Chicago Health and Aging in Chicago Neighborhoods

The Internet of Things (IoT) and recent advances in low-cost sensors have made it possible to monitor environmental conditions at very high resolution, such as those collected by the Array of Things project (Benedict, Wayland, & Hagler 2017). The U.S. federally funded

Array of Things (AoT) project has installed more than 140 such sensors across the City of Chicago (Catlett et al., 2017). One challenge is how one can link such high-resolution environmental information to survey data collected from households or individuals. We attempted to do so using survey data from the Chicago Health and Activity in Real-Time (CHART) project, which has been collecting household, Ecological Momentary Assessment (EMA), and GPS tracking data from 450 elderly Chicagoans in order to assess the impact of daily activity spaces and social networks on health outcomes[1]. CHART has been collecting data in three waves using both an in-person survey and five EMA surveys per day over the course of a week in each wave. The project has also used the GPS feature of provided smartphones to track respondents and provide a measure of their "activity spaces" (York Cornwell & Cagney 2017).

The goal of our study was to use the network of AoT sensors deployed across Chicago neighborhoods to more accurately measure individual exposure to pollutants. Our approach was to link AoT sensor data with survey data collected by the CHART study as described above and detailed in English et al. (2020). In this paper we use our experience to illustrate the contribution of GIS and geographic data to the TSE framework.

## 3. Data and Methods

We downloaded sensor-level environmental data from https://arrayofthings.github.io/ representing the period July 2018 to July 2019, excluding those data points where temperature, humidity, or air-pollutant readings were outside a reasonable range. For example, some excluded data points reported air pollutant concentrations above the maximum concentration levels found in the literature. We then derived an annual average measure of temperature, humidity, PM2.5, PM10, O3, CO, SO2, H2S, and NO2 by calculating an unweighted average of monthly measures for each sensor that collected observations (English et al. 2020). Monthly values were then aggregated to create sensor-specific annual averages. We then used inverse distance weighting (IDW) in GIS to interpolate annual average values from sensors to a 200m-by-200m raster grid cell across the study area (Burrough & McDonnell 1998). Finally, we calculated mean raster values within a 250m radius of each respondent's home address which were used as their assigned environmental exposure value.

As noted above, the CHART sample was randomly selected from a frame of addresses in 10 neighborhoods in the City of Chicago. A team of field interviewers then screened each address for having a resident aged 65+, with recruitment ending after 450 completed interviews containing a series of self-reported health questions. For example, In the CHART household survey, respondents were asked whether or not they have ever been diagnosed with any respiratory disease, including emphysema, asthma, chronic bronchitis, or chronic obstructive pulmonary disease. Out of 343 elderly respondents, 11.7% said they had been diagnosed with respiratory disease. We then fit four logistic regression models to understand the relationship between environmental exposure and health outcomes. A much more detailed analysis may be seen in English et al. (2020), while this paper is designed to use the results to describe the relationship to TSE.

[1] http://www.norc.org/Research/Projects/Pages/chicago-health-and-activity-in-real-time.aspx

## 4. Results and Discussion

As described in English et al. (2020), our analysis revealed considerable variation in temperature, humidity, and air-pollution across space based on data derived from the array of things (AOT) sensors. Table 1 below shows one regression analysis that was used to explore the relationship between air pollution exposure and self-reported lung conditions. One key finding in the below model was that environmental variables were not significant alone, but were significant when interacted with residence time spent in the neighborhood. While Table 1 shows one of several models discussed in English et al. (2020), we use it here to illustrate that there is some relationship between air pollution exposure and self-reported lung health in our study as captured by linking AOT to CHART. One clear limitation of such an exploration is that we do not know their prior environmental exposure, whether or not they have resided in the same location. Moreover, our analysis demonstrated the challenges of data availability and how best to link environmental data to households. In addition, specific environmental expertise is necessary to understand the salience of specific variables, alone or in concert, and how to treat outliers.

**Table 1:** Logistic regression: environmental exposure and respiratory health[2]

| *Independent Variable* | *Odds Ratio* | *P* Value |
|---|---|---|
| (Intercept) | 0.205 | 0.028* |
| $O_3$ | 0.523 | 0.134 |
| $SO_2$ | 1.524 | 0.560 |
| Temperature | 1.143 | 0.429 |
| Humidity | 1.038 | 0.561 |
| Smoke Regularly | 2.484 | .005** |
| $NO_2$ : Neighborhood 25-50 years | 16.824 | .040* |
| *Hosmer-Lemeshow Test* | *Statistics = 2.398* | *P < .969* |

** P < 0.05, ** P < 0.01, *** P < 0.001*

As the purpose of this discussion relates to TSE and error sources, we may summarize specific error sources in Table 2 below. Clearly, the nature of geographic data and geographic data processing introduced some ambiguity into our analysis, both implicitly and explicitly. Most of the errors in Table 2 concern measurement error on the TSE framework. That said, our approach was highly novel with measurements that would not have been possible previously using different technology. Consequently, we argue that having potentially limited data may be superior to not having any or those of low spatial resolution.

[2] Controlling for gender, race, age, education level, time resident in the neighborhood, and smoking status while considering interaction effects

**Table 2:** Error Sources in the CHART Study

| *Error Source* | *Discussion* |
|---|---|
| GPS Position | Waypoints potentially in incorrect place, linked to more distant environmental data |
| Sensor Data Capture | Pollution, temperature, or humidity readings may be incorrect in unpredictable ways |
| Interpolation of Surfaces | Our inverse-distance weighting (IDW) model will differ from ground-truth |
| Linkage of Surface to Individuals | The method of assigning individual location to inverse-distance weighted estimates may mis-state their exposure |
| Survey Measurement Error | The CHART survey itself could have gathered erroneous self-reported health information |
| Interpretation of Environmental Data Impact | The salience of exposure to specific pollutants is complex and nuanced |

## 5. Conclusion

GIS and spatial data may make a substantial contribution to the TSE framework, both with respect to contributing to specific error sources as well as understanding and limiting errors. For example, coverage error is a specific error source in the representation process, which is shown in the literature to be impacted by geocoding error during frame construction (Eckman and English 2012). Modern geocoding, however, is considerably more accurate than earlier methods such as paper-and-pencil household listing, and thus introduces less coverage error at aggregate. Sampling error, another component of the representation process, may be ameliorated by more effective stratification during the design process which is enabled by small-area data linkage through GIS. Measurement error, a component of the measurement process, may be reduced through data linkage and spatial modeling, both of which are made possible through GIS. GIS and spatial data also create the opportunity for new measures and derive inferences, with less measurement-error than could have been done before

Maps and GIS also present the opportunity to understand and remediate error through visualization and novel approaches to quality-control. Maps allow us to visualize components of TSE, such as where non-response bias or coverage error may be most concentrated in specific sampled units. GIS allows us to be more specific about location, and even may permit us to visualize uncertainty in our estimates. The detailed level of understanding that maps provide represent a new level in understanding survey error sources.

In conclusion, GIS and geospatial data contribute to the TSE framework with specific error sources and the potential for error remediation. Fundamentally, GIS allows for new approaches to measurement and quality control that were previously possible, and should be considered a valuable and necessary component of the survey and analysis process.

## References

Benedict, K., Wayland, R., & Hagler, G. (2017). Characterizing air quality in a rapidly changing world. EM Magazine, November. Retrieved from the EPA website at https://www.epa.gov/sites/production/files/2017-11/documents/wayland_with_citation.pdf.

Burrough, P.A. and R. A. McDonnell. 1998. Principles of Geographic Information Systems. Oxford: Oxford University Press.

Catlett, C. E., Beckman, P. H., Sankaran, R., & Galvin, K. K. (2017). Array of Things: A scientific research instrument in the public way: Platform design and early lessons learned. Proceedings of the 2nd International Workshop on Science of Smart City Operations and Platforms Engineering, 26-33. ACM.

Eckman, S and English, N. 2012. Creating Housing Unit Frames from Address Databases: Geocoding Precision and Net Coverage Rates Field Methods. 24(4): 399-408 http://fmx.sagepub.com/content/early/2012/07/02/1525822X12445141.abstract?papetoc.

English, Ned, Chang Zhao, Kevin Brown, Charlie Catlett, Kathleen Cagney. Making Sense of Sensor Data: How Local Environmental Conditions Add Value to Social Science Research. 2020. Social Science Computer Review. https://doi.org/10.1177/0894439320920601

Groves, Robert and Lars Lyberg. Total Survey Error Past, Present, and Future. 2010. *Public Opinion Quarterly* 74(10), 849-879.

Groves, Robert, Floyd Fowler, Mick Couper, Eleanor Singer, and Roger Tourangeau. 2004. Survey Methodology. New York: Wiley.

Monmomier, Mark. 1991. How to Lie with Maps. Chicago: University of Chicago Press.

National Academies of Sciences, Engineering, and Medicine 2017. Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps. Washington, DC: The National Academies Press. https://doi.org/10.17226/24893.

Openshow S. A million or so correlation coefficients, three experiments on the modifiable areal unit problem. Statistical applications in the spatial science. 1979:127-44.

Wood, Denis. 1992. The Power of Maps. New York: The Guilford Press.

York Cornwell, E. & Cagney, K. A. (2017). Aging in activity space: Results from smartphone-based GPS-tracking of urban seniors. Journals of Gerontology: Social Sciences 72, 864-875.