SURVEYS · TECHNOLOGY AND THE FUTURE

Linking Survey Microdata to Geospatial Layers for Spatially Explicit Knowledge

Attaching geographic information to survey responses lets researchers study how places shape people's lives. This article explains how survey microdata are linked to geospatial layers, the main analytical pitfalls and how to protect respondents' confidentiality.

ModQu research article · 744 words · 6 references (APA 7th edition)

Every survey respondent lives somewhere. Their neighbourhood's air quality, access to shops and services, crime levels, green space and local economy may all affect their health, opinions and behaviour. **Geospatial layers** are data sets that describe such features of places: maps of pollution, locations of hospitals, census figures for small areas or satellite measures of land use. Linking survey **microdata**, the individual responses, to these layers makes survey knowledge **spatially explicit**, so that researchers can ask how where people live relates to how they live.

What spatial linkage adds

Linking to geospatial layers adds information that respondents could not easily report. Few people know the pollution level on their street, the exact distance to the nearest hospital or the unemployment rate in their district. Layers supply these measures objectively and consistently for every respondent, without lengthening the questionnaire.

Linkage also allows the same survey to be analysed at different geographic scales and combined with other data that share the same geography. A national survey linked to local service locations, for example, can show how access to services varies between regions and how it relates to people's use of those services and their satisfaction with them. For businesses, linking customer surveys to store locations and local population data can show how distance and local competition shape customers' views.

How the linkage works

The process has three steps, shown in Figure 1.

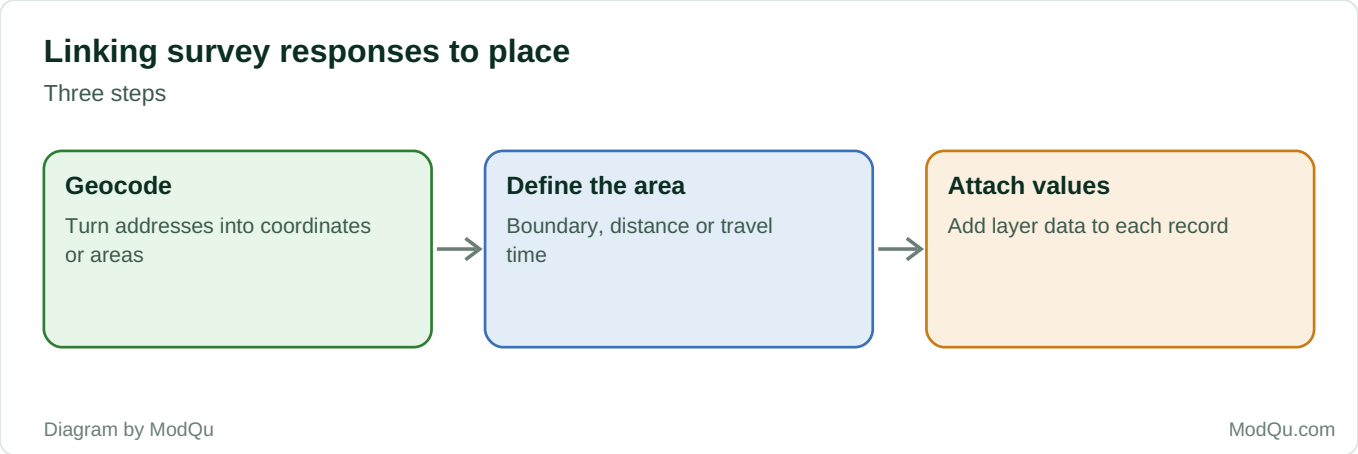


Figure 1. Steps in linking survey responses to geospatial layers. Diagram by ModQu.

Geocoding. Each respondent's address is converted into map coordinates or assigned to an area, such as a census tract. Accuracy varies with the method: matching to individual address points is generally more precise than estimating a position along a street (Zandbergen, 2008). Errors in geocoding can place respondents in the wrong area.

Defining the area. The researcher decides which area around each respondent is relevant: an administrative unit, a circle of a set distance or the travel time to a service.

Attaching values. Values from each layer, such as average pollution, distance to the nearest clinic or the area's unemployment rate, are attached to each respondent's record.

Analytical pitfalls

Results can depend on how areas are defined. The **modifiable areal unit problem** describes how the same data can produce different results when grouped into different areas or at different scales (Openshaw, 1984). A relationship between neighbourhood deprivation and health may look strong with one set of boundaries and weak with another.

A related issue is the **uncertain geographic context problem**. People do not spend all their time in their home neighbourhood; they work, shop and socialise elsewhere. Measuring exposure only around the home may misrepresent the places that actually affect them (Kwan, 2012).

Area effects are also hard to separate from the characteristics of the people who live in an area. People choose, or are constrained in, where they live, so differences between neighbourhoods may reflect who lives there rather than the effect of the place itself (Diez Roux, 2001). Surveys help here, because they measure individual characteristics that can be taken into account in the analysis.

Protecting confidentiality

Precise locations make respondents identifiable. A point on a map, combined with a few survey answers, may point to a single household. VanWey et al. (2005) described the tension between the scientific value of spatially explicit data and the duty to protect respondents. Common safeguards include:

- releasing only area-level identifiers or attached layer values, not coordinates;
- **geographic masking**, which moves each location by a small random distance so that points cannot be traced to households while broad patterns remain (Armstrong et al., 1999);
- providing precise locations only in secure research environments.

Respondents should be told if their location will be linked to other data.

An example

A city health survey asks residents about asthma symptoms. Researchers geocode each address, calculate average air pollution within 500 metres and attach it to each record. Respondents in the most polluted fifth of locations report symptoms more often than those in the least polluted fifth, after accounting for smoking and income. The researchers repeat the analysis with a 1,000-metre radius and find a similar pattern, which gives them more confidence that the result does not depend on the chosen area. The released data set contains only the pollution values, not coordinates (illustrative example).

Conclusion

Linking survey microdata to geospatial layers adds the dimension of place to survey knowledge. Careful geocoding, thoughtful choice of areas and strong confidentiality protection make the linked data both useful and safe.

References

Armstrong, M. P., Rushton, G., & Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine, 18*(5), 497–525.

Diez Roux, A. V. (2001). Investigating neighborhood and area effects on health. *American Journal of Public Health, 91*(11), 1783–1789. https://doi.org/10.2105/AJPH.91.11.1783

Kwan, M.-P. (2012). The uncertain geographic context problem. *Annals of the Association of American Geographers, 102*(5), 958–968. https://doi.org/10.1080/00045608.2012.687349

Openshaw, S. (1984). *The modifiable areal unit problem*. Geo Books.

VanWey, L. K., Rindfuss, R. R., Gutmann, M. P., Entwisle, B., & Balk, D. L. (2005). Confidentiality and spatially explicit data: Concerns and challenges. *Proceedings of the National Academy of Sciences, 102*(43), 15337–15342. https://doi.org/10.1073/pnas.0507804102

Zandbergen, P. A. (2008). A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems, 32*(3), 214–232. https://doi.org/10.1016/j.compenvurbsys.2007.11.006