

## Visual error detection in geocoded point data

Jukka Matthias Krisp, Terhi Ahola, Olga Spatenkova

Geoinformation & Positioning Technology, Helsinki University of Technology, Helsinki, Finland,  
[jukka.krisp@hut.fi](mailto:jukka.krisp@hut.fi), [Terhi.Ahola@tkk.fi](mailto:Terhi.Ahola@tkk.fi), [Olga.Spatenkova@tkk.fi](mailto:Olga.Spatenkova@tkk.fi)

### Extended Abstract

Point patterns, where the only data is the location of a set of point objects, represent the simplest possible spatial data. As pointed out by O’Sullivan and Unwin (2003) there are a number of requirements for a set of events to constitute a point pattern. Among these requirements, the event locations must be “proper”. They should not be, for example the centroids of units chosen as representation. They really should represent the point location of entities that can be sensibly being considered points at the scale of the study. In common GIS analysis operations location points can be obtained by geocoding (e.g. based on addresses). Geocoding is the “process of creating geometric representations for descriptions of locations” (ESRI, 2002). Therefore geocoding provides a spatial reference to data, which is not initially recorded. In some cases it is referred to as “address matching.” Geocoding allows the address point to be linked to other geospatial databases. The resulting point distributions can be analyzed using any of a variety of spatial, geostatistical or exploratory spatial data analysis techniques.

In some cases the geocoding is an automated process and errors that may occur during this process are not directly obvious. Within a GIS environment “playing” with the input data, using available geovisualization tools may reveal a sudden pattern that makes the user stop and think, “This cannot be!” The visual output (a map, curve, etc.) may lead the user to assume an error in the data at hand and lead to further investigations. Meng (2003) suggests that maps work as thinking instrument that visually supports users to confirm known facts, detect unknowns and finally value-add the database either by inserting the new knowledge into it or removing the redundancy from it. Numerous authors have investigated the use of Geovisualization and its usefulness in the detection of relationships between spatial clusters and variables (Fuhrmann, Ahonen-Rainio et al., 2005; Gahegan, 1999; Griffin, 2004). We suggest that visual methods may be useful in the detection of errors within geocoded point datasets. A large number of publications have dealt with the positional error in automated geocoding, among them Cayo and Talbot (2003) who investigating the distances, which are measured from the true point locations of a house to the automated geocoded points (generally based on the street address) and to the residential parcel points to determine the positional error.

Generally the visual inspections can detect the presence of a mistake in the date, the absence of data, or positional accuracy of data. Visual quality analysis can be performed using hardcopy plots or on-screen views. The hardcopy plotting of data might be in some cases the best method for checking for missing features, misplaced features, and

registration errors. On-screen views are an excellent way to verify that edits to the database were made correctly, but are not a substitute for inspecting plots. Visual inspection should occur during initial data capture, at feature attribution, and again at final data delivery. At initial data capture, review the data for missing or misplaced features and alignment problems that could point to a systematic error. Each error type needs to be evaluated along with the process that created the data in order to determine the appropriate root cause and solution (McCain and Masters, 1998). Depending on the data context, there are classic misplacements of points, like a house in the water, which can be detected by visual analysis, but also by automatic consistency checks.

Our input data are fire & rescue service missions, which are geocoded by the administrative body within these organizations. The geocoding is based on the address data, which is recorded during the rescue mission. The initial dataset from the fire and rescue services included no proper metadata, only a small description of the dataset and a list describing the coding for the accidents types. No further information on the geocoding process or of the data quality had been provided. Referring to Gould in O'Sullivan and Unwin (2003) geographical data is not random and (because of autocorrelation) geographical data is not independent random. When dealing with geocoded fire & rescue service mission data, we assume that the data is potentially randomly distributed, meaning that theoretically the incidents could happen anywhere in the study area. In reality this is not the case, as incidents are related to a number of reasons. Incidents are to some degree related to population density (Krisp and Karasova, 2005), traffic density etc., and also time of the day / year (Ahola, Virrantaus et al., 2007). Therefore we can assume clusters in this data and we can also assume these are irregular clusters.

The first step in the analysis of the point data is to take a look at what we have at hand. The data contains 11.435 points. Generally there seem to be some clusters in the data, with a high concentration in the Helsinki centre area. Due to the fragmented city structure of Helsinki we can recognise few accidents in the water areas. That includes incidents in water areas, as quite frequently the rescue services have save people from boats or some did break through the ice. Due to the number of points it is impossible to "see", if there are geocoding errors in the data. A grid analysis function reveals a regular pattern in the data. Within this analysis a 250 grid is laid over the point file and the points within each grid cell are counted. These numbers are classified and the colour is given on the class base. We can notice a very unusual pattern in this data, showing high peaks of incident occurrence between a very regular number of cells. This indicates an error within our dataset and needs further investigations. So far this approach has been tried only on data about fire rescue incidents in Helsinki. Applying this to other datasets, which are based on address geocoding might show similar patterns and assist the identification of geocoding errors in a potentially randomly distributed points datasets.

---

**References**

- Ahola, T., Virrantaus, K., Krisp, J.M. and Hunter, G., 2007. A spatio-temporal population model to support risk assessment and damage analysis for decision-making. *International Journal of Geographical Information Science (IJGIS)*, accepted for special issue.
- Aronoff, S., 1993. *Geographic Information Systems: a management perspective*. WDL Publications, Ottawa.
- Cayo, M.R. and Talbot, T.O., 2003. Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics*, 2(10).
- ESRI, 2002. *ArcGIS 8.1 Online Manual*. Environmental Systems Research Institute Inc.,
- Fuhrmann, S., Ahonen-Rainio, P., Edsall, R., Fabricant, S.I., Koua, E.L., Tolon, C., Ware, C. and Wilson, S., 2005. Making Useful and Useable Geovisualization: Design and Evaluation Issues. In: J. Dykes, A.M. MacEachren and M.J. Kraak (Editors), *Geovisualization*. Elsevier, Amsterdam, pp. 553-566.
- Gahegan, M., 1999. Four barriers to the development of effective exploratory visualisation tools for the geosciences. *Geographical Information Science*, 13(4): 289-309.
- Griffin, A.L., 2004. *Understanding How Scientists Use Data-Display Devices for Interactive Visual Computing with Geographical Models*. PhD Thesis, Pennsylvania State University, Pennsylvania.
- Krisp, J.M. and Karasova, V., 2005. The relation between population density and fire & rescue service incidents in urban areas, *Proceedings on the 10th Scandinavian Research Conference on Geographical Information Science (ScanGIS)*, Stockholm, Sweden, 13. -15. June, pp. 237-246
- McCain, M. and Masters, W.C., 1998. Integrating Quality Assurance into the GIS Project Life Cycle. *ESRI ArcUser Magazine*, July-Sep.
- Meng, L., 2003. Missing Theories and Methods in Digital Cartography, *The 21st International Cartographic Conference (ICC)*, Durban, South Africa, pp. pages pending
- O'Sullivan, D. and Unwin, D.J., 2003. *Geographic Information Analysis*. Wiley, New Jersey.
- Zhang, J. and Goodchild, M.F., 2002. *Uncertainty in Geographical Information*. Research Monographs in Geographic Information Systems. Taylor & Francis, New York.