POSITION PAPER

Michael F. Goodchild
University of California, Santa Barbara

Potential of dense-tracking data

There seems to be no doubt that densely sampled tracks have many powerful applications in transportation and social science. We know already that even with standard GPS (no differential correction) the average of a small number of tracks provides the cheapest and fastest method of maintaining currency in street centerline databases (compare the *Los Angeles Times* story on problems of map currency, 21 August 2005), and this potential is already being exploited by some vendors. We know that averaging tracks is capable of providing positional accuracies at the meter and even sub-meter levels, which would be sufficient to support lane-level modeling and guidance. We know that dense tracking provides the most rapid way of detecting and responding to accidents and other obstructions. We also know that tracks can be parsed to determine activities (particularly mode) with some accuracy, and that comparison of tracks can provide useful information on interactions. So what impediments make this potential problematic?

The IRB problem

The potential for tracking to compromise privacy is clearly a major issue, and institutional review boards are likely to make it very difficult to conduct research in this area without stringent safeguards. A recent article (VanWey *et al.*, 2005) argues that all current methods for protecting confidentiality in microdata, such as aggregation, positional distortion, and firewalls are flawed, and that major funding is needed to support the exploration and examination of possible alternatives. At the same time, there is a conspicuous difference between the stringency likely to be exhibited by IRBs, and the environment in which the private sector and many government agencies operate.

Analytic techniques

Dense tracks represent a novel type of data that is not compatible with any of the current analytic environments—GIS, statistical packages, OLAP, etc. Simple null hypotheses, such as CSR for cross-sectional point data, have no equivalent for tracks. We need to build a toolkit of simple analysis techniques, to bring the analysis of tracking data to somewhere near the level of support available for cross-sectional data. We need a library of metrics of similarity between tracks, to support a basic equivalent of cluster analysis, and to identify anomalies and outliers—what do we mean by saying that two tracks are "similar"? We also need better methods of visualization, as anyone who has tried to make sense of the Hagerstrand plot of large numbers of tracks will attest.

Traditional analytic techniques have tended to focus on the "double negative" approach of null hypothesis rejection. Thus although it is normally uninteresting, for example, to establish that a point pattern is distinctly different from the random pattern of CSR, we tend to be satisfied with analyses that result in a rejection of the null hypothesis, and to move only rarely to a more positive acceptance of a well-defined alternative. In part this

is the result of lacking tests with adequate power, since there will always be an infinity of acceptable alternatives, but only one rejected null.

With today's GIS and massive computing power there is the potential to adopt a more positive approach, as is being demonstrated by various types of dynamic models. But we need a toolkit of techniques appropriate for tracking data—stochastic models of tracks, for example.

Although the idea of an average track is compelling, in reality there is no obvious and widely accepted method for averaging tracks, just as there is no widely accepted method for averaging lines in two dimensions (Goodchild, Cova, and Ehlschlaeger, 1995). We need a standard model of uncertainty in tracks, comparable to the Gaussian distribution for scalar measurements. Other simple geometric tools are also needed, such as the polyline-to-arcs tool developed by Noronha and Church for preparing GIS data to support Paramics simulation, and should become a standard part of a computing environment for tracks.

In short, the widespread availability of dense-tracking data has exposed the lack of suitable tools, models, and theories in the current scientific apparatus, and has created an urgent need for fundamental research.

References

Goodchild, M.F., T.J. Cova, and C.R. Ehlschlaeger, 1995. Mean objects: extending the concept of central tendency to complex spatial objects in GIS. *Proceedings GIS/LIS '95, Nashville, TN*, pp. 354–364.

VanWey, L.K., R.R. Rindfuss, M.P. Gutmann, B. Entwisle, and D. Balk, 2005. Confidentiality and spatially explicit data: concerns and challenges.