FUTURE VIEWS OF FIELD DATA COLLECTION IN STATISTICAL SURVEYS

Sarah Nusser
Department of Statistics & Statistical Laboratory
Iowa State University
nusser@iastate.edu

Leslie Miller
Department of Computer Science
Iowa State University
lmiller@iastate.edu

Keith Clarke
National Center for Geographic Information and Analysis & Department of Geography
University of California, Santa Barbara
kclarke@geog.ucsb.edu

Michael Goodchild
National Center for Geographic Information and Analysis & Department of Geography
University of California, Santa Barbara
good@ncgia.ucsb.edu

Project Battuta Website
www.statlab.iastate.edu/dg

Abstract

Recent research in mobile computing for survey data collection has focused on implementing paper-based procedures for field activities on handheld and tablet devices using a structured client-server interaction model developed for computer-assisted telephone interviewing labs. As these applications have evolved, little attention has been given to new and emerging information technologies. The goal of Project Battuta is to conduct research that focuses on the developing next-generation field and infrastructure technologies for mobile data collection, with an emphasis on using digital geospatial information resources within the field data collection process. To do so requires a new paradigm for field data collection that draws upon the ability to integrate distributed geospatial information resources in a mobile environment, modern computing paradigms for mediated infrastructures that increase the simplicity and flexibility of the user's interactions with data resources, and future mobile technologies such as wearable computers and augmented reality in the computer-assisted field data collection environment. We discuss research in these three areas and their potential for transforming the field data collection process for statistical surveys.

1 Introduction

Interest in mobile computing for statistical survey data collection is rapidly increasing. Early research in mobile applications has focused on handheld and tablet devices using design principles developed for desktop telephone interviewing laboratories. As such, these endeavors have failed to anticipate emerging technologies such as augmented vision and wearable computers, infrastructures that facilitate more flexible and user-friendly interactions, and the exploitation of digital information resources such as geospatial data and other information beyond that currently supplied as part of a central survey database.

To create a foundation for incorporating next-generation mobile field technologies in statistical survey data collection settings, research is being conducted in using geospatial data and emerging technologies to support survey field data collection. The shift towards increasingly capable mobile technologies and highly available digital geospatial resources presents numerous opportunities and challenges for survey data collection, which has historically relied on highly structured protocols and prescribed interactions. A new paradigm for field data collection is needed that takes full and appropriate advantage of integrated and distributed geospatial information resources, modern computing paradigms for mediated infrastructures, and emerging mobile technologies. The potential impact of these developments on the survey process is considerable and must be evaluated carefully with respect to its effect on data quality and the information contained in the survey product.

2 A New Model for Computer-Assisted Field Data Collection

The client-server model used in computer-assisted field data collection systems relies on predefined interaction rules and data sources. Although this system design largely reflects existing technologies when the design was adopted, it also mirrors the structured nature of survey data collection. In general, strict protocols govern scientific data collection, including scripted interviews, specific definitions supporting questions and response options, edit rules to check the consistency of responses for a respondent, and so on. To the extent possible, protocols and specifications are administered via the survey instrument software as a means of reducing human error in data collection.

An alternative model is to augment the client-server setting with an infrastructure that enables more flexible access to a broader suite of information resources. Under this paradigm, a field user may take advantage of digital repositories prepared for data collection, as well as information resources more generally available via the Web. Geospatial data represent an especially rich set of resources of interest in field data collection, and are a special focus of this paper. This model raises issues with respect to three areas of field data collection for statistical surveys: the user, the field computing environment, and the survey process.

First, the primary user in a statistical survey is a field representative with specific data collection skills such as interviewing, pricing, or biological data collection. These skills are of primary

importance in collecting high quality data. However, it is unlikely that field staff will also be technologically literate. This limitation manifests itself in two ways.

With geospatial data in particular, the richness of the information resource is far higher than is generally available with text and numeric information. If a more complex set of resources is made available, it must be in a form that is usable for persons who may have limited training for handling such resources. This raises questions about what types of information are most useful for various aspects of the survey process, and how best to display such information within a given field computing environment. A special challenge occurs when multiple sources of information are needed, requiring intelligent integration of potentially conflicting information sources.

A second dimension from the user perspective is the increased complexity in interacting with relatively unstructured and diverse information resources. In order to provide a practical benefit for field data collectors, the infrastructure must seamlessly negotiate for the user, hiding the complexities of interacting with ad hoc information resources. A more flexible architecture than the client-server model is needed to provide this kind of functionality to users.

The second issue in pursuing this vision for field data collection systems is the mobile computing environment. In most mobile settings, the field computing environment is extremely limited in computing capacity, bandwidth, and display interfaces relative to the large volumes of information that exist in distributed geospatial data repositories. These limitations have two kinds of impacts. Historically, computer-assisted data collection applications and geospatial data processing software have relied on clients with ample computing capacity. The thin client environment is far more restrictive, and calls for rethinking basic infrastructure design as well as strategies of implementation, such as where computing loads are borne. Second, the nature of field settings can be quite varied, where field setting is broadly defined to be the field data collection application plus the computational environment. For this new model for field data collection to be effective across a broad range of settings, extensibility is a primary issue. Of special importance is that the information requested by the user is prepared and delivered in a manner accounts for the field context.

The third area impacted by this model is the survey process. Under a client-server model, interactions with a database server are prescribed via communications software and mobile client application (survey instrument), and the field user has little if any access to information other than that provided by the agency. As noted earlier, this paradigm mirrors and supports the structured nature of a scientific data collection process. However, it is not uncommon for field staff to use ad hoc information resources to plan field activities, identify new housing units, and track respondents. For example, unofficial paper-based information resources, such as maps and telephone books, are widely used by field staff for locating sample units in establishment and household surveys. More generally, there are portions of the data collection process that stand to benefit from more flexible access to information resources than is typically offered by client-server systems, provided that such access does not interfere with the principal goal of collecting data of high quality.

The availability of a broad array of geospatial information resources plus the varied and potentially limiting nature of the field computing environment raises a number of research questions relating to infrastructure design, extensibility to emerging technologies, and application within the survey domain. The following sections discuss these areas in relation to selected Project Battuta research activities.

3 Infrastructure Design

A key feature of any environment designed to give field workers access to geospatial data is the infrastructure used to connect the field devices to the data sources. Under our model, the infrastructure must be very flexible in its ability to obtain data. At the same time, it must be capable of minimizing the amount of data flowing through the network. To do this, we have chosen to make use of object-oriented views implemented as mobile agents. The views provide an excellent basis for deriving the data for the user's request and the mobile agent approach creates a great deal of flexibility in the location for integrating or analyzing data.

The implementation of our view agent infrastructure model makes use of wrappers, mediation, and XML. The wrappers are used for encapsulating not only the data sources, but the mobile field devices as well. As is generally the case, the wrappers allow the details associated with the heterogeneity of the data source (or device) to be localized. The result is that within the boundaries of the wrappers, the mobile view agents work in a relatively homogeneous environment of manipulating XML encoded data.

The internal infrastructure environment is also populated with a set of computation servers. Each computation server has a local object-oriented data warehouse equipped with a set of tools designed to work with geospatial data. Since the prospect of query reuse is likely for a field worker, we store the final and intermediate results in the data warehouse, allowing the warehouse to act as an active cache. The warehouse is a key component to providing options that allow the infrastructure to accommodate field limitations and to prepare appropriate views for the field.

Taken as a whole, these tools form the basis for a dynamic, adaptable infrastructure for handling geospatial data in varied and potentially limited mobile field applications. These features are demonstrated in the testbed environment discussed in Section 6.

4 Field Tools

The use of geospatial data in the field environment requires integration of heterogeneous resources. Regardless of whether data sources are combined in a warehouse or the mobile field environment, issues of positional accuracy arise. The positional accuracy of any geospatial data set depends on many factors, including the spatial resolution of the data, the scale or representative fraction if the data were derived from paper maps, and general issues of accuracy. We use the term conflation to refer to the combination of geospatial data sets that represent at

least in part the same phenomena, over overlapping geographic areas. For example, a DOQ image and a digital road map both portray roads, in the former case as variations in the image, and in the latter case as objects.

Because positional accuracies can never be perfect, it is inevitable that the positions shown for the shared phenomena will be different. The amount of difference may be unimportant or unacceptable, depending on the spatial resolution required for the application. Data sets can also differ in the attributes they assign to objects, and in the spectral properties assigned to locations in images.

We are investigating conflation as a general problem associated with access to multiple geospatial data sets, and as a specific problem for field workers retrieving such data sets from digital libraries. Conflation can occur in several distinct contexts: 1) as a means of correcting or reducing errors in one data set through comparison with a second; 2) as a process of averaging, in which the product is more accurate than either input; 3) as a process of concatenation, in which the output data set preserves all of the information in the inputs; and 4) as a means of removing visually unacceptable differences when data sets are overlaid.

As part of this research, prototypes have been developed for the specific problem of conflating data sets in the field that disagree in positional accuracy to an unacceptable level. One assumes that an existing data set must be conflated with measurements of position obtained in the field from the Global Positioning System (GPS), to remove any unacceptable differences. The second prototype allows two data sets obtained from different servers to be conflated, again in the client and in order to remove unacceptable differences of position. Both prototypes are being ported to the wearable system discussed below.

5 Wearable Systems for Field Computing

A handheld device coupled with a GPS receiver enables user context (e.g., position) to be combined with geospatial resources in a mobile environment. Wearable computers offer an extension of this integration of user perspective, digital information resources, and surrounding reality. For example, visual augmentation offers new possibilities for supporting data collection activities by combing user position, digital maps, and auxiliary information about a household with views of the physical environment.

As part of Project Battuta, we are researching how state-of-the-art wearable computing technology can be of use in field data collection. We are currently building a prototype of such system, using the CharmKit PC-104 system. The prototype includes input from a single-hand keyboard called a Twiddler; from a digital compass with roll, pitch and yaw tracking; and from a small GPS receiver. Output is delivered to a MicroOptical eyepiece, which clips onto a regular pair of glasses.

Further work with the prototype relates to modes of operation. A navigation mode is being explored that leads the user through geographic space, using prior paths, or computed navigation

information. A second mode focuses on positioning, where assistance to the user in geographic positioning and spatial alignment are given. In our prototype, data collection will be conducted by filling in on-line Internet-based forms, presented to the user in context by Java applets running on a Mobile Internet platform.

We are conducting research into aspects of the user interface that such a system would require, including icons and graphics for the two modes, and map-based displays within the visual interface.

6 Integration for Survey Applications

Testbeds for demographic and land-based surveys are being developed to integrate and demonstrate Batutta research activities within the survey setting. Current research focuses on prototyping the infrastructure architecture with alternative field clients to support survey tasks that focus on the use of geospatial data, such as navigating to and identifying specific sample locations. Parameters used to demonstrate the flexibility of the infrastructure include the user's physical and computing environment, network constraints, repository characteristics, and the types of tools and products required to support data collection activities in the field. The wearable computing environment will be added as an alternative field environment as this prototype evolves. In the future, we expect to explore issues related to using geospatial data for survey data collection and the impact of providing ad hoc access to information resources in a mobile survey environment. This includes augmenting conflation tools with field sampling technologies for natural resource settings.

Acknowledgements

This research is supported in part by funds provided via NSF Digital Government grant EIA-9983289 and the following collaborating agencies: Bureau of the Census, Bureau of Labor Statistics, U.S. Department of Agriculture (Forest Service, National Agricultural Statistics Service, Natural Resources Conservation Service), and U.S. Geological Survey.