Using Linear Programming Techniques for Balancing Automatic Case Assignments in Field Surveys

Edward English, Steven Pedlow
Statistics and Methodology, NORC, 55 E. Monroe St., Suite 2000, Chicago, IL, 60603

Abstract

National field surveys attempt to hire field interviewers who reside proximate to selected cases. Traditionally, case assignment was a very labor-intensive activity, with staff manually comparing national maps to determine where cases are in comparison with interviewers. NORC has been using an automatic distance-based case assignment procedure that generally assigns cases to the closest interviewer. However, many interviewers have been allocated unreasonably high work loads, while others had been assigned none. The current research uses linear programming to create constraints during allocation in order to lessen the degree of manual intervention and thus balance loads automatically. Our method realized an increase in operational efficiency without a significant loss in accuracy, and thus shows promise as a method for integrating GIS, programming, and survey research.

KEY WORDS: Area Probability Sample, Interviewer Allocation

1. Introduction

Fundamentally, all face-to-face surveys require the hiring of field staff and the subsequent assignment of sample in advance of field interviewing. Regional field managers have traditionally been responsible for executing both processes on their own. The manual exercise of traditional assignment-making had been challenged, however, by fundamental structural limitations. Complications result from the use of obsolete technology to manually match cases to the appropriate interviewer by ZIP code or city.

There was thus the potential to develop an automated method of case assignment that would consider distance and proximity within a database structure. This type of solution could be facilitated through the use of GIS technology. Geographic Information Systems, or 'GIS', embodies a set of tools that permits the automatic linkage of data sets based on geographic variables, such as distance, adjacency, or the quality of being within given census or postal areas.

Through the use of GIS, NORC developed a method to assign cases to interviewers based on distance. In this relatively simple approach, cases were automatically assigned the nearest interviewer. While the distance-based approach was a significant improvement over manual case assignment, it resulted in severely unbalanced caseloads. Specifically, some field interviewers could be assigned very large case loads while others were assigned no cases at all, as determined by the distribution of field interviewers and cases. The distance-only method therefore required a manual re-allocation step for a reasonable end result. One would thus significantly benefit by an automated method that would both be able to take advantage of GIS-calculated distances and automatically balance case allocation to generate optimal assignments.

The purpose of this paper is to describe our linear programming-based approach to optimizing case assignment. We integrate Geographic Information Systems (GIS) and mathematical programming in *SAS* to develop a novel method for field surveys. In so doing we broaden on earlier research (English and Pedlow, 2005) that emphasized the advantages of the distance method. In addition, we compare the linear-programming based method of auto-assignment to the distance-only approach previously undertaken. Lastly, we suggest constraints beyond distance that could be included to further enhance the results.

2. Background

Traditionally, the assignment of selected cases (household addresses) to field interviewing staff was a very labor-intensive activity. Field managers and/or central office staff consulted national maps to determine where the cases were in comparison to the staff actually hired. After hiring, field managers would then assign all cases in a particular ZIP code or city to a field interviewer with that, or a similar, ZIP code or city. The traditional method assumes that entities within the same ZIP code or city are more proximate than those between ZIP codes or cities, which is not always true. For example, if we had one interviewer in Salt Lake City and one in the suburb of West Valley City, a non-optimal assignment might be made for cases on the borders, especially if the interviewers were not centrally located within their cities. So, the traditional method does not ensure that cases are assigned to the closest interviewer. The consequence of this time-consuming approach was over-allocation of cases to certain field interviewers, and of general geographic inaccuracy and imprecision in areas of irregular ZIP code or city geographies.

For the 2004 Survey of Consumer Finances (SCF 2004), Pedlow and English implemented a first attempt at automatically assigning cases (and segments) to the nearest interviewer. In so doing they also provided further data (e.g., the closest 10 interviewers and their distances from each case/segment) for manual adjustments by statistical staff and the regional field managers. This first effort in automated case assignment also included a feedback loop that allowed field managers to adjust the allocations; processing derived information was determined to be time-consuming and inefficient. Results from these analyses were presented as part of an invited paper at JSM 2005 (English and Pedlow, 2005).

Subsequently, English and Pedlow have performed automatic case assignment for the list of National Longitudinal Survey of Youth 1997 Cohort (NLSY97) cases three times (2004-2006) with no field manager feedback loop. The process has become very efficient, but one limitation remains: some interviewers receive far too many cases, while others receive none. Some effort has gone into redistribution of the most extreme assignments, but it is a time-consuming endeavor relying on informal knowledge.

We would therefore benefit greatly by being able to automatically balance the interviewer loads to minimize the total distance from interviewer to case. Such a method would employ GIS-calculated distances and a programmatic means to automatically balance the cases. Linear programming represents a theoretical method to accomplish this task.

3. Linear Programming

The basic task for linear programming is to maximize or minimize a function subject to a set of constraints. For example, if one needed to schedule all JSM presentations and needed to prevent the same speaker from having more than one event in the same time slot, they could do so. Our application of linear programming to automatic case assignment attempted to minimize the sum of distance traveled by field interviewers to cases, subject to three constraints.

We use the matrix D to contain our input matrix of distances between every case and every interviewer:

$$\begin{array}{c} \\ \text{Case1} \\ \text{Case2} \\ \ldots \\ \text{Case}n \end{array} \begin{array}{c} \begin{array}{cccc} \text{FI1} & \text{FI2} & \ldots & \text{FI}m \end{array} \\ \begin{pmatrix} d_{11} & d_{12} & \ldots & d_{1m} \\ d_{21} & d_{22} & \ldots & d_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ d_{n1} & d_{n2} & \ldots & d_{nm} \end{pmatrix} \end{array}$$

This matrix has m x n entries, where m is the number of interviewers and n is the number of cases to be assigned.

We use the matrix A to contain the output matrix of assignments for each case to an interviewer:

$$\begin{array}{c} \\ \text{Case1} \\ \text{Case2} \\ \ldots \\ \text{Case}n \end{array} \begin{array}{c} \begin{array}{cccc} \text{FI1} & \text{FI2} & \ldots & \text{FI}m \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{pmatrix} \end{array}$$

All the entries in the matrix A are binary (the case is assigned to this interviewer or not). These m x n parameters are the unknown parameters solved via linear programming.
The function that we minimized is the total distance traveled by the interviewers in one visit to every case:

$$\text{minimize} \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij} a_{ij}$$

There are two important constraints that are key to our linear programming effort. Firstly, every case needed to be assigned to exactly one field interviewer:

$$\text{for every j,} \quad \sum_{i=1}^{n} a_{ij} = 1$$

Secondly, no field interviewer could be assigned to "too many" cases, which was determined from past experience to be any caseload greater than 75:

$$\text{for every i,} \quad \sum_{j=1}^{m} a_{ij} \leq 75$$

Our third (optional) constraint was that field interviewers could only be assigned cases within their same administrative region, of which there were 12 nationally. Our application of this constraint was to change every d_{ij} entry where the region of the interviewer was different from the region of the case to 9999, a large enough distance to prevent the assignment of any case to an interviewer from a different region.

Linear programming algorithms can then solve this series of simultaneous equations iteratively.

4. Methodology

We used data from the 2007 Survey of Consumer Finances (SCF) as input to the auto-assignment procedure. The 2007 SCF had an initial sample draw of 8,902 cases spread nationally and 153 field interviewers hired to administer them, and so an average of 58.2 cases per interview. Both sets of addresses were geocoded using the *MapMarker Plus* geocoding package for *MapInfo Professional*. Then, we developed software in *SAS* to do the two steps necessary for automatic case assignment. First, the software calculated the spherical distance between all cases and all field interviewers in a 153 by 8,902 matrix. Second, the software used PROC LP to conduct the automatic assignment subject to our above three constraints.

For purposes of comparison, we also conducted "distance only" assignment using an algorithm developed in *MapBasic* and implemented in *MapInfo Professional*.

5. Results and Discussion

Figure 1 shows the distribution of cases per field interviewer, using the Distance-only method. These results contrast with figure 2, which shows those automatically assigned using linear programming.

Figure 1- Cases per Field Interviewer Using the Distance-only Method

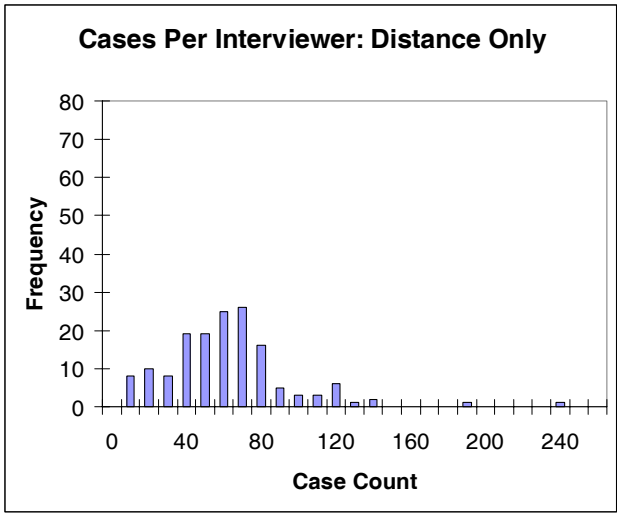

Figure 1 shows that most interviewers received 90 cases or fewer with the distance-only method, but many received 120 or more. If a caseload of up to 75 is considered to be reasonable, one interviewer received more than what three interviewers should be assigned.

Figure 2- Cases per Field Interviewer Using the Linear Programming Method

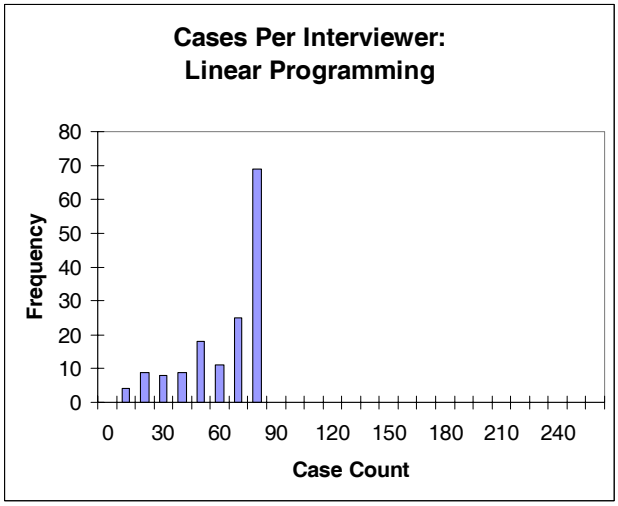

Figure 2 shows that using linear programming almost half of the interviews were assigned the maximum of 75 cases. In addition, many interviewers were assigned an impractically low number of cases. It is theoretically possible to set a minimum load constraint, but there is a subset of interviewers who are hired to travel to a specific area, so it is preferable to assign these interviewers zero cases.

As an example of how the differences between the two methods are manifest in practice, Figures 3 and 4 show a map of the Philadelphia, Pennsylvania area. Field interviewers are illustrated by the human icons; an icon is inverted to indicate they are in a different region than Philadelphia. Cases are colored to show the interviewer they are assigned to.

Figure 3- Assigned Cases from Distance-only Method

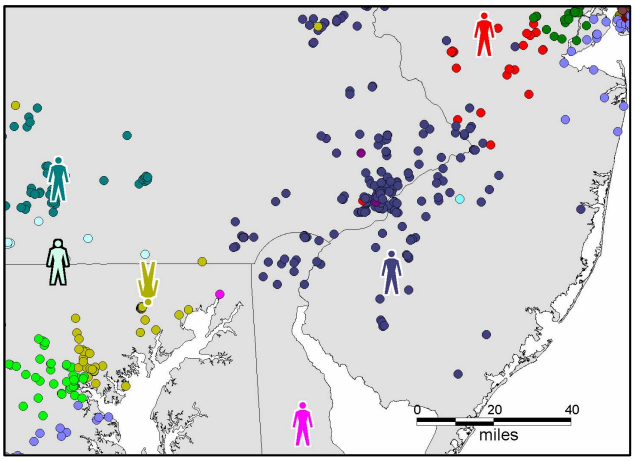

Figure 3, the results of the distance-only method, shows that there is one interviewer in southern New Jersey that would be assigned all (dark blue) cases in Philadelphia. Clearly, it would be necessary to manually re-allocate cases in this area to avoid an impractical assignment. Note that there are a few cases in metro Philadelphia that are not dark blue as they are still associated with different regions due to recent movement.

Figure 4- Assigned Cases from Linear Programming Method

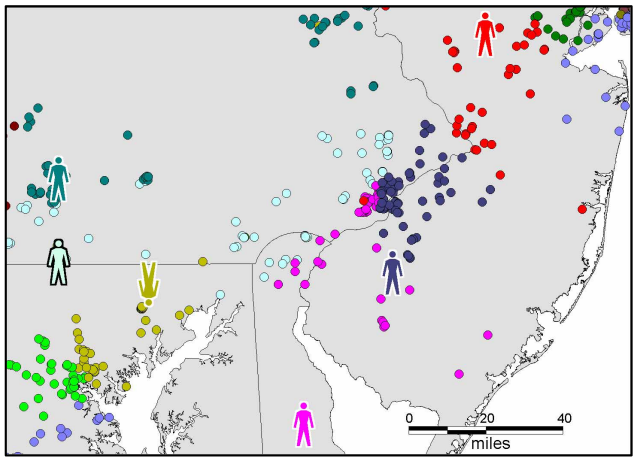

Figure 4 shows the Philadelphia cases equitably allocated between interviewers in New Jersey, Delaware, and Maryland. Realistically, both approaches demonstrate the need to hire at least one additional interviewer in south eastern Pennsylvania. Nonetheless, the contrast indicates the benefit of the integrated case balancing that accompanies linear programming.

In fact, for the distance-only method used in 2006, about 20% of auto-assigned cases were transferred to a new FI during the first few days of data collection. Using the linear programming-based method in 2007, only about 10% of auto-assigned cases were moved to a new FI during the first days of data collection, cutting in half the need for quick re-allocation.

Implementation was time-consuming as a result of the matrices for processing being of size 8,902 by 153, and thus more than 1.3 million parameters. Two factors that may have increased the number of iterations needed were the third optional constraint by field region and the binary output matrix. We were able to simply increase the maximum number of iterations from the default, however, to obtain a successful outcome. Set-up was also tedious, but new SAS procedures like PROC OPTMODEL should be easier to use.

6. Conclusions

We argue that our experience with automated case assignment demonstrates that it is an effective replacement for the traditional method, especially with the added efficiency brought by linear programming. Even with the automated balancing feature, it is still recommended to have a feedback loop with the actual field staff in order to incorporate valuable holistic knowledge.

It would be possible to make improvements to the approach by adding more relevant variables and constraints. A first priority would be to determine distance along road networks rather than the "crow flies" distances calculated thus far. Using road networks should improve overall accuracy to some degree, especially in areas with impassable natural features such as water bodies or mountains. Also, we could add additional constraints to improve the general utility, such as matching language, race-ethnicity, or relative difficulty of cases to the appropriate interviewers. Nonetheless, we feel our current simple approach is effective as implemented

References

English, Ned, and Steven Pedlow. Using GIS to Improve Field Interviewing Efficiency: Enhanced Interviewer Selection and Sample Allocation. 2005 Proceedings of the American Statistical Association, [CD-ROM].

Murty, Katta G. 1983. Linear Programming. John Wiley & Sons: New York.