

# Simulating NHIS Field Operations\*

**Bor-Chung Chen**

Federal Railroad Administration<sup>†</sup>

1200 New Jersey Avenue, SE, Washington, DC 20590

Bor-Chung.Chen@dot.gov

## 1 Introduction

Discrete-event simulation modeling has become the most commonly used tool for performance evaluation of stochastic dynamic systems in science and engineering, including such complex systems as manufacturing and material handling systems ([17] and [19]), logistics and transportation systems ([14]), healthcare and service systems ([15] and [11]), computer and communication systems ([9]). These applications of simulation modeling are results of significant achievements in electronic and computer technologies that have led to broad proliferation of powerful computers and computer networks, and significant achievements in software technology, that have resulted in simple but very efficient human-computer interfaces. However, no technological innovation can release simulators from their responsibility of ensuring that their simulation experiments produce credible final results.

In this paper, we will describe how to use the simulation techniques for the field operations application of a national household survey. We will discuss main problems and solutions of quantitative stochastic discrete-event simulation, i.e. the stochastic simulation in which the emphasis is put on statistical correctness of the final results. Whole spectrum of the problems will be covered: from generators of uniformly distributed pseudo-random numbers, which play the role of original sources of randomness in stochastic simulation, to methods of generation of system variables, such as interview length, contact time, in field representative's visits of sample households.

At the U.S. Census Bureau, the mission of the Field Division is to collect quality data at the right time for the lowest cost. Therefore, there is a need to have a valid method of predicting cost, response rates, and timing of new or continuing surveys for the field operations (per discussions with Bitzer ([3]) and others). This project is intended to develop such a method.

In complex field operations for a household survey or census, the scheduling function is typically concerned with determining the starting time and the sequence of visiting the cases assigned to the interviewers in which system performance is to be optimized. The system performance is defined as controlling the cost and timing and maximizing the response rates. The complexity and practical importance of the field operations scheduling problem has motivated the development of models appropriate for a broad range of surveys and censuses, and has focused attention on the impact of scheduling decisions on contact time and travel time. Most importantly, the model would provide a tool for predicting costs, response rates, and timing before the survey begins.

Currently, regression analysis is used on the data set from the Field Division's CARMN (Cost And Response Management Network) and Population Division's Planning Data Base ([16]) to explore survey-related cost drivers for the CPS (Current Population Survey). Also, Shimizu and Lan [18] use a simplified overall cost model based on the NHIS multistage sample.

---

\*This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau or Federal Railroad Administration.

<sup>†</sup>This research had been done while the author was affiliated with the U.S. Census Bureau.

For example, one of the system performance measures is the cost. If  $TDC$  is the total direct cost of the operation, then

$$TDC = \sum_{i=1}^r \sum_{j=1}^d C_{ij}, \quad (1)$$

where  $C_{ij}$  is the daily cost of FR (Field Representative or interviewer)  $i$  on  $j$ th day,  $r$  is the number of FRs, and  $d$  is the number of days.

$$C_{ij} = (H_{ij} \times R_{ij}) + (M_{ij} \times P_{ij}), \quad (2)$$

$H_{ij}$  = Total Hours

$R_{ij}$  = Hourly Rate

$M_{ij}$  = Total Mileage

$P_{ij}$  = Reimbursement Per Mile.

To accurately predict the total time spent for each FR each day, we further decompose the total hours,  $H_{ij}$ , into traveling time, contact time, interview time, and so on. Each time segment has its own statistical distribution. For example, the traveling time follows a statistical distribution to be determined given the traveling distance. We will describe how to obtain the statistical distribution for each time segment in the next section.

The objective of the project is to build a model that will take the inputs of workload, staffing, traveling time, productivity, etc., for the field operations of a survey. The output of the model will be the cost, response rates, and timing to complete the operations based on the assumptions given to the input of the model. The performance measures identified are, therefore, the cost, response rates, and timing. The final goal of the field operations is to have Low cost, High response rates, and Short timing; it is referred to as *LHS*.

The current version of the model contains about 1888 lines of C++ code. The compiler used is Microsoft Visual Studio .NET 2003. Section 2 gives a brief description of the simulation modeling concept. Section 3 describes the proposed approaches for the project. Section 4 gives a description of how to perform the input modeling. Section 5 discusses random number and random variate generations. Section 6 briefly describes the simulation model for the simplified field operations. Section 7 provides preliminary results of simulation runs with different seeds and gives an example of sensitivity analysis of the model. Section 8 is the conclusion and summary.

## 2 Preliminary Concept for Simulation Modeling

The concept of simulation modeling for the field operations is described in this section. The reader who is familiar with simulation modeling should skip ahead. Simulation is the use of computations to implement a model of some dynamic system or phenomenon, such as field operations. It is using a model implemented as a computer program, rather than experimenting with a real system. In order to study the field operations scientifically, we need to make a set of assumptions about how the operations work. These assumptions are usually in the form of mathematical or logical relationships, and are therefore called mathematical or logical models. A valid model will help decision makers gain some understanding of how the system behaves. The field operations system is too complex to allow us to obtain a realistic model to be evaluated analytically. Instead, a simulation solution is obtained by using a computer to evaluate the model *numerically* over a time period of interest, and data are collected to *estimate* the desired characteristics of the model, i.e., the operating cost, response rates, timing, etc.

We would like to propose a simulation model for the field operations. Specifically, we would like to use the discrete event simulation technique to model the field operations system. Discrete event simulation concerns

the modeling of a system as it evolves over time by a representation in which the variables change only at a countable number of points in time. These points in time are the ones at which an event occurs, where an *event* is defined as an instantaneous occurrence which may change the state of a system. In this paper, simulation will be used to describe and analyze the behavior of the field operations system, ask what-if questions about the system, and aid in the modification of the field operations when needed. A sensitivity analysis will also be conducted to find out which potential solutions are the most cost effective methods to implement.

Although a discrete event simulation is usually done by computer, an example of simulation by hand is given in [4]. If a simulation of field operations is done by hand, there is a limit to the complexity of the operations that can be solved in this manner. Also the number of cases and the number of FRs that must be included in the simulation model could be much larger and the number of times that the simulation must be run for statistical purposes (output analysis) could be large. Therefore, the simulation modeling using computers for the field operations is appropriate.

### 3 Proposed Approaches

So far, no similar work has been found in the literature describing the analytical or simulation modeling of the operations. The field operation is a unique system in the operations research field. Developing an analytical model for the operation requires an extensive investigation of the operation itself as well as the investigation of the operations research techniques. In this project, we will use a computer simulation model based on the concept described in Section 2 and this section.

As indicated before, we would like to propose a simulation model for the field operations. Developing a valid simulation model involves three basic entities: the real system under consideration (the field operations for a particular survey); a theoretical model of the real system; and a computer-based representation of the model, the simulation program. The activities of developing a theoretical model from the real system are referred to as simulation modeling, and the activities of developing a computer-based representation for the theoretical model are referred to as simulation programming. We will use C++ as the programming language. C++ is an object-oriented programming language that can be used to program an FR as an object. FRs are the key persons to the success of the operations. The object-oriented simulation is a technique to view the real system as being composed of various objects ([10]). The FR objects will be the core component of the simulation model of the field operations. Other C++ classes related to the object-oriented simulation, such as random number generation class, will be defined as well.

In the simulation programming, we will concentrate on modeling the behavior of interacting objects, such as FRs, respondents, etc., over time. The behavior of the interaction also involves other important steps in the simulation modeling and programming: random numbers and random variates generation, input data analysis, output data analysis, etc.

The example in Section 2 used input values that were generated by spinners and a coin. In computer simulation, the computer will generate independent random numbers that are distributed continuously and uniformly between 0 and 1 [i.e.,  $U(0,1)$ ]. These random numbers can be converted to the desired statistical distributions, or random variates. The random numbers and random variates generation can be easily implemented with the object-oriented C++ language because the *class* definition in C++ can determine the objects' (random variates') characteristics or properties.

Input data analysis is another important step in the simulation modeling and programming. Input data modeling uses statistical methods to determine the desired statistical distributions needed for the random numbers and random variates generation. We will give a more detailed description of input data analysis in Section 4.

The analysis of simulation output begins with the selection of performance measures. As indicated before, the performance measures of interest in field operations are cost, response rates, and timing. The primary purpose of most simulation studies is the approximation of prescribed system parameters with the objective of identifying parameter values that optimize some system performance measures. Because some of the input

processes in the field operations simulation study are random, the output data are also random and runs of the simulation result in *estimates* of performance measures. Unfortunately, a simulation run does not usually provide independent, identically distributed (IID) observations; therefore, “classical” statistical techniques are not directly applicable to the analysis of simulation output. Statistical techniques of the simulation output analysis can be found in [1].

Finally, we will also perform a sensitivity analysis to determine the impact on the performance measures if some of the input variables (parameters) can be controlled. This analysis is valuable in determining what types of potential solutions are the most cost effective to implement. It will also be a feasibility study to determine the limitations of the simulation modeling applied to the field operations.

## 4 Input Data Analysis

The most difficult aspect of simulation input modeling is gathering data of sufficient quality, quantity, and variety to perform a reasonable analysis. After a preliminary study, we have identified part of the required and available data sets as described below. If a data set is not available for the project, we will have to make reasonable assumptions with help from the subject matter experts of Field Division. The following is a list of data sets so far identified and required for the project:

1. the average speed distribution for an FR driving between households;
2. the time distribution for an FR to make contact with respondents;
3. the time distribution for an FR to complete an interview if the respondent is contacted;
4. contact histories ([2]).

A random input variable to a simulation model can be viewed as a *stochastic process*. A stochastic process is often defined as a collection of random variables. In simulation modeling, the strongest assumptions of a stochastic process that we can make are: (1) all of the random variables are probabilistically *independent* of one another; (2) all of the random variables follow the same probability distribution and thus are said to be *identically distributed*. In other words, FR’s are independent and follow the same rules. Also, the same type of random variables associated with each FR are also independent and follow the same rules. For example, the interview time of respondent A conducted by a particular FR is independent and identically distributed as that of respondent B conducted by the same FR. Therefore, we propose the following methods to perform the input data modeling for the simulation study: (a) tables and/or plots of estimated lag (linear) correlations and (b) scatter diagrams for assessing independence and stability of distributions [20].

The input data modeling also includes fitting a probability distribution to the data. We will assume that distributions are defined by their distribution functions, or equivalently, by their related density (continuous) or mass (discrete) functions. If the “best” of the fitted distributions provides a reasonable representation of the data, we will use it in the simulation. Otherwise, an empirical distribution will be used to represent the data directly.

### 4.1 Outcome Frequency Distribution of NHIS

In NHIS (National Health Interview Survey), each sample household is assigned one of the 28 outcomes after the visits of FRs. Table 1 lists the 28 possible outcomes and their frequency distribution from the 2004 interviews.

The following are the definitions used in the NHIS surveys:

- **Eligible cases** = total cases – (Type B’s + Type C’s);
- **Complete cases** = 201’s + 203’s;

Table 1: The 2004 NHIS Frequency Distribution by Outcome

<i>i</i>	Outcome	Freq.	Original		Adjusted	
			$\%(f_i)$	Cumul.	$\%(g_i)$	Cumul.
1	201 (completed interview)	30992	43.58	43.58	44.48	44.48
2	203 (sufficient partial interview, no follow-up)	5916	8.32	51.90	8.49	52.97
Type A						
3	213 (language problem)	83	0.12	52.01	0.12	53.09
4	215 (insufficient partial interview)	519	0.73	52.74	0.74	53.83
5	216 (no one home, repeated calls)	1224	1.72	54.46	0.00	53.83
6	217 (temporarily absent, no follow-up)	312	0.44	54.90	0.45	54.28
7	218 (refused)	2604	3.66	58.56	3.74	58.02
8	219 (other-Type A)	570	0.80	59.36	0.82	58.83
Type B						
9	223 (all arm force)	122	0.17	59.54	0.18	59.01
10	225 (all URE)	934	1.31	60.85	1.34	60.35
11	226 (vacant, nonseasonal)*	6330	8.90	69.75	8.90	69.25
12	228 (to be demolished)*	243	0.34	70.09	0.34	69.59
13	229 (under construction)*	245	0.34	70.44	0.34	69.94
14	230 (temporarily business or storage)*	203	0.29	70.72	0.29	70.22
15	231 (unoccupied site)*	226	0.32	71.04	0.32	70.54
16	232 (construction not started)*	41	0.06	71.10	0.06	70.60
17	233 (other-Type B)*	138	0.19	71.29	0.19	70.79
18	235 (vacant, seasonal)*	1295	1.82	73.11	1.82	72.61
19	236 (screened out)	13813	19.42	92.53	19.82	92.43
Type C: unit is not there						
20	240 (demolished)*	236	0.33	92.87	0.33	92.77
21	241 (house-trailer mover)*	193	0.27	93.14	0.27	93.04
22	242 (out of segment bounds)*	137	0.19	93.33	0.19	93.23
23	243 (converted permanent business/storage)*	332	0.47	93.80	0.47	93.70
24	244 (merged)*	170	0.24	94.04	0.24	93.94
25	245 (condemned)*	26	0.04	94.07	0.04	93.97
26	246 (built after 4/1/1990)	3418	4.81	98.88	4.91	98.88
27	247 (other-Type C)*	261	0.37	99.24	0.37	99.24
28	248 (spawned in error)*	537	0.76	100.00	0.76	100.00
Total		71120	100.00	100.00	100.00	100.00

- **Response rate** =  $1 - (\text{Non-response rate})$ ;
- **Non-response rate**: proportion of eligible cases that were noninterviews (Type A's)

$$\text{Equation} = \frac{\text{Type A's}}{\text{Eligible Cases}} \times 100; \quad (3)$$

- **Interview rate**: proportion of eligible cases that were completed interviews (outcome = 201)

$$\text{Equation} = \frac{\text{201's}}{\text{Eligible Cases}} \times 100; \quad (4)$$

- **Partial rate**: proportion of eligible cases that were sufficiently completed interviews (outcome = 203)

$$\text{Equation} = \frac{\text{203's}}{\text{Eligible Cases}} \times 100; \quad (5)$$

- Therefore, **Response rate** = **Interview rate** + **Partial rate**.

In the simulation study, the outcome frequency distribution needs to be adjusted and taken as the input of the simulation model. Some of the 28 outcomes can be determined when the FR visits only once regardless of the result of contact or no-contact. These outcomes are called one-visit outcomes. We have identified the following outcomes as the one-visit outcomes:

226 (vacant, nonseasonal)	240 (demolished)
228 (to be demolished)	241 (house-trailer mover)
229 (under construction)	242 (out of segment bounds)
230 (temporarily business or storage)	243 (converted permanent business/storage)
231 (unoccupied site)	244 (merged)
232 (construction not started)	245 (condemned)
233 (other-Type B)	247 (other-Type C)
235 (vacant, seasonal)	248 (spawned in error)

The other outcomes, except 216 (no one home), are determined as soon as the respondent is contacted. These outcomes are called contact outcomes and 216 is called no-contact outcome.

In the simulation model, we will assume a zero probability of the no-contact outcome. The final percentage of the no-contact outcome is determined by the contact/no-contact distribution discussed in Section 4.3. Therefore, We need to adjust the frequency distribution of the contact outcomes and make the no-contact outcome 0.0% for the simulation modeling. We will keep the one-visit outcome distribution unchanged and assume that the percentage of the no-contact outcome is redistributed according to the ditribution of the contact outcomes. Let  $V$  be the index set of the one-visit outcomes and  $U$  be the other outcomes. Also, let  $f_i$  be the percentage of the  $i$ th outcome, then

$$\sum_i f_i = 100.0$$

and the adjusted frequency distribution of the outcomes is computed as following:

$$g_i = \begin{cases} 0.0, & \text{if } i = 5; \\ f_i \times \frac{\sum_{j \in U} f_j}{\sum_{k \in U - \{5\}} f_k}, & \text{if } i \in U - \{5\}; \\ f_i, & \text{if } i \in V. \end{cases}$$

Table 1 also shows the 2004 adjusted frequency distribution by outcome at the national level. In the table, the one-visit outcomes are marked with an asterisk(\*). The final percentage of 1.72% for code 216 (no one home) is determined by the distribution of contact/no-contact. Each of the 12 regional offices (and eventually, each PSU) will be handled in the same way. Table 2 shows the 2004 Quarter 2 (Q2) frequency distribution and its adjusted frequency distribution by outcome for the Denver Regional Office.

## 4.2 Interview Length Distributions of Outcomes in NHIS

In this section, we will try to decide what general family of distributions appears to be appropriate for each outcome's interview length. The methods we use for this purpose are scatter diagrams and probability plots. A scatter diagram is constructed for assessing the independence of observations and was described earlier in Section 4. A probability plot is a graphical comparison of an estimate of the distribution function of the interview length data  $X_1, X_2, \dots, X_n$  with the distribution function of one of the standard distributions being considered as a model for the data. Before we perform the input analysis using the probability plots and other methods, we would like to remove the outliers from the observed data. Some of the data are not good and considered as outliers for a variety of reasons. One of the reasons is that some of the observed data gave much longer time than the actual interview time because the computer was kept running without "the end of interview" being entered at the end of interview. Another is that some of the interview lengths are negative values.

There are  $m = 30992$  observations for outcome 201 (completed interview) of NHIS in 2004. To remove the outliers, we have truncated the observations that are beyond two standard deviations from the mean. The truncation of removing the outliers has been repeated 4 times ( $k = 4$  iterations) for outcome 201. The final number of observations used in the input analysis for outcome 201 is  $n = 26741$ . Table 3 shows the numbers of  $m$ ,  $k$ , and  $n$  for each of the outcomes. The entries with "-" indicate that there was no interview time needed even though some observations were still captured.

Table 2: 2004 Q2 NHIS Frequency Distribution by Outcome (Denver RO)

$i$	Outcome	Freq.	Original		Adjusted	
			$\%(f_i)$	Cumul.	$\%(g_i)$	Cumul.
1	201 (completed interview)	512	46.80	46.80	47.10	47.10
2	203 (sufficient partial interview, no follow-up)	70	6.40	53.20	6.44	53.54
Type A						
3	213 (language problem)	0	0.00	53.20	0.00	53.54
4	215 (insufficient partial interview)	3	0.27	53.47	0.28	53.81
5	216 (no one home, repeated calls)	6	0.55	54.02	0.00	53.81
6	217 (temporarily absent, no follow-up)	8	0.73	54.75	0.74	54.55
7	218 (refused)	28	2.56	57.31	2.58	57.13
8	219 (other-Type A)	5	0.46	57.77	0.46	57.59
Type B						
9	223 (all arm force)	4	0.37	58.14	0.37	57.95
10	225 (all URE)	32	2.93	61.06	2.94	60.90
11	226 (vacant, nonseasonal)*	86	7.86	68.92	7.86	68.76
12	228 (to be demolished)*	2	0.18	69.10	0.18	68.94
13	229 (under construction)*	5	0.46	69.56	0.46	69.40
14	230 (temporarily business or storage)*	9	0.82	70.38	0.82	70.22
15	231 (unoccupied site)*	2	0.18	70.57	0.18	70.40
16	232 (construction not started)*	2	0.18	70.75	0.18	70.59
17	233 (other-Type B)*	0	0.00	70.75	0.00	70.59
18	235 (vacant, seasonal)*	17	1.55	72.30	1.55	72.14
19	236 (screened out)	198	18.10	90.40	18.21	90.35
Type C: unit is not there						
20	240 (demolished)*	5	0.46	90.86	0.46	90.81
21	241 (house-trailer mover)*	4	0.37	91.22	0.37	91.18
22	242 (out of segment bounds)*	0	0.00	91.22	0.00	91.18
23	243 (converted permanent business/storage)*	2	0.18	91.41	0.18	91.36
24	244 (merged)*	1	0.09	91.50	0.09	91.45
25	245 (condemned)*	2	0.18	91.68	0.18	91.63
26	246 (built after 4/1/1990)	83	7.59	99.27	7.64	99.27
27	247 (other-Type C)*	4	0.37	99.63	0.37	99.63
28	248 (spawned in error)*	4	0.37	100.00	0.37	100.00
Total		1094	100.00	100.00	100.00	100.00

#### 4.2.1 Assessing Independence of Interview Length of Outcomes in NHIS

A lag  $k$  correlation plot and a scatter diagram could have been constructed to assess the independence of the interview lengths. However, the interviews were conducted by different interviewers (FRs) at different households and at different time as described earlier in Section 4. Therefore, it is reasonable to assume that the interview lengths conducted by different FRs at the sample households for all outcomes are independent samples.

#### 4.2.2 Probability Plots of Interview Length of Outcomes in NHIS

Let  $X_{(i)}$  be the smallest of the  $X_j$ 's, called the  $i$ th order statistic of the  $n$   $X_j$ 's. The distribution function  $F$  of a random variable  $X$  is defined so that for any  $x$ ,  $F(x) = P\{X \leq x\}$ . If  $X$  has the same distribution as the  $X_j$  data, a reasonable approximation to  $F(x)$  is thus the proportion of the  $X_j$ 's that are less than or equal to  $x$ . Therefore, we might want to define an empirical distribution function  $\tilde{F}_n(x)$  so that  $\tilde{F}_n(X_{(i)}) = i/n$ , or  $\tilde{F}_n(X_{(i)}) = (i - 0.5)/n$  to avoid an empirical distribution function that is equal to 1 for a finite value of  $x$ .

For  $0 < q < 1$ , the  $q$  quantile of a distribution  $F$  is a number  $x_q$  that satisfies  $F(x_q) = q$ . Thus, if  $F^{-1}$  denotes the inverse of the distribution function  $F$ , a formula for the  $q$  quantile of  $F$  is  $x_q = F^{-1}(q)$ , where  $F^{-1}$  exists if  $F$  is continuous and strictly increasing. If  $F$  and  $G$  are two distribution functions, it is clear that  $F = G$  if and only if each of the quantiles of  $F$  is the same as the corresponding quantile of  $G$ . Thus, if

Table 3: Number of Observations Used for Input Analysis

outcome	$m$	$k$	$n$	outcome	$m$	$k$	$n$
201	30992	4	26741	231	226	—	—
203	5916	4	5055	232	41	—	—
213	83	2	78	233	138	—	—
215	519	3	460	235	1295	—	—
216	1224	—	—	236	13813	3	12470
217	312	—	—	240	236	—	—
218	2604	—	—	241	193	—	—
219	570	3	496	242	137	—	—
223	122	3	101	243	332	—	—
225	934	3	866	244	170	—	—
226	6330	—	—	245	26	—	—
228	243	—	—	246	3418	3	3120
229	245	—	—	247	261	—	—
230	203	—	—	248	537	—	—

$x_q$  and  $y_q$  are the  $q$  quantiles of  $F$  and  $G$ , respectively, a plot of the points  $(x_q, y_q)$  for various values of  $q$  will produce points along a straight line having slope 1 (a  $45^\circ$  line) and passing thru the origin, since  $x_q = y_q$  for all  $q$ . Furthermore, if the random variables corresponding to  $F$  and  $G$  differ only in location and scale, then for some real numbers  $\gamma$  and  $\beta > 0$ , we have  $G(x) = F((x - \gamma)/\beta)$  for all  $x$ . In this case, it is easy to see that for all  $q$ ,  $y_q = \gamma + \beta x_q$ , so that a plot of the points  $(x_q, y_q)$  produces a straight line of points which has a slope not necessarily 1 and which need not pass thru the origin. Thus, distributions having the same shape (but which may differ in location and scale) have quantiles which are linearly related. A plot of pairs of quantiles such as  $(x_q, y_q)$  is called probability plot, or Q-Q plot.

Probability plots provide a way of assessing whether the empirical distribution function  $\tilde{F}_n$ , defined at the  $X_{(i)}$  points, has the same shape as a distribution function from one of the theoretical families (see [12] for more details). For survey interview length, suppose that we are considering a particular distribution form and that if this distribution has shape parameters, they have already been estimated from the data. Let the resulting distribution function be denoted by  $F$ , which represents a trial hypothesized distribution shape, with unspecified location and scale. We would like to compare  $\tilde{F}_n$  with  $F$ , and we can do so by a (Q-Q) probability plot of the quantile pairs for  $q = (i - 0.5)/n$  for  $i = 1, 2, \dots, n$ , as follows. By definition, the  $(i - 0.5)/n$  quantile of  $\tilde{F}_n$  is precisely  $X_{(i)}$ . The  $(i - 0.5)/n$  quantile of  $F$  is simply  $F^{-1}((i - 0.5)/n)$ . Thus, we plot the points

$$\left( X_{(i)}, F^{-1}\left(\frac{i - 0.5}{n}\right) \right)$$

for  $i = 1, 2, \dots, n$ , and if the resulting points appear to lie along a straight line, we have informal confirmation that, except for adjustments in location and scale,  $F$  is a good distribution function for our interview length data.

Figure 1 shows the beta distribution probability plot for interview length of outcome 201. The plot indeed appears to have a straight line, supporting the beta distribution. To provide an idea of what a probability might look like when an inappropriate distribution is hypothesized, we made probability plots for the Weibull and gamma distributions. The resulting Weibull probability plot in Figure 2 displays obvious nonlinearity at the upper end while the gamma plot in Figure 3 displays nonlinearity at both ends.

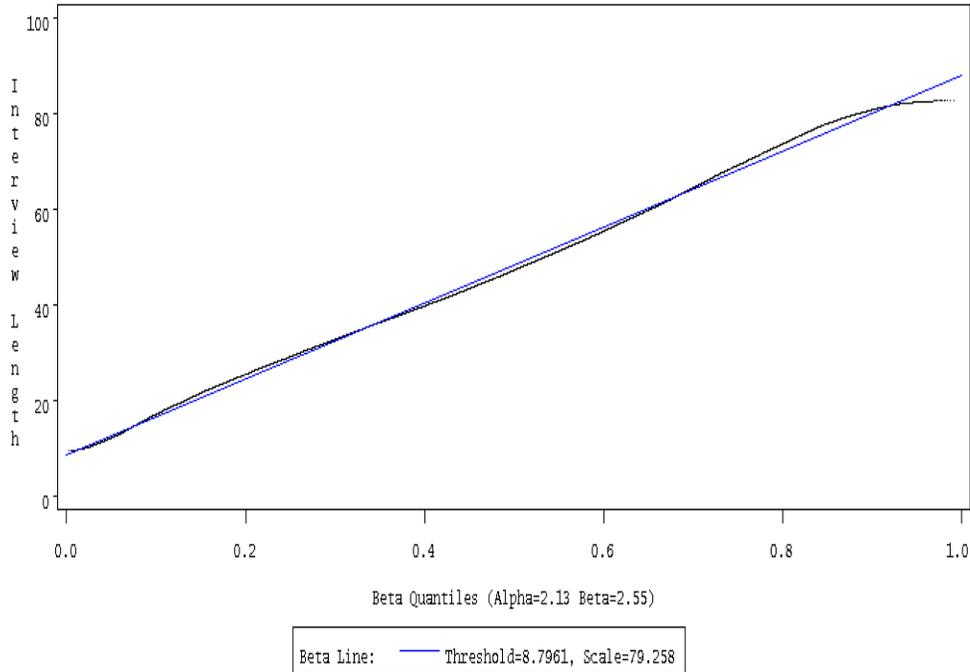


Figure 1: The Beta Probability Plot for Interview Length of Outcome 201.

#### 4.2.3 Estimation of Parameters for Interview Length of Outcomes in NHIS

After a family of distributions has been hypothesized, we must specify the value(s) of its parameter(s) in order to determine completely the distribution from which we shall sample during the simulation. Our hypothesized distribution is a beta distribution,  $\text{Beta}(\alpha, \beta, \theta, \lambda)$ , where  $\alpha$  and  $\beta$  are the shape parameters,  $\theta$  is the threshold parameter, and  $\lambda$  is the scale parameter. For both distributions,  $\alpha > 0$  is the shape parameter,  $\theta$  is the threshold parameter, and  $\lambda > 0$  is the scale parameter.

Table 4: Parameter Estimates for the Three Distributions and Outcome 201 Data

Distribution	Range	Parameters			
		Shape( $\alpha$ )	Shape( $\beta$ )	Threshold( $\theta$ )	Scale( $\lambda$ )
$\text{Beta}(\alpha, \beta, \theta, \lambda)$	$\theta < x < \theta + \lambda$	2.127	2.549	8.796	79.258
$\text{Weibull}(\alpha, \theta, \lambda)$	$x > \theta$	2.484		6.566	43.030
$\text{Gamma}(\alpha, \theta, \lambda)$	$x > \theta$	17.392		-24.802	3.997

The parameter estimates using the *maximum-likelihood estimators* (MLEs) for the three distributions and the interview lengths of outcome 201 are given in Table 4. Other distributions used for testing other outcome data are exponential and lognormal distributions.

#### 4.2.4 Goodness-of-Fit Tests for Interview Length of Outcomes in NHIS

After we have hypothesized a distribution form for our data and have estimated its parameters, we must examine whether the fitted distribution is in agreement with our observed data  $X_1, X_2, \dots, X_n$ . If  $F(x)$  is the distribution function of the fitted distribution, a hypothesis test is addressed with a null hypotheses of

$$H_0 : \text{The } X_i\text{'s are IID random variables with distribution function } F(x) \quad (6)$$

This is called a *goodness-of-fit test* since it tests how well the fitted distribution “fits” the observed data. We

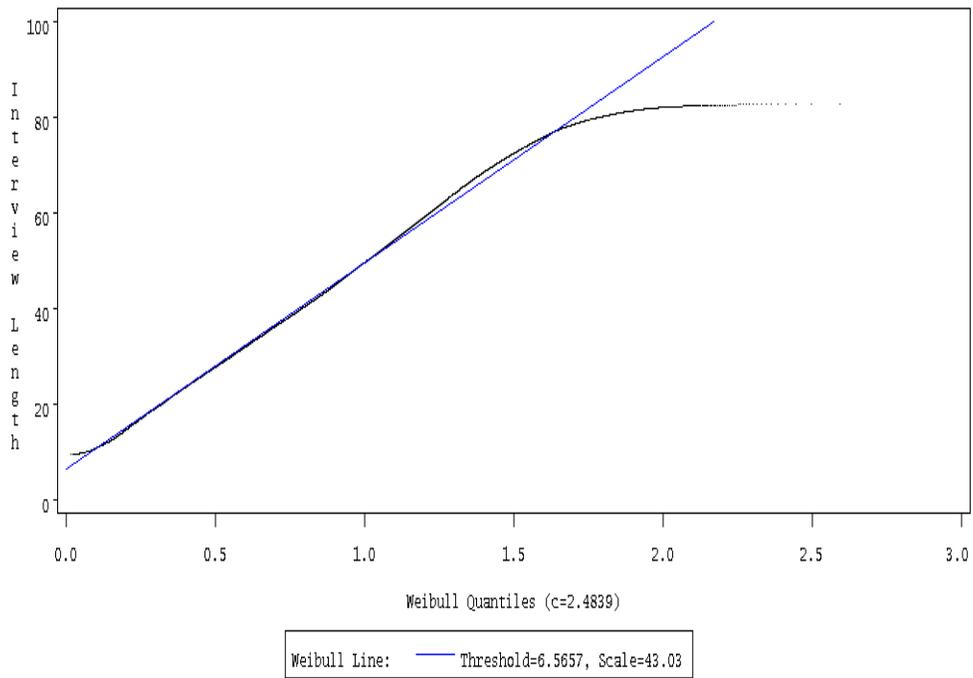


Figure 2: The Weibull Probability Plot for Interview Length of Outcome 201.

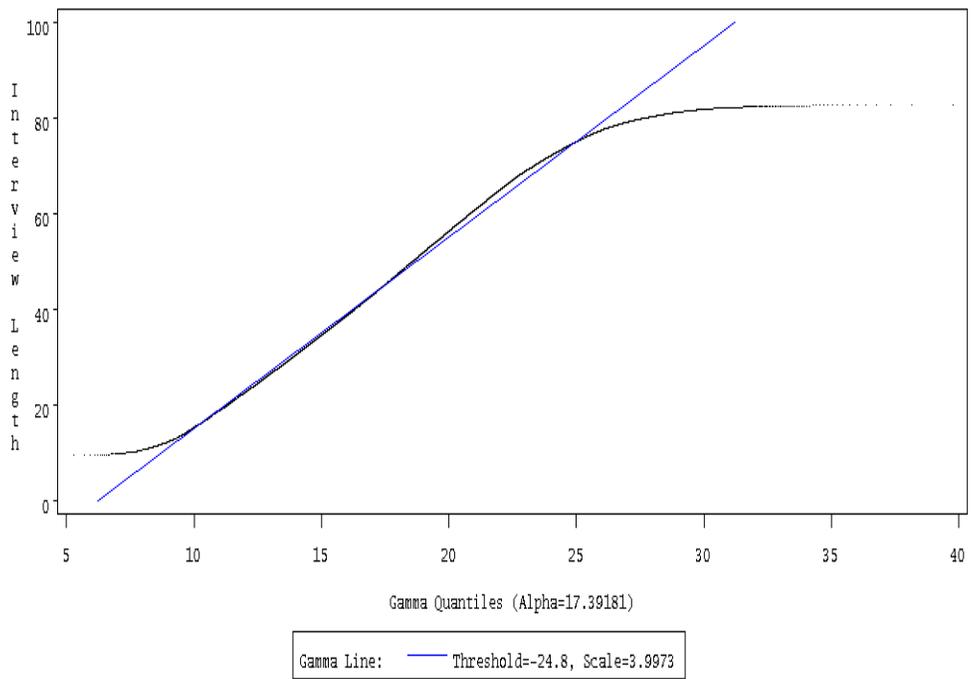


Figure 3: The Gamma Probability Plot for Interview Length of Outcome 201.

used four goodness-of-fit test methods to perform the tests. The four methods used are (1) Chi-Square Tests, (2) Kolmogorov-Smirnov Tests, (3) Cramér-von Mises Tests, and (4) Anderson-Darling Tests.

Figure 4 shows the histogram of the interview lengths for outcome 201. Table 5 show the results of the four goodness-of-tests, indicating that the outcome 201 interview lengths do not fit the three distributions tested. In the simulation, we will use an empirical distribution from which the samples are drawn.

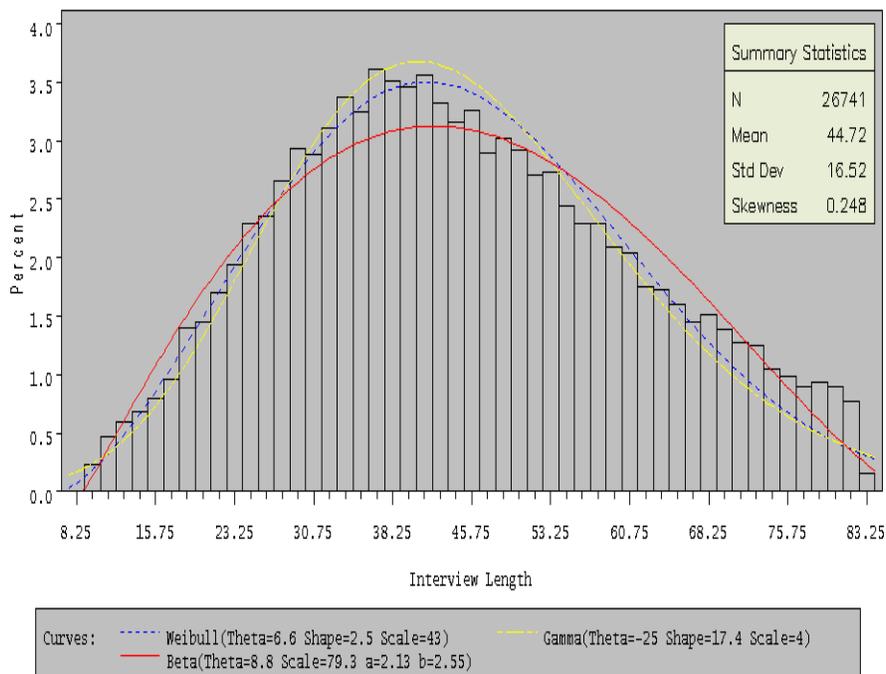


Figure 4: The Histogram Plot for Interview Length of Outcome 201 (Beta, Weibull, and Gamma Distributions).

#### 4.2.5 Data Analysis for Interview Lengths of Other Outcomes in NHIS

The scatter diagrams, not shown, for the interview lengths of outcomes 203, 213, 215, 219, 223, 225, 236, and 246 indicate that it is reasonable to assume that they are independent samples.

We also performed the Q-Q plots and goodness-of-fit tests for each of those outcome data. Table 6 shows the results of the goodness-of-fit tests of all the outcome data analyzed.

The fitted distributions in Table 6 will be used for the random variate generation described in Section 5.2.

### 4.3 Analysis of Contact Histories with NHIS

In this section, we describe the analysis of the contact attempt history data collected with the 2004 NHIS. Dahlhamer, Simile, Stussman, and Taylor [6] give a detailed analysis of this CHI (Contact History Instrument) data set. They found weekday evenings and weekends to be the best times to make contact with households in the NHIS, at least for the first four attempts (where prior attempts were no-contacts). In their work, for all analyses involving time of contact attempt, mornings are defined as 12:00 AM to 11:59 AM, afternoons as 12:00 PM to 4:49 PM, and evenings as 5:00 PM to 11:59 PM. The input modeling of the simulation is based on their work. However, we are only interested in the contact/no-contact distributions based on the time by the hour of a day and the days of a week. We divide a day by the hour because the starting time of each day by the FRs will be a decision variable when a sensitivity analysis is performed.

There are 197607 observations from the 2004 CHI data sets. More than 78% of them (154741 observations) are personal visits, the other 22% (42866 observations) are telephone calls. In our simulation model, we assume

Table 5: Results of the Four Goodness of Fit Tests for Outcome 201 Data

Distribution Test	Statistic	DF	p Value		Test Result
Beta( $\alpha, \beta, \theta, \lambda$ )					
Chi-Square	$\chi^2 = 645.802862$	47	$\Pr(p > \chi^2)$	< 0.001	rejected $H_0$
Kolmogorov-Smirnov	$D = 0.024418$		$\Pr(p > D)$	< 0.001	rejected $H_0$
Cramér-von Mises	$W^2 = 5.205604$		$\Pr(p > W^2)$	< 0.001	rejected $H_0$
Anderson-Darling	$A^2 = 31.448265$		$\Pr(p > A^2)$	< 0.001	rejected $H_0$
Weibull( $\alpha, \theta, \lambda$ )					
Chi-Square	$\chi^2 = 687.104887$	47	$\Pr(p > \chi^2)$	< 0.001	rejected $H_0$
Kolmogorov-Smirnov	$D = 0.019136$		$\Pr(p > D)$	< 0.001	rejected $H_0$
Cramér-von Mises	$W^2 = 2.413048$		$\Pr(p > W^2)$	< 0.001	rejected $H_0$
Anderson-Darling	$A^2 = 21.969588$		$\Pr(p > A^2)$	< 0.001	rejected $H_0$
Gamma( $\alpha, \theta, \lambda$ )					
Chi-Square	$\chi^2 = 834.565941$	47	$\Pr(p > \chi^2)$	< 0.001	rejected $H_0$
Kolmogorov-Smirnov	$D = 0.022515$		$\Pr(p > D)$	< 0.001	rejected $H_0$
Cramér-von Mises	$W^2 = 4.223140$		$\Pr(p > W^2)$	< 0.001	rejected $H_0$
Anderson-Darling	$A^2 = 36.594500$		$\Pr(p > A^2)$	< 0.001	rejected $H_0$

Table 6: Results of the Goodness of Fit Tests for Outcome Data Tested

Outcome	Sample Size ( $n$ )	Distribution(s) Fitted (Not Rejected)	Distribution Will Be Used in Simulation Runs	Alternative Distribution(s) Will Be Used
201	26741	None	Empirical	Beta( $\alpha, \beta, \theta, \lambda$ ) =(2.127, 2.549, 8.796, 79.258)
203	5055	None	Empirical	Beta( $\alpha, \beta, \theta, \lambda$ ) =(1.988, 3.233, 0.969, 81.207)
213	78	Weibull Exponential Gamma	Weibull( $\alpha, \theta, \lambda$ ) =(1.000, 0.367, 2.064)	Empirical Exponential( $\theta, \lambda$ ) =(0.340, 2.090) Gamma( $\alpha, \theta, \lambda$ ) =(0.913, 0.367, 2.261)
215	460	Gamma Exponential Weibull	Gamma( $\alpha, \theta, \lambda$ ) =(1.052, 1.298, 15.257)	Exponential( $\theta, \lambda$ ) =(1.265, 16.076) Weibull( $\alpha, \theta, \lambda$ ) =(1.016, 1.299, 16.148)
219	496	Gamma Weibull	Gamma( $\alpha, \theta, \lambda$ ) =(1.017, 0.233, 2.965)	Weibull( $\alpha, \theta, \lambda$ ) =(1.000, 0.233, 3.014)
223	101	Gamma Weibull Beta Lognormal	Gamma( $\alpha, \theta, \lambda$ ) =(1.708, 0.462, 2.195)	Weibull( $\alpha, \theta, \lambda$ ) =(1.328, 0.531, 3.999) Beta( $\alpha, \beta, \theta, \lambda$ ) =(1.339, 6.897, 0.533, 22.667) Lognormal( $\zeta, \theta, \lambda$ ) =(1.319, -0.288, 0.623)
225	866	Lognormal	Lognormal( $\zeta, \theta, \lambda$ ) =(0.802, 0.194, 0.982)	Weibull( $\alpha, \theta, \lambda$ ) =(1.031, 0.267, 3.512)
236	12470	None	Empirical	Lognormal( $\zeta, \theta, \lambda$ ) =(1.538, -0.086, 0.596)
246	3120	Lognormal	Lognormal( $\zeta, \theta, \lambda$ ) =(0.677, -0.054, 0.825)	Gamma( $\alpha, \theta, \lambda$ ) =(1.302, 0.166, 1.910)

that there are no telephone calls in the field operations activities. In reality, phone calls may be used for follow-up interviews after the first contact personal visit. We will add the activities of phone calls into the model in the future.

Table 7: The Frequency Distributions of Contact/No-Contact

	Hour	Sun(%)	Mon(%)	Tue(%)	Wed(%)	Thur(%)	Fri(%)	Sat(%)
No-Contact	00-01	66.67	34.38	64.41	62.14	54.44	58.90	55.56
Contact		33.33	65.62	35.59	37.86	45.56	41.10	44.44
No-Contact	01-02	64.52	39.13	47.37	63.92	51.65	50.70	52.17
Contact		35.48	60.87	52.63	36.08	48.35	49.30	47.83
No-Contact	02-03	66.67	50.00	68.18	59.34	71.74	61.82	77.78
Contact		33.33	50.00	31.82	40.66	28.26	38.18	22.22
No-Contact	03-04	50.00	57.89	69.23	48.15	34.88	58.62	42.86
Contact		50.00	42.11	30.77	51.85	65.12	41.38	57.14
No-Contact	04-05	44.44	33.33	70.27	71.88	37.50	63.16	57.14
Contact		55.56	66.67	29.73	28.12	62.50	36.84	42.86
No-Contact	05-06	88.89	50.00	58.82	50.00	37.50	29.41	0.00
Contact		11.11	50.00	41.18	50.00	62.50	70.59	100.00
No-Contact	06-07	58.33	71.43	60.87	65.99	49.28	57.41	53.85
Contact		41.67	28.57	39.13	34.01	50.72	42.59	46.15
No-Contact	07-08	29.41	53.57	51.38	65.53	54.17	53.15	61.54
Contact		70.59	46.43	48.62	34.47	45.83	46.85	38.46
No-Contact	08-09	66.04	38.46	49.19	44.99	45.27	42.15	56.69
Contact		33.96	61.54	50.81	55.01	54.73	57.85	43.31
No-Contact	09-10	48.04	54.49	44.97	46.30	51.34	46.68	41.18
Contact		51.96	45.51	55.03	53.70	48.66	53.32	58.82
No-Contact	10-11	51.12	47.15	48.52	47.16	46.08	47.14	46.42
Contact		48.88	52.85	51.48	52.84	53.92	52.86	53.58
No-Contact	11-12	50.89	49.23	51.45	50.85	46.61	50.16	43.75
Contact		49.11	50.77	48.55	49.15	53.39	49.84	56.25
No-Contact	12-13	48.04	49.48	49.15	52.27	47.70	50.81	45.44
Contact		51.96	50.52	50.85	47.73	52.30	49.19	54.56
No-Contact	13-14	47.10	49.37	47.47	48.51	51.00	47.45	46.39
Contact		52.90	50.63	52.53	51.49	49.00	52.55	53.61
No-Contact	14-15	46.39	49.29	49.98	46.23	50.06	46.27	45.40
Contact		53.61	50.71	50.02	53.77	49.94	53.73	54.60
No-Contact	15-16	49.53	46.95	46.83	46.98	46.88	45.10	45.13
Contact		50.47	53.05	53.17	53.02	53.12	54.90	54.87
No-Contact	16-17	47.22	43.46	44.52	42.90	42.89	45.44	47.06
Contact		52.78	56.54	55.48	57.10	57.11	54.56	52.94
No-Contact	17-18	45.60	42.70	42.88	39.82	41.91	45.24	49.57
Contact		54.40	57.30	57.12	60.18	58.09	54.76	50.43
No-Contact	18-19	47.90	42.59	40.20	41.91	42.92	46.42	51.58
Contact		52.10	57.41	59.80	58.09	57.08	53.58	48.42
No-Contact	19-20	50.07	47.26	46.13	44.50	44.79	48.79	50.94
Contact		49.93	52.74	53.87	55.50	55.21	51.21	49.06
No-Contact	20-21	47.96	49.06	49.24	48.34	49.77	52.41	59.21
Contact		52.04	50.94	50.76	51.66	50.23	47.59	40.79
No-Contact	21-22	52.97	50.73	49.40	52.77	51.70	58.70	54.18
Contact		47.03	49.27	50.60	47.23	48.30	41.30	45.82
No-Contact	22-23	46.67	56.52	55.74	55.74	56.03	60.00	61.06
Contact		53.33	43.48	44.26	44.26	43.97	40.00	38.94
No-Contact	23-24	53.63	57.84	58.36	52.09	54.44	65.71	58.76
Contact		46.37	42.16	41.64	47.91	45.56	34.29	41.24

Table 7 shows the personal visit frequency distributions of contact/no-contact based on the time of a day and the day of a week. In the table, the column of “Hour” shows the local time of the PSUs that the FRs visit the households. It shows that the best times to make contact with households in the NHIS are between 1:00 PM to 8:00 PM on weekdays, 9:00 AM to 5:00 PM on Saturdays, and 12:00 PM to 7:00 PM on Sundays. The table also indicates that the personal visits occurred at any time of a day.

## 5 Random Number and Random Variate Generations

A random number is a single observation of the continuous uniform distribution on the interval  $(0, 1)$ . The random number is then transformed as needed to simulate a random variate from different probability distributions, such as the normal, exponential, Poisson, binomial, Weibull, gamma, lognormal, etc. Random number generation is a computational procedure designed to generate a sequence of numbers. In contrast, random variate generation always refers to the generation of variates whose probability distribution is usually different from that of the uniform on the interval  $(0, 1)$ .

### 5.1 Random Number Generation

Random numbers are the basic building blocks of simulation study. A random number generator is needed to generate a sequence of independent and identically distributed (iid)  $U(0, 1)$  random variables. This sequence of random numbers can be obtained thru deterministic algorithms with a solid mathematical basis. The numbers produced by these algorithms are in fact not random at all. They should be called pseudorandom. For more detailed description of pseudorandom number generations, see L'Ecuyer [13]. For simplification, the term random is used instead of pseudorandom in the simulation contexts. A *random number* is always meant a uniform random variable, denoted by  $U(0, 1)$  (or `rand()` in our C++ code of the simulation model), whose distribution function is

$$F(u) = \begin{cases} 0, & \text{if } u \leq 0; \\ u, & \text{if } 0 < u < 1; \\ 1, & \text{if } u \geq 1. \end{cases} \quad (7)$$

The algorithm we used to generate a sequence of random numbers is given in [4].

### 5.2 Random Variate Generation

In Section 5.1 the generation of (pseudo) *random numbers* was briefly discussed. In this section, we will briefly discuss the random variate generations, see Cheng [5] for more detailed descriptions.

Random variate generation refers to the generation of variates whose probability distribution is different from that of the uniform on the interval  $(0,1)$ . The basic concept is to generate a random variable,  $X$ , whose distribution function

$$F(x) = \Pr(X \leq x) \quad -\infty < x < \infty \quad (8)$$

is assumed to be completely known, and which is different from that of Equation (7). A list of the random variate generations used in our C++ code of the simulation model is given in [4].

## 6 Description of the Model

First, one thousand and fifty cases (households) are generated for the model. Ten field representatives (FRs) are assumed, each of them is assigned a hundred and five cases and a PSU of  $60 \times 60$  square miles <sup>1</sup>. Each of the one thousand and fifty cases is identified with its case number and its location  $(x, y)$  within its own PSU, where  $0 \leq x \leq 60$  and  $0 \leq y \leq 60$ . The values of  $x$  and  $y$  come from a *uniform input distribution between 0 and 60*,  $U(0, 60)$  <sup>2</sup>. The field office and/or the FRs' homes can be located any where in a PSU of 3600 square miles. The simulation results are independent of the locations because the sample households are randomly selected. In this simulation study, they are assumed to be located at  $(0, 0)$ .

---

<sup>1</sup>The area of a PSU should not exceed 3,000 square miles except in cases where a single county exceeds the maximum area; we use 3,600 square miles for our experimental runs

<sup>2</sup>All the uniform distributions described in this paper are discrete uniform distributions

Each FR selects the first  $n$  cases (ascending order of case numbers of the incomplete cases) for each day's work. The value of  $n$  comes from a *uniform input distribution*  $U(8, 16)$ . The FR has to visit each of the  $n$  selected cases once for that day. The visiting order of the  $n$  cases is determined by the following:

1. The direct distance ( $d_{ij}$ ) between each pair ( $(x_i, y_i)$  and  $(x_j, y_j)$ ) of the  $n$  cases is calculated by

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

2. A distance matrix of the  $n$  cases is formed, and therefore, a traveling network is formed with interconnections between the nodes (cases).
3. Starting from the field office or FR's home located at  $(0, 0)$  and ending at  $(0, 0)$ , the FR visits each case only once. This becomes a traveling salesman problem<sup>3</sup>.
4. The objective is to minimize the total distance traveled. Instead of this measure of distance, any other measure of effectiveness may be substituted, such as time, likelihood of contact, and so on.
5. A branch and bound algorithm<sup>4</sup> is used to determine the visiting order.

With the shortest path to visit each of the  $n$  households determined, each of the FRs is to visit each household to conduct an interview for the survey. Table 8 shows the detailed information for conducting interviews of 10 cases by FR 7 at day 5 for the simulation run with seed 23. The distances (determined by the two locations  $(x_i, y_i)$  and  $(x_j, y_j)$ ) of traveling to each of the households are given in column (4) of Table 8. The average speed (mph) for the distances is a *uniform distribution*  $U(30, 40)$ . The contact time (minutes) at each household is a *uniform distribution*  $U(3, 7)$ . At each household, it is either contacted or not contacted. The contact/no-contact distributions are described in Table 7; no-contact if 0 and contact otherwise (potential refusal is considered contact) so that the probability of contact depends on the time of a day and the day of a week. If it is contacted, the interview length (minutes) is generated from the distributions given in Table 9 (also see Table 6) depending on the outcomes of the visits.

In Table 8, columns (1) and (2) list the  $n$  cases that need to be visited by the FR for the field operations. The dashes after Today's Seq. 10 and Case Number 738 are given to indicate the mileage and time needed for the FR to drive back to the office at  $(0, 0)$ . The first row shows that the FR drives 13 miles at average speed of 32 mph to the first household (case number 750) arriving at minute 24, computed from columns (3) and (4). Column (6) shows the simulated arrival time of the FR at each household. Column (7) shows the time needed to make contact with the respondent in the household. Column (8) shows the clock time that the interview began. Column (9) indicates that the number of visits to complete the interview so far. Column (10), contact or no-contact, shows the binary values of contact = 1 and no-contact = 0. The values of column (11) are time needed for the interview if there was a contact. Otherwise, there was no interview and the time needed was 0. Finally, column (12) shows the clock time that the interview ended.

## 7 Preliminary Output Analysis Results

FRs are given 17 days, starting with the Monday of the assignment week for each month, to complete each assignment. Therefore, the simulation model starts in a state of no personal visits for all cases assigned

---

<sup>3</sup>The traveling salesman problem can be stated as follows. A salesman, starting from a city, intends to visit each of  $(n-1)$  other cities once and only once and return to the start. The problem is to determine the order in which he should visit the cities to minimize the total distance traveled, assuming that the direct distances between all city pairs are known. The structure of the problem shows that there are  $(n-1)!$  possible tours, of which one or more should be optimal

<sup>4</sup>The method of the branch and bound algorithm is to first identify a feasible solution and then to decompose the set of all remaining feasible tours into smaller and smaller subsets. At each step of the decomposition, a lower bound on the length of the current best tour is readily available. The bounds provide a guide for the partitioning of the subsets of feasible tours and eventually for the identification of an optimal tour. When a tour with length less than or equal to the minimum lower bound of all other tours is found, this intermediate solution becomes the best available. This process of bounding tours, eliminating suboptimal alternatives, and branching to new (better) tours is the basis of the algorithm.

Table 8: Field Representative 7 at Day 5 with seed 23

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Today's Seq.	Case Number	Average Speed	Distance	Time for Traveling	Arrival Time	Time to Contact	Time Interview Begins	Current Visit	Contact	Time for Interview	Time Interview Ends
1	750	32	13	24	24	5	29	1	1	22	51
2	746	36	38	63	114	6	120	1	1	21	141
3	747	30	51	102	243	5	248	1	1	23	271
4	748	35	12	20	291	3	294	1	0	0	294
5	752	37	27	43	337	3	340	1	0	0	340
6	749	39	3	4	344	5	349	1	1	23	372
7	751	38	27	42	414	4	418	1	1	21	439
8	745	38	18	28	467	5	472	1	1	23	495
9	737	39	57	87	582	6	588	2	1	25	613
10	738	37	11	17	630	3	633	2	0	0	633
-	-	31	81	156	789	-	-	-	-	-	-
Total	-	-	338	586	-	45	-	-	7	158	789

Table 9: Distributions Used in the Model for Interview Lengths

Outcome	Probability Distribution Used in the Model
201	$\text{Beta}(\alpha, \beta, \theta, \lambda) = \text{Beta}(2.127, 2.549, 8.796, 79.258)$
203	$\text{Beta}(\alpha, \beta, \theta, \lambda) = \text{Beta}(1.988, 3.233, 0.969, 81.207)$
213	$\text{Weibull}(\alpha, \theta, \lambda) = \text{Weibull}(1.000, 0.367, 2.064)$
215	$\text{Gamma}(\alpha, \theta, \lambda) = \text{Gamma}(1.052, 1.298, 15.257)$
219	$\text{Gamma}(\alpha, \theta, \lambda) = \text{Gamma}(1.017, 0.233, 2.965)$
223	$\text{Gamma}(\alpha, \theta, \lambda) = \text{Gamma}(1.708, 0.462, 2.195)$
225	$\text{Lognormal}(\zeta, \theta, \lambda) = \text{Lognormal}(0.802, 0.194, 0.982)$
236	$\text{Lognormal}(\zeta, \theta, \lambda) = \text{Lognormal}(1.538, -0.086, 0.596)$
246	$\text{Lognormal}(\zeta, \theta, \lambda) = \text{Lognormal}(0.677, -0.054, 0.825)$

each month. We will assume the simulation of field operations is the type of finite-horizon simulations. The estimations of the performance measures via independent replications will be used for the output analysis.

For the one thousand observations, a point estimate and the 95% confidence interval estimation of the mean and variance of the performance measures, such as response rate, average number of visits per case, and cost are analyzed using the sample mean, the sample variance, and the confidence interval with  $k = 1000$  and  $\alpha = 0.95$ . Table 10 shows the estimates and their 95% confidence intervals of the aforementioned performance measures.

Table 10: The Estimation of the Cost, Response Rate, and Number of Personal Visits

Performance Measure	Mean			Variance		
	Estimate	95% Conf.	Limits	Estimate	95% Conf.	Limits
Cost (\$)	25,475	25,454	25,495	111870	102673	122367
Response Rate (%)	86.04	85.94	86.15	3.02	2.77	3.30
Average # of Personal Visits	1.74188	1.74020	1.74356	0.0007356	0.0006751	0.0008046

Next, we show an example of how the performance measures would change when we change some of the parameters. Those parameters are controllable:

1. The starting time of each day by the field representatives: we assume that all the FRs start at 10:00 AM, 12:00 noon, or 3:00 PM. These parameter settings are based on the contact/no-contact distributions

given in Table 7. We repeat part of the distributions in Table 11. Note that we consider *Potential Refusal* as *Contact*. Table 11 shows that the overall contact probability (55.32%) is higher during the hours of 3:00 to 8:00 PM than during the hours of 10:00 AM to 3:00 PM.

Table 11: The Selected Frequency Distributions of Contact/No-Contact

	Hours	Sun(%)	Mon(%)	Tue(%)	Wed(%)	Thur(%)	Fri(%)	Sat(%)	Overall(%)
No-Contact	10:00-12:00	50.98	48.46	50.25	49.33	46.38	48.91	44.83	48.12
Contact		49.02	51.54	49.75	50.67	53.62	51.09	55.17	51.88
No-Contact	12:00-15:00	47.03	49.37	48.90	48.78	49.69	48.04	45.75	48.36
Contact		52.97	50.63	51.10	51.22	50.31	51.96	54.25	51.64
No-Contact	15:00-20:00	48.05	44.50	43.95	43.06	43.74	46.14	48.23	44.68
Contact		51.95	55.50	56.05	56.94	56.26	53.86	51.77	55.32

2. The number of field representatives: we increase the number of FRs from 10 to 15 and keep the same number of cases assigned at 1050. The covered geographical area is changed from 3,600 square miles to 2,401 square miles. The number of cases assigned to each FR is also changed from 105 to 70. The number of working days is reduced from 17 to 11.

Table 12 shows the nine parameter settings discussed above. For each parameter setting, we generate 1000

Table 12: The Nine Parameter Settings for the Experiments

Setting	Starting Time	# of FRs	Days	Area	FR-Days	Adjusted Days
1	10:00	10	17	3600	170	17.00
2	12:00	10	17	3600	170	17.00
3	15:00	10	17	3600	170	17.00
4	10:00	15	11	2401	165	11.33
5	12:00	15	11	2401	165	11.33
6	15:00	15	11	2401	165	11.33
7	10:00	15	17	2401	255	11.33
8	12:00	15	17	2401	255	11.33
9	15:00	15	17	2401	255	11.33

observations, the estimates of the performance measures are given in Table 13. Table 13 also shows the

Table 13: The Estimates of the Performance Measures of the Nine Parameter Settings

Setting	Cost	Response Rate(RR)	Average Visits	Adjusted to 170 FR-Days			Cost Savings
				Cost	Response Rate(RR)	Average Visits	
1	\$25,375	86.19%	1.72	\$25,375	86.19%	1.72	--
2	\$25,238	86.86%	1.71	\$25,238	86.86%	1.71	--
3	\$25,475	86.04%	1.74	\$25,475	86.04%	1.74	--
4	\$20,722	82.23%	1.68	\$21,349	84.72%	1.73	15.86%
5	\$20,575	83.50%	1.66	\$21,199	86.03%	1.71	16.00%
6	\$20,589	83.88%	1.67	\$21,213	86.42%	1.72	16.73%
7	\$24,545	89.93%	1.78	RR gain	3.74%	--	3.27%
8	\$24,085	89.96%	1.75	RR gain	3.10%	--	4.57%
9	\$23,926	89.98%	1.75	RR gain	3.94%	--	6.08%

adjustments of the performance measures to 170 FR-Days for the parameter settings of 4, 5, and 6. By visual inspection, there is no significant difference among the parameter settings of starting time for each day for

all three performance measures. However, there is cost saving if more field representatives are assigned to the 1050 cases as indicated in Table 13 that the settings of 4, 5, and 6 have potential cost savings of 15.86%, 16.00%, and 16.73% over the settings of 1, 2, and 3, respectively. We also examine where the cost saving is coming from. Table 14 shows the cost estimates with seed 169001 for the parameter settings 3 and 6, where five more FRs are assigned to the 1050 cases. The last row labeled **Adjusted** is the adjustments to 170 FR-Days for parameter setting 6. The total traveling distance is 35,012 miles for parameter setting 3 and 27,922 miles for parameter setting 6. It is a saving of 20.25%. Therefore, a smaller PSU area would reduce the traveling time for the FRs, meaning *less time on the roads and more time knocking on the doors*.

Parameter settings 7, 8, and 9 are used to examine the effect of the response rate if we would like the FRs to work 17 days instead of 11 days. The results indicate that these three parameter settings have cost savings of 3.27%, 4.57%, and 6.08% over parameter settings 1 to 3, respectively. The response rates also have increases of 3.74%, 3.10%, and 3.94%, respectively. These are strong evidences that reducing the cost while increasing the response rate is feasible for the field operations if the parameters are properly set.

Table 14: The Cost Estimates of the Replication with Seed 169001

FR	Total time (hours)	Wages (\$)	Total distance (miles)	Mileage (\$)	Total cost (\$)
<b>Parameter Setting 3</b>					
0	129.35	1293.50	3304	1156.40	2449.90
1	135.40	1354.00	3569	1249.15	2603.15
2	138.95	1389.50	3691	1291.85	2681.35
3	136.83	1368.33	3710	1298.50	2666.83
4	127.33	1273.33	3229	1130.15	2403.48
5	130.67	1306.67	3255	1139.25	2445.92
6	134.82	1348.17	3412	1194.20	2542.37
7	133.17	1331.67	3503	1226.05	2557.72
8	132.15	1321.50	3448	1206.80	2528.30
9	150.70	1507.00	3891	1361.85	2868.85
<b>Total</b>	<b>1349.37</b>	<b>13493.67</b>	<b>35012</b>	<b>12254.20</b>	<b>25747.87</b>
<b>Parameter Setting 6</b>					
0	71.83	718.33	1747	611.45	1329.78
1	73.42	734.17	1803	631.05	1365.22
2	71.75	717.50	1772	620.20	1337.70
3	73.85	738.50	1885	659.75	1398.25
4	71.38	713.83	1701	595.35	1309.18
5	73.05	730.50	1835	642.25	1372.75
6	77.12	771.17	1865	652.75	1423.92
7	73.37	733.67	1801	630.35	1364.02
8	79.25	792.50	1918	671.30	1463.80
9	69.75	697.50	1673	585.55	1283.05
10	79.18	791.83	1926	674.10	1465.93
11	75.90	759.00	1861	651.35	1410.35
12	69.17	691.67	1713	599.55	1291.22
13	70.08	700.83	1706	597.10	1297.93
14	78.42	784.17	1895	663.25	1447.42
<b>Total</b>	<b>1107.52</b>	<b>11075.17</b>	<b>27101</b>	<b>9485.35</b>	<b>20560.52</b>
<b>Adjusted</b>	<b>1141.08</b>	<b>11410.78</b>	<b>27922</b>	<b>9772.78</b>	<b>21183.57</b>

## 8 Conclusion and Summary

In conclusion, we have shown that the simulation model can be used for optimizing the field operations by setting the controllable parameters before a decision is made and implemented. The cost savings might be enormous as shown in the example (about 16%) of Section 7 and would not be at the expense of the response

rate. If more working days are needed by FRs, a cost saving with higher response rate is also feasible.

Figure 5 shows how the optimization of field operations cost can be achieved. In the figure, the solid line shows the direct cost of FRs vs. the number of FRs<sup>5</sup>. The preliminary result indicates that the direct cost is a decreasing function of the number of FRs. If the hiring and training cost (or the overhead) of FRs is an increasing function of the number of FRs, shown in Figure 5 as a dot line, then the minimum total cost can be located by examining the dash line, which is the sum of the solid and dot lines, of the figure.

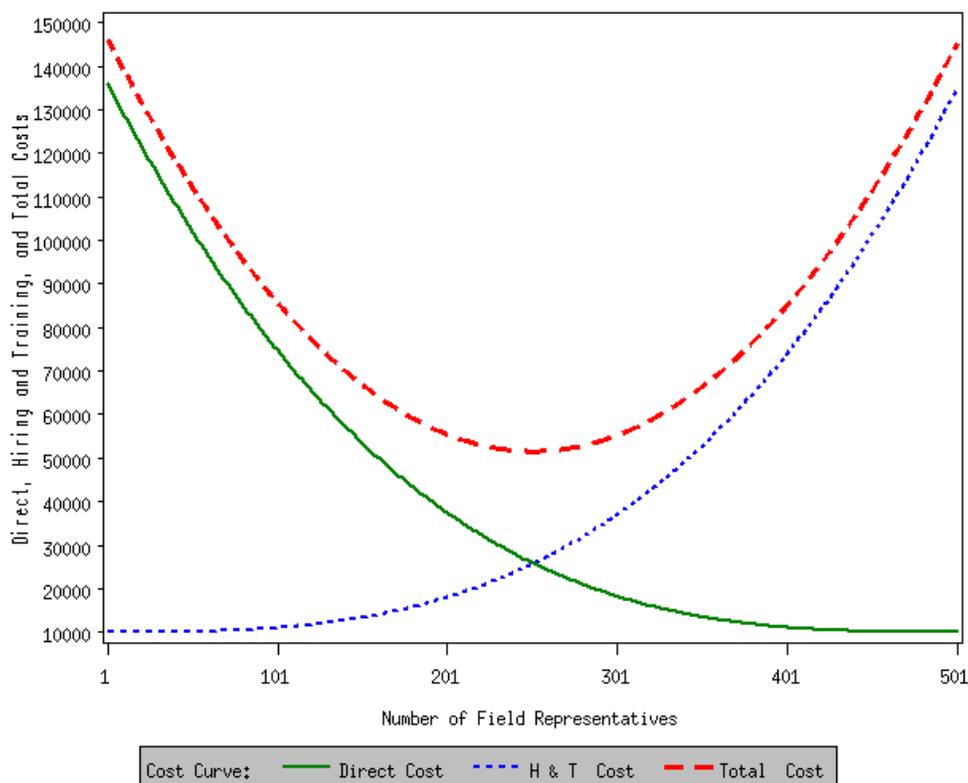


Figure 5: Optimization of Field Operations Cost.

In summary, the following preliminary tasks have been used for this study:

1. Model conceptualization: The model will begin simply and grow until a model of appropriate complexity has been developed.
2. Data collection: A data set for each variable from the NHIS is collected with help from Demographic Surveys Division.
3. Input data analysis: Determine the distribution function of the data set collected for each variable.
4. Model translation: The conceptual model constructed in Step 1 is coded into a computer-recognizable form, an operational model.
5. Verification and validation: Verification concerns the operational model. Is it performing properly? Validation is the determination that the conceptual model is an accurate representation of the field operations. The process of verification and validation is an iterative one. New details will be added to the model and new results are presented to Field Division (or field operations experts). If the results are

<sup>5</sup>We use 501 as the total number of FRs in the figure for illustration purpose only, it is not the actual number of FRs for the NHIS.

not sufficiently accurate, Field Division/experts identify other details that should be included. These details are added, and the cycle starts anew. At some point, we “must” agree that the model is “close enough” to provide useful information. The agreement can be based on the simulation output data and the Field Division historical data.

6. Experimental design: For each scenario that is to be simulated, decisions need to be made concerning the length of the simulation run, the number of runs (also called *replications*), and the manner of initialization, as required.
7. Production runs and output analysis: Production runs and their subsequent output analysis are used to estimate the performance measures for the scenario that are being simulated.
8. Sensitivity and feasibility study.
9. Documentation and reporting.

The simulation model will be modified according to the aforementioned Step 5 to make the model valid for a better tool of decision making.

As mentioned in Section 6, the simulation model described in this paper is for the simplified field operations of surveys. For example, we explored the possible simulation models at the national level. However, other geographic attributes such as region and MSA (Metropolitan Statistical Area) status should be explored in the future. These attributes have been shown to be a measure associated with contact [7]. Other future work that should be included is listed as following:

1. physical impediments, at-home patterns of households as described in Section 4.3;
2. interviewer strategies that influence contact such as advance letters and notices of visit [8] and telephone interviews after the first contact;
3. multiple visits of completed interviews (outcome 201), a completed interview may need several visits of the same household in which the interview lengths of the visits may be correlated;
4. a sample household may have several unrelated persons living in the same house, it is required by NHIS to interview each one of them;
5. classification of interviewers (field representatives or supervisory field representatives) based on their experience and training;
6. GPS with live traffic used by field representatives to cut traveling time between sample households.

## References

- [1] Alexopoulos, Christos and Seila, Andrew F. Output Data Analysis. In Jerry Banks, editor, *Handbook of Simulation*, pages 225–272. John Wiley & Sons, Inc., New York, 1998.
- [2] N. Bates. Contact Histories in Personal Visit Surveys: The Survey of Income and Program Participation (SIPP) Methods Panel. Demographic Surveys Division, U.S. Bureau of the Census, Washington, DC 20233, May 7, 2003.
- [3] R. L. Bitzer. Personal Communications, 2003.
- [4] Chen, Bor-Chung. Stochastic Simulation of Field Operations in Surveys. Research Report Computing #2008-1, Statistical Research Division, Bureau of the Census, Washington, D.C., 2008.
- [5] Cheng, Russell C. H. Random Variate Generation. In Jerry Banks, editor, *Handbook of Simulation*, pages 139–172. John Wiley & Sons, Inc., New York, 1998.

- [6] James M. Dahllamer, Catherine M. Simile, Barbara J. Stussman, and Beth Taylor. Determinants and Outcomes of Initial Contact in the National Health Interview Survey, 2004. National Center for Health Statistics, Hyattsville, MD, May 14, 2005.
- [7] James M. Dahllamer, Barbara J. Stussman, Catherine M. Simile, and Beth Taylor. Modeling Survey Contact in the National Health Interview Survey (NHIS). National Center for Health Statistics, Hyattsville, MD, May 14, 2005.
- [8] Robert M. Groves, Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey Methodology*. John Wiley & Sons, Inc., New York, 2004.
- [9] Alfred Hartmann and Herb Schwetman. Discrete-Event Simulation of Computer and Communication Systems. In Jerry Banks, editor, *Handbook of Simulation*, pages 659–676. John Wiley & Sons, Inc., New York, 1998.
- [10] J. A. Joines and S. D. Roberts. Design of Object-Oriented Simulation in C++. In *Proceedings of the 1995 Winter Simulation Conference*, 1995.
- [11] Ron Laughery, Beth Plott, and Shelly Scott-Nash. Simulation of Service Systems. In Jerry Banks, editor, *Handbook of Simulation*, pages 629–644. John Wiley & Sons, Inc., New York, 1998.
- [12] A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, New York, 2nd edition, 1991.
- [13] Pierre L’Ecuyer. Random Number Generation. In Jerry Banks, editor, *Handbook of Simulation*, pages 93–137. John Wiley & Sons, Inc., New York, 1998.
- [14] Mani S. Manivannan. Simulation of Logistics and Transportation Systems. In Jerry Banks, editor, *Handbook of Simulation*, pages 571–604. John Wiley & Sons, Inc., New York, 1998.
- [15] Frank McGuire. Simulation in Healthcare. In Jerry Banks, editor, *Handbook of Simulation*, pages 605–627. John Wiley & Sons, Inc., New York, 1998.
- [16] H. G. Meyers. Estimating Survey Cost Drivers with FLD Administrative Records and POP Demographic Data. Applied to the Current Population Survey, U.S. Bureau of the Census, Field Division Memo, January 31, 2003.
- [17] Matthew W. Rohrer. Simulation of Manufacturing and Material Handling Systems. In Jerry Banks, editor, *Handbook of Simulation*, pages 519–545. John Wiley & Sons, Inc., New York, 1998.
- [18] I. Shimizu and F. Lan. Approximation of Variable Costs for the National Health Interview Survey. *Proceedings of the Annual Meeting of the American Statistical Association*, August 5-9, 2001.
- [19] Onur Ulgen and Ali Gunal. Simulation in the Automobile Industry. In Jerry Banks, editor, *Handbook of Simulation*, pages 547–570. John Wiley & Sons, Inc., New York, 1998.
- [20] Stephen Vincent. Input Data Analysis. In Jerry Banks, editor, *Handbook of Simulation*, pages 55–91. John Wiley & Sons, Inc., New York, 1998.