Estimating Toll-Free Telephone Demand: Customer Retrial Behavior and Blocking Rate

Robin H. Lee, Internal Revenue Service

The Internal Revenue Service (IRS) toll-free telephone system has provided American taxpayers assistance in tax law since 1965. In fiscal year 1992, the IRS received over 120 million call attempts and answered roughly 35 million calls. During the filing season alone (January - April), the IRS answered over 18 million net calls (to be explained later) out of 68 million call attempts. In fiscal year 1993 (October 1, 1992 - September 30, 1993), 23.5 million calls were answered by the end of April with almost 78 percent of them (18.3 million) completed during the filing season.

The statistics show that the IRS answered only 30 percent of call attempts in 1992. However, this doesn't mean only 30 percent of taxpayers were served because some of the unanswered call attempts represent the same individuals making multiple tries to get through. Thus, equating the total number of call attempts to the number of unique taxpayers would mistakenly count many repeat callers several times.

The IRS is interested in estimating true demand, the actual number of taxpayers trying to access the toll-free telephone system in any given time period. The reliability of the estimation methodology is very important because the data are used to determine the staffing and circuitry requirements for budget purposes and are cited in the General Accounting Office's (GAO) Congressional report every year.

The purpose of this paper is to overview the current IRS demand estimation methodology and its assumptions. The first section will describe the IRS phone system and the measurement data available from the system. The second section will review two basic approaches that IRS has taken in the past to estimate true demand. The third section will discuss the current model in greater detail to examine the validity of its assumptions. The implications for the data sets used at Bell Labs will also be explored. Finally, the IRS new method (planned for implementation in October of 1993) and future research plans will be presented in the last section.

■ IRS System Overview

When a taxpayer calls the IRS 800-829-1040 number, the local telephone company, recognizing the call belongs to Sprint (the IRS toll-free number carrier), sends it to the closest Sprint switching center. Sprint, then routes the call to one of 32 answering sites automatically, based upon a predetermined routing tree that matches area codes of originating calls to answering sites.

Each answering site is furnished with hardware, known as Automatic Call Distributor (ACD), that receives and distributes the calls to open telephone lines. If a taxpayer calls while all the trunks (bundles of phone lines) are occupied, the ACD rejects the call and the taxpayer receives a busy signal. These calls are referred to as *overflows* or *blocked calls*.

If there are open lines but no assistor is available--the ratio of the number of phone lines to assistors is about 1.3, then the caller hears a recorded message that all representatives are busy and is put on hold until the next assistor becomes free. In sites with Automated Response Units (ARUs), the touch tone callers can choose from the listed menu to directly route themselves to assistors specializing in the content area of their questions. If a caller hangs up before speaking to an assistor during this process, the call is considered *abandoned.*

The data available from the automated system are the total calls attempted (T), completed (C), blocked (B), and abandoned (AB) in a given period of time. The number of net calls answered by the IRS, mentioned earlier, is the number of completed calls (C) minus the abandons (AB).

■ Two Basic IRS Approaches

Over the years, the IRS has used several different models to estimate true demand. The current IRS model specifies true demand as the total number of

callers making their first attempt to reach IRS on a particular issue. The proportion of first time callers is estimated by asking a sample of taxpayers whether or not they have had busy signals before. The estimated proportion of first time callers is then multiplied by the total call volume in each hourly period to estimate the total number of taxpayers making initial attempts.

The earlier models used a different approach. When a taxpayer can't get through, he has two options: either keep trying until connected, or give up. Thus, true demand was defined as the sum of the number of completed calls and a portion of blocked callers who stop redialing, which is denoted by (1-r)B where r is an average redial probability (Harris, Hoffman, and Saunders, 1987). A challenge here was estimating r. Carl Harris of George Mason University applied queuing theory to estimate an overall retry probability; however, a single fixed value of r didn't work well when the system congestion level varied a lot. As a result, the basic formula was later modified to allow r to vary as a function of the blocking probability based upon the assumption that redial and blocking probabilities are correlated.

A major difference between the two approaches is that the Harris method counts the callers by their last attempt; hence, the alternative method counts the callers by their first attempt, hence it doesn't need the redial probability (Stone, 1989).

Both models make various assumptions regarding the dynamics of the incoming call processing system and the individual's redialing behaviors. However, it has been hard to test these assumptions empirically until now. A recent development in caller ID technology made it possible to count the unique phone numbers and the average number of attempts generated from each number, to calculate the actual retry probability, and test many of these assumptions.

■ The Treasury Model

The alternative model was developed in 1988 by Daniel J. Opitz (Opitz, 1988), working in the Treasury Office of Planning and Management Analysis and is referred to as the Treasury model from here on.

Underlying Logic

According to Opitz, the basic logic for the Treasury formula can be explained by an analogy. Suppose you want to count the fish population in a lake. One way to do it is using what is known as a capture/recapture method. First, you take a sample of fish, color them red, note the number of colored fish, C, and release them back into the lake. When you think they are totally mixed in with other fish in the lake, you take another random sample of fish and observe the proportion of colored fish (p). The total fish count is estimated by dividing C by p, assuming that the sample is large enough to catch at least one or more colored fish.

Likewise, we can think of the IRS calling population at any given time as the mixture of two different types of callers--first time callers and repeat callers. Since we know the total number of call attempts, if we can estimate the percentage of callers making their first attempt, then we can simply multiply these two together to estimate the total number of callers making their first attempt.

Survey

In the Treasury method, the percentage of first attempt callers is determined by directly asking a sample of taxpayers the following survey question.

"Before [you go/I transfer you], may I ask you a quick question to help us improve our service? When you called us about [this problem/question/topic], did you get a busy signal?"

Callers saying 'NO' are tallied as first time callers and the proportion of these callers is computed for each hourly period and multiplied by the corresponding hour's total call attempts. Daily or weekly demand is just the cumulative sum of these hourly estimates.

Assumptions

The model makes several assumptions:

(a) The probability of completing an arbitrary call is (within a small tolerance) independent of the time it is placed within the sampling period.

(b) The probability of completing an arbitrary call at any fixed point in time is independent of the number of previous calls.

(c) Redials are made within the same sampling period as the original call--or equivalently, the process is sufficiently stable that "carryover" call attempts by callers from one period to the next are relatively small and fairly equally balanced at both ends of the sampling window.

(d) The IRS sample is sufficiently large and reflects the variation in phone traffic throughout the sampling period.

Assumptions a, b, and c basically imply that the system is completely stationary, in that the accessibility and the rate of call traffic flowing in and out are constant within each hourly sampling window.

Derivation of the Formula

Using the first and second assumptions, the average probability of completing a call without being blocked is C/T and C_i/T_i is assumed to be equal to C/T for all i. The subscripts denote the level of attempt. For example, T_1 and C_1 are the number of total call attempts and calls completed on the first attempt. This implies that $T_1/T = C_1/C$. Assuming the IRS sample is random and adequate in size (assumption d), the sample-based estimates (c_i/c) will be approximately equal to the population values (C_i/C). Combining all the steps,

$$T_1/T \approx c_1/c, \text{ and}$$

$$T_1 \approx T\,(c_1/c).$$

According to this formula, true demand is defined as the total number of first attempts (T_1) and can be obtained by multiplying total call attempts by the percentage of calls completed on the first attempt (C_1/C), which is estimated by the survey as (c_1/c).

Further Discussions on Treasury Assumptions

One assumption implicit in this formula is that C_1/C represents the proportion of all the first time callers, not just for completed calls but for blocked calls, as well. Using our notation, this can be written as $C_1/C = T_1/T = B_1/B$.

This assumption is closely related to assumption (b) -- that the probability of a call being completed or blocked is the same for everyone, regardless of their previous attempt history. It seems like a reasonable assumption because incoming calls are distributed to the available assistors randomly by the telephone hardware, the ACD, which does not keep track of how many prior attempts there were.

However, this assumption is not supported by the Bell Lab data obtained from Automated Number Identification (ANI) records (Table 1). This table shows the distribution of total call attempts in a monthly window by attempt level and each attempt's outcome.

Table 1.--Call Blocking Rate by Attempt Level

Attempt	Total Attempt	Complete Attempt	Blocked Attempt	Blocking Prob.	Retry Prob.
1	112,687	92,842	19,845	0.18	0.80
2	15,845	8,267	7,578	0.48	0.84
3	6,351	2,968	3,383	0.53	0.86
4	2,908	1,228	1,680	0.58	0.90
5	1,520	604	916	0.60	0.91
6	829	329	500	0.60	0.91
7	457	146	311	0.68	0.93
8	289	101	188	0.65	0.94
9	176	53	123	0.70	0.94
10	115	25	90	0.78	0.93
Total	141,430	106,636	34,794		

These data indicate that the conditional blocking probabilities at each attempt level actually increase with the attempt level. This means, as the callers make more attempts, it becomes harder to get through. In addition, blocking probability and retrial probability are positively correlated, implying that the callers' redialing tendency increases with the blocking rate.

The table also shows that T_1/T is .8 and neither C_1/C (.87) nor B_1/B (.57) is equal to .8, as assumed under the Treasury model. The average blocking probability was 24 percent in the data. A couple of other Bell Lab data sets with higher average blocking rates showed even greater discrepancy between T_1/T and C_1/C.

Does this invalidate the Treasury model? Not really, because a monthly window was used in these data whereas an hourly window is used in the Treasury model. Would the independence assumption and the identity have held, if the data had an hour window? The answer is yes, **if and only if** the blocking probability didn't change at any point in time during an entire hour. In other words, a necessary condition is a constant blocking rate within each window, not the same window size, although it is true that the wider the window, the more fluctuation will be present in the system.

A Necessary Condition for the Independence Assumption

The independence assumption implies that the conditional completion probability given the i'th call attempt (C_i/T_i) is equal to the unconditional probability (C/T) for all i. This assumption seems perfectly valid if you reference only a single point in time. We can conceptualize this instantaneous distribution of incoming calls at a particular point in time, t_j, as a snapshot picture of the call distribution by attempt level and outcome during a brief moment where all callers have just enough time to make one call attempt. Within this instantaneous window, every caller has the same chance of completing a call, no matter how many prior attempts there have been. Thus, the proportion of first time calls among the blocked and the completed should be equal.

The distribution of call attempts made in an hour can be thought of as many of these instantaneous distributions at t_1, t_2, t_3,..., t_j stacked up. What the Treasury model assumes regarding this distribution is that the sum of C_{1j} over j, divided by the sum of C_j, is equal to the proportion of the first attempts computed from the hourly totals. Using the same notation with subscript j denoting the time period, this can be expressed as:

$$\frac{\sum C_{1j}}{\sum C_j} = \frac{\sum T_{1j}}{\sum T_j}$$

$$\text{if} \quad \frac{C_{1j}}{C_j} = \frac{T_{1j}}{T_j} \quad \text{for all } j .$$

Even if the condition $(C_{1j}/C_j)=(T_{1j}/T_j)$ is true for all j, the ratios based upon hourly totals in the numerator and denominator are not the same between the completed and total ***unless*** the blocking probabilities remain constant for all j. This can be shown in a simple numerical example.

An Example

For the sake of illustration, suppose we have a sampling window partitioned into 3 mini windows with the following statistics for each window:

at t_1, $C_1 = 20$, $T_1 = 40$,
$C = 30$, $T = 60$,

at t_2, $C_1 = 95$, $T_1 = 475$,
$C = 100$, $T = 500$ and

at t_3, $C_1 = 100$, $T_1 = 180$,
$C = 500$, $T = 900$.

$$\frac{\sum C_{1j}}{\sum C_j} = \frac{20+95+100}{30+100+500} = .34$$

$$\frac{\sum T_{1j}}{\sum T_j} = \frac{40+475+180}{60+100+900} = .48 .$$

As you can see, the ratios of the sums over j are not the same, although (C_{1j}/C_j) was equal to (T_{1j}/T_j) for all j.

The blocking rates are 50, 80, and 45 percent, respectively, using 1-(C/T). When the blocking rates are constant for all instantaneous windows j, C_j can be expressed as KT_j and C_{1j} as KT_{1j}. Then

$$\frac{\sum C_{1j}}{\sum C_j} = \frac{\sum K \times T_{1j}}{\sum K \times T_j}$$

and the identity holds.

So the question is, how reasonable is this assumption and how can we test it?

Testing the Treasury Assumption

One way to test the Treasury assumption is to measure the instantaneous blocking every 5 to 10 minutes or so. Since the assumption is required throughout an entire hour, the finer the intervals are, the more conclusive the results will be. However, as the intervals get shorter, the measurement of blocking rates will become less reliable. So an optimal interval length to balance between these two factors should be determined somehow.

Another way to test this assumption is to examine the ANI data in an hourly window to see if the actual conditional blocking rate is independent of the attempt level. This approach, however, needs some caution. The calls made at the beginning or end of each hourly window may not be accurately accounted for, unless a series of call attempts generated from the same number is made within the same sampling window. As assumed in the Treasury model, if redials are made within the same sampling period as the original call or "carryover" calls are relatively small and fairly equally balanced at both end of the sampling window, bias should be minimized. A more direct way to deal with these carryover calls might be to reorganize the weekly window data into hourly periods, based upon one of the ANI data fields, the time of the day.

■ Future Plans

Currently, the IRS is pursuing an alternative methodology that will measure true demand by counting the number of unique telephone numbers attempted on the IRS toll-free network. The IRS will not receive the actual ANI records but only the aggregated counts of unique phone numbers. Sprint reports will have a new measure of accessibility in a separate column--the cumulative percentage of callers getting answered on their initial, second, third call attempt, and so on. Table 2 shows this statistics from the same Bell Lab data. This statistic will tell us the percentage of "taxpayers" eventually getting through by the end of the same week they start calling. It is important to distinguish this caller-based service level indicator from a call-based service level measure--the percentage of "call attempts" served.

As pointed out earlier, the 30 percent call completion rate in 1992 doesn't necessarily mean that only 30 percent of taxpayers got through. If the average number of attempts per caller were 2, the 30 percent call completion rate translates to a caller completion rate of 60 percent. The sixty percent figure tells us what percent of true demand was served, whereas the call completion rate indicates how congested the system was in providing that level of service. Obviously these indices measure two different aspects of service level.

Table 2.--Cumulative Percent Completed Callers by Attempt Level

Attempt Level	Cum. % Completed Callers
1	82
2	89
3	92
4	93
5	93
6	94
7	94
//	
Total	94

One limitation in the Sprint reports is that they exclude non-800 local numbers, which account for about 30 to 40 percent of total IRS call traffic. The IRS local numbers are serviced by seven independent Regional Bell Operating Companies. Although the same reporting system as the toll-free reports would be ideal, cost and system compatibility, among other factors, need to be studied. A temporary solution is to apply our knowledge gained from Sprint data, such as the redial probability and its relationship to blocking probability, to estimating local line demand, assuming the customers' calling behaviors are alike between the two circuitries. The IRS is planning to do a study to examine this assumption and continue research on the models and assumptions.

■ References

Harris, C. M., Hoffman, K. L., and Saunders, P. B. (1987), "Modeling the IRS Telephone Taxpayer Information System," *Operations Research*, 35, 504-523.

Opitz, D. J. (1988), "Alternative Methodologies for Estimating Demand for IRS Toll-free Phone Service," Unpublished manuscript, Department of the Treasury.

Stone, R. (1989), "Comments on Two Approaches to Estimating the True Demand for the IRS Taxpayer Phone Service," Unpublished manuscript, AT&T Bell Lab. ■