# Time Series Analysis of Census Internet Response

**Fred Highland**
Lockheed Martin, Transportation and Security Solutions

## Abstract

Collection of census data over the internet promises to be more respondent friendly, accurate and cost effective than other methods. However, the characteristics of internet response over time are not well understood limiting the ability to accurately model, predict and manage take-up. This work theorizes that census internet response can be modeled as a time series of exponentially decreasing response probabilities associated with stimulating events. The response can be modeled as statistical distributions that characterize the events. This modeling is applied to census internet response data provided by Canada and the US to determine the common characteristics of response events and modeling parameters. The approach provides a means to model internet response patterns over time when calibrated to population and the survey methodology characteristics.

KEY WORDS: Census, Internet, Time Series

## 1. Background

Collecting census data over the internet promises to be the most cost effective and accurate response mode available. Once the initial development and infrastructure are paid for, the incremental costs are small compared to paper, telephone or personal interview (field data collection) methods. Studies have shown that the data collected is more consistent, accurate and complete than other methods (Roy and Laroche, (2006)).

However, the time-series profiles and rates for a census internet response are not well understood. These profiles are characterized by extreme usage peaks over short collection periods. These characteristics can cause significant problems for the design and management of the systems that support internet response as well as impacting user experience.

This paper seeks to understand the nature of census internet response and develop models of its behavior. Modeling census internet response will allow more accurate prediction of technology needs thereby optimizing system design. It will also provide a basis for monitoring and predicting response during the short collection periods allowing systems to adapt to unexpected behaviors. Finally, understanding the nature of census internet response will enable improvements in the response process and potentially increase overall response rates

The paper will first discuss three major tests of internet data collection: the US 2005 National Census Test, the Canada 2004 Census Test and Canada 2006 Population Census[1]. It will then develop a theory and models of the nature of internet response and apply the models to data from US and Canada internet tests. The results will be analyzed and recommendations made to aid in prediction of the response profiles for future internet census activities.

### 1.1 Canada 2006 Census

In 2006, Canada conducted a Census of Population which is done very 5 years. Canada's 2006 Census represents the most complex and sophisticated census data collection effort up to that date incorporating responses from paper questionnaires, internet response, and telephone interviews.

In order to ensure that processes and systems for the 2006 census worked together correctly, Canada conducted a Census Test in 2004. The Census Test involved 240,000 households from selected areas of Canada. Each household was given the option of responding by paper, internet or telephone. The internet was planned to be 18% of total take-up (~43,000 responses). Response to the test was voluntary. Lessons learned from the test were applied to the systems and processes being developed for the Canada 2006 Census of Population. Actual results showed approximately 9% internet take-up which was in line with expectations from a voluntary census test.

In 2006, Canada conducted a nationwide Census of Population. Participation involved approximately

[1] This paper contains response trend data provided courtesy of Statistics Canada and the US Census Bureau. These data are internet response statistics and contain no confidential or personal information. Statistics Canada exclusively operated and administered these systems and processes using its own staff and facilities.

31,600,000 residents in approximately 13,000,000 households across the six time zones of Canada. Internet was planned to comprise 15% of the total response. Approximately 2,000,000 internet responses were expected with the peak occurring on Census Day (April 18, 2006). A 15% internet take-up was achieved in the first month of internet availability. At the completion of the census internet take-up was 18% of all responses.

### 1.2 US 2005 Test

As part of its program to perform the US 2010 Census, the US Census Bureau conducted the 2005 National Census Test (NCT). The overall objectives of this test were to improve reporting completeness and accuracy, improve coverage accuracy, determine the feasibility of targeted mailing of replacement questionnaires, improve self response while maintaining data quality using bilingual questionnaires and reduce respondent and data capture errors (Boone, (2005) and Tancreto (2006)). As part of this activity, tests of internet data collection were performed. The NCT had a sample population of 420,000 households from selected areas across the US. The expected internet take-up rate was 10%. The actual internet response rate was 7.3% of the sample population or 12.1% of the total responses after adjusting for non-response.

## 2. The Nature of Internet Response Behaviour

### 2.1 Behavioral Hypothesis

Based on observations of census internet response time series data the following hypothesis is presented:

*Hypothesis: Census internet response patterns are defined by a series of exponentially decreasing response probabilities associated with stimulating events.*

This hypothesis has three major elements. First, that the response probability is exponential in nature. Second, activities are triggered by stimulating events. Third, that census internet response is not a single activity but a series of activities. Each of these elements is discussed further below.

The underlying mechanism that defines census response is the queuing of tasks to be serviced by the respondents. The response to a census over the internet can be viewed as one of many tasks respondents must perform in their busy lives. These tasks arrive at a roughly linear rate over time. The probability that a task gets addressed at a given point in time is a function of the tasks priority on the queue of other tasks. The priority of the task is affected by many things but primarily recency. That is, the more recently a task is placed on the queue or the respondents attention is drawn to the task, the more likely it is to be serviced. The probability of responding to a task in this way is well known in queuing theory and follows an exponential distribution.

Response is triggered by an event. An event is any stimulus that raises the priority (probability) of response by the respondent. This could be the initial arrival of the census form or invitation letter containing internet access information, a reminder letter, Census Day (providing publicity makes the respondent aware of Census Day and its importance) or other events that cause the respondent to fill out the census form on the internet. A less obvious stimulus apparent in the data is weekends. Weekends are periods of free (non-work) time that allow respondents to reassess their work queues and priorities. This stimulates response beginning on the weekend and reminds them to respond in the coming days.

The nature of response and non-response leads to the internet return profile being a series of events rather than a single event. More accurately, the overall response profile follows an exponentially decreasing probability distribution with stimulating events increasing the probability of the remaining non-responders to act. This results in a series of exponential response probability profiles defined by the stimulating events with a decreasing amplitude based on the remaining non-responders.

### 2.2 Exponential Representation

The profile of census internet response derived form the hypothesis discussed above can be modeled in standard exponential form as:

$$R_t = \frac{\omega_{event}}{\lambda_{event}} e^{-\frac{(t - t_{event})}{\lambda_{event}}}$$

for $t \geq t_{event}$ and $t < t_{event\ n+1}$

Where

- $R_t$ – *responses at time t*
- $t$ – *time*
- $\omega_{event}$ – *scaling parameter for the event*
- $t_{event}$ – *start time of the event*
- $t_{event\ n+1}$ – *start time of the next event*
- $\lambda_{event}$ – *average response time (time by which 63.2% have responded)*

This equation defines a series of exponentially shaped peaks based on the characteristics of the stimulating event defined by the event time ($t_{event}$), weight or impact ($\omega_{event}$) and the average response time ($\lambda_{event}$). The values for each of these parameters are not constant but depend on the nature of the event.

### 2.3 Gamma/Weibull Representations

The modeling of census internet response as an exponential probability distribution assumes that the triggering event occurs at a single point in time and applies to all respondents simultaneously. This is not always the case. For example, the mail system may not deliver census forms or reminders to all respondents on the same day, awareness of census day may be affected by work schedules and availability of weekend time may be impacted by other priorities. This suggests that the stimulus be applied across time following a distribution that is approximately normal making the overall distribution more complex than a simple exponential form. One such distribution is the Gamma distribution which represents the composition of a series of Poisson processes. The formulation for census internet response based on a Gamma distribution is as follows using the same variables as above with the addition of the shape parameter $\alpha_{event}$:

$$R_t = \frac{\omega_{event}}{\lambda_{event}} \frac{t^{(\alpha_{event}-1)}}{\lambda_{event}^{\alpha_{event}} \Gamma(\alpha_{event})} e^{-\frac{(t-t_{event})}{\lambda_{event}}}$$

for $t \geq t_{event}$ and $t < t_{event\ n+1}$

Another possible distribution to represent this phenomenon is the Weibull distribution (Weibull (1951)). While not having a general theoretical justification, the Weibull distribution has been found well suited to model many naturally occurring events and is widely used in reliability engineering. A Weibull formulation for census internet response is presented below:

$$R_t = \frac{\omega_{event}}{\lambda_{event}} \frac{\alpha_{event} t^{(\alpha_{event}-1)}}{\lambda_{event}^{\alpha_{event}}} e^{-\left(\frac{(t-t_{event})}{\lambda_{event}}\right)^{\alpha_{event}}}$$

for $t \geq t_{event}$ and $t < t_{event\ n+1}$

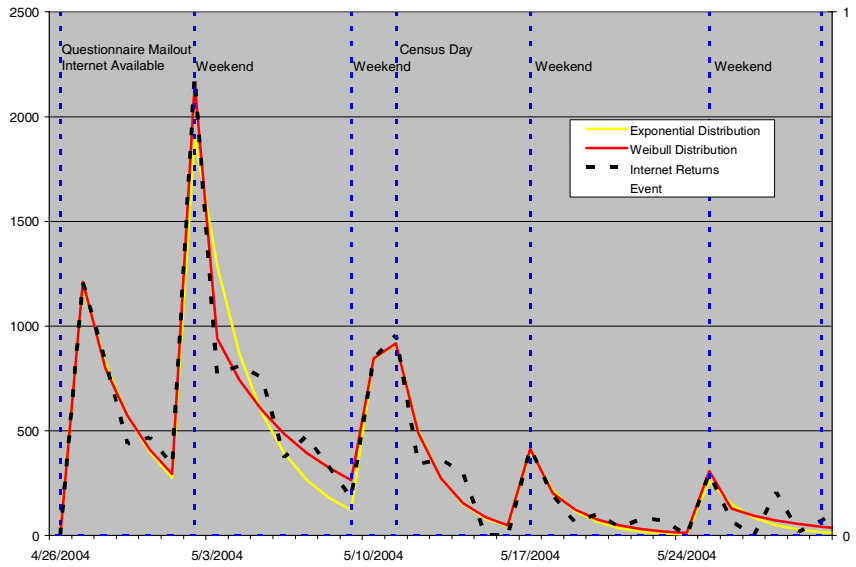
As will be seen below, in application, the two distributions are nearly identical in terms of results. The Weibull distribution model will be used for examples in the remainder of the paper.

## 3. Application to Real Data

The models discussed above were applied to census data from the Canada 2004 Census Test, the Canada 2006 Census and the US 2005 National Census Test. The results of these applications can be seen in Figures 1-3 below.

In each of these figures, the chart shows the actual distribution of response volumes as black dashed line plots, the best fit exponential model as yellow line plots and the best fit Weibull model as red line plots. The table below each chart shows the events and the best fit values for event time ($t_{event}$), weight ($\omega_{event}$) and average response time ($\lambda_{event}$). The table also shows cumulative Root Means Square Error (RMSε) and cumulative error for the distributions as percentages providing a relative measure of the accuracy of the prediction. Best fit distributions were determined by adjusting the model parameters to obtain minimum RMSε for each event.

As can be seen from each of the plots, the exponential model provides a good fit for the events selected. The Weibull model provides a slightly better fit in all cases exhibiting better RMSε and total error. Similar results were also obtained for a Gamma model with performance only slightly worse than the Weibull but still better than the exponential model.



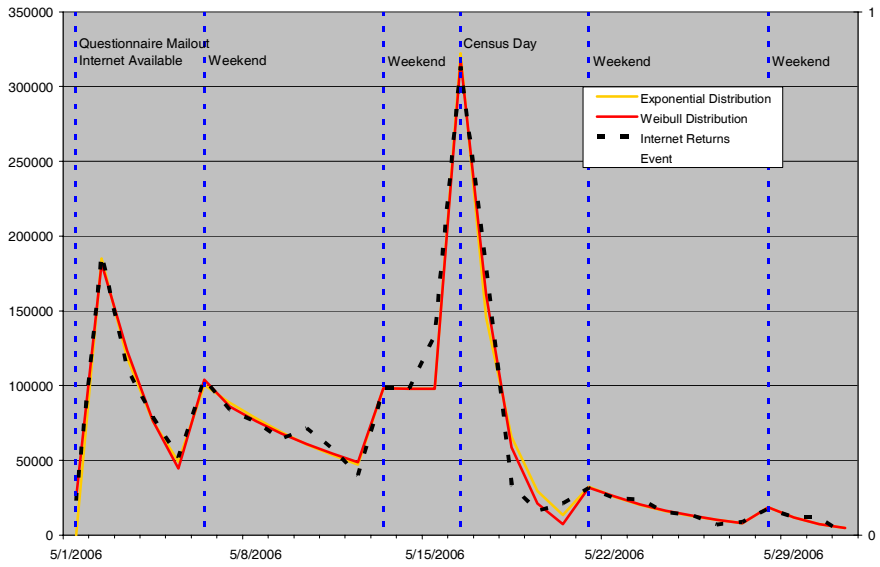

Exponential Distribution

| Event | Initial Quest | Weekend | Weekend | Census Day | Weekend | Weekend | | |
|---|---|---|---|---|---|---|---|---|
| $t_{event}$ | 04/27/2004 | 05/02/2004 | 05/10/2004 | 05/11/2004 | 05/17/2004 | 05/25/2004 | | |
| λ | 2.74 | 2.55 | 1000.00 | 1.66 | 1.64 | 1.80 | RMSε% | 0.9% |
| w | 1187 | 1906 | 847 | 912 | 399 | 261 | Σε% | -3.2% |

Weibull Distribution

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| α | 0.99 | 0.96 | 1.00 | 1.00 | 0.98 | 0.96 | | |
| λ | 3.03 | 4.94 | 1000.00 | 1.74 | 2.09 | 3.77 | RMSε% | 0.5% |
| w | 1110 | 1139 | 847 | 872 | 330 | 167 | Σε% | 0.3% |

**Figure 1 - Canada 2004 Census Test Internet Response - Actual, Exponential and Weibull Models**



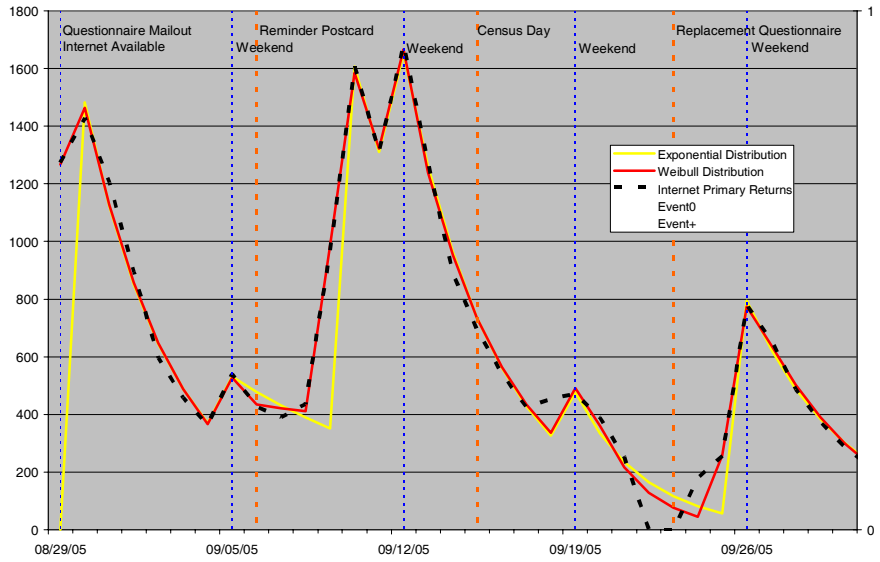

Exponential Distribution

| Event | Initial Quest | Weekend | Weekend | Census Day | Weekend | Weekend | | |
|---|---|---|---|---|---|---|---|---|
| $t_{event}$ | 05/02/2006 | 05/06/2006 | 05/13/2006 | 05/16/2006 | 05/21/2006 | 05/28/2006 | | |
| λ | 2.26 | 7.94 | 1000.00 | 1.26 | 4.34 | 2.22 | RMSε% | 0.62% |
| w | 184756 | 100620 | 98055 | 322315 | 32340 | 18535 | Σε% | -2.3% |

Weibull Distribution

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| α | 1.17 | 0.99 | 1.00 | 1.02 | 1.00 | 1.00 | | |
| λ | 2.17 | 8.98 | 1000.00 | 1.01 | 4.32 | 2.22 | RMSε% | 0.49% |
| w | 265016 | 95460 | 98055 | 424565 | 32735 | 18535 | Σε% | -1.7% |

**Figure 2 - Canada 2006 Census Internet Response - Actual, Exponential and Weibull Models**




Exponential Distribution

| Event | Initial Quest | Weekend | Reminder | Weekend | Weekend | Weekend | | |
|---|---|---|---|---|---|---|---|---|
| $t_{event}$ | 08/30/2005 | 09/05/2005 | 09/10/2005 | 09/12/2005 | 09/19/2005 | 09/26/2005 | | |
| λ | 3.61 | 9.64 | 4.91 | 3.70 | 2.84 | 4.10 | RMSε% | 0.7% |
| w | 1482 | 531 | 1606 | 1652 | 476 | 795 | Σε% | -7.9% |
| Weibull Distribution | | | | | | | | |
| α | 1.03 | 0.99 | 1.05 | 1.00 | 1.02 | 1.10 | | |
| λ | 3.56 | 46.20 | 4.60 | 3.84 | 1.93 | 3.84 | RMSε% | 0.1% |
| w | 1933 | 430 | 1993 | 1608 | 602 | 1010 | Σε% | 0.1% |

**Figure 3 - US 2005 National Census Test Internet Response - Actual, Exponential and Weibull Models**

## 4. Analysis

The above applications of the modeling to real census data show a very good fit of the models confirming the underlying hypothesis. The exponential model provides a good fit and relatively simple analysis due to its smaller number of parameters. The more complex Weibull model provides a better fit at the expense of more complex parameter estimation. Both models map well into plausible major events that occur as part of census data collection.

In order to normalize the parameter results from this data, the model components were averaged in the table below.

| Distribution | Event | $\alpha_\mu$ | $\alpha_\sigma$ | $\lambda_\mu$ | $\lambda_\sigma$ |
|---|---|---|---|---|---|
| Exponential | Initial Quest | | | 2.87 | 0.68 |
| | Weekend 1st | | | 6.71 | 3.70 |
| | Weekend | | | 2.95 | 1.11 |
| | Census Day | | | 1.46 | 0.28 |
| Weibull | Initial Quest | 1.07 | 0.09 | 2.92 | 0.70 |
| | Weekend 1st | 0.98 | 0.02 | 20.04 | 22.75 |
| | Weekend | 1.01 | 0.04 | 3.14 | 1.02 |
| | Census Day | 1.01 | 0.02 | 1.37 | 0.52 |

From the resulting model parameters, census events can be classified into two general categories: Weekends and Census Day. Weekends, which include the initial questionnaire distribution, have a λ of approximately 2.9 days using the exponential model. An exception to this is the data for the first weekend after distribution which sometimes has a greater λ value. The parameters for the Weibull model are similar for with an $\alpha \approx 1$ and $\lambda \approx 3$.

Census Day exhibits different parameters with a λ of 1.4 for the exponential model and $\alpha \approx 1$ and $\lambda \approx 1.4$ for the Weibull model. It should be noted that $\alpha=1$ for a Weibull and Gamma distribution results in an exponential distribution. Hence, while the more complex Weibull and Gamma models produce a better fit, they are almost exponential in nature.

The weight ($\omega_{event}$) parameters of both models can also be partitioned into the same two categories based on their observed characteristics. For Census Day a weight of approximately 15% of total response is suggested by the data from Canada 2006 Census. This

data is heavily influenced by a number of factors including advertising and cultural factors of the respondent population.

The weight parameters for weekends decrease over time appearing to follow the expected exponential decay as shown in Figure 4 - Normalized Internet Weekend Peak Responses. This graph of average peak response on weekends as a percentage of total response shows that the response levels exponentially decrease over time from a peak of ω=8.3% with a decay period of λ=2.8 weeks (excluding census week)

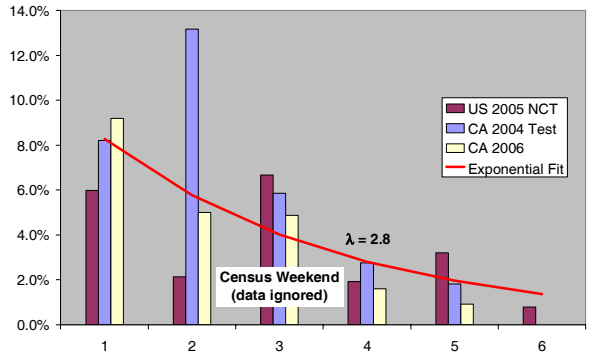



**Figure 4 - Normalized Internet Weekend Peak Responses**

Although the Weibull and Gamma models fit the data better than the exponential models, in actual application they are impractical. With the limited data available and the high degree of variation in the data due to both population and survey characteristics, use of average values for α and λ for Weibull and Gamma distributions produces results with larger RMSε variation than simple exponential models. Hence the exponential models are preferred for estimating future census activities.

As an example of the use of this modeling approach and the recommended parameters, Figure 5 - Canada 2006 Census Internet Response - Actual vs Recommendations shows the application of this data to the 2006 Canada Census data. The dark blue line in this figure is the predicted response prior to the census. The black dashed line is the actual response. The yellow line is the predicted exponential response based on the parameters presented here. Not that this tracks closely to the actual data. The red line is the predicted Weibull response based on the parameters presented here. The Weibull prediction tracks closely to the actual data but not as closely as the exponential prediction.

## 5. Conclusions

The paper hypothesized that census internet response patterns can be defined as a series of exponentially decreasing response probabilities associated with stimulating events. Application of that hypothesis to Canada and US census data using models based on exponential, Weibull and Gamma distributions has shown a good fit with the data justifying the approach as a way to model census Internet response.

While Weibull and Gamma based models produce a better fit to real data with lower error characteristics, the complexity of their parameters makes them inadequate as predictors of census Internet response patterns.

The preferred modeling approach is based on an exponential distribution model defining response timeframes based on two classes of stimulating events. The first includes the initial distribution and all weekends. It is defined by an exponential function with λ≈2.9 and ω exponentially decaying with time according to the rate ω = 8.3% Exp (2.8 weeks).

The second class is Census Day which has a higher peak (weight) and faster decay rate than normal weekends due to its significance. For Census Day modeling should be done with λ≈1.4 and ω≈15%.

Finally, it should be noted that, while this analysis provides a general basis for modeling census internet response, calibration to the methodology and the culture of the respondent population are necessary to adjust for factors beyond the scope of this analysis.

## Acknowledgements

The author would like to thank Statistics Canada, particularly Graeme Gilmore and Anil Arora, and the US Census Bureau, particularly Suzanne Fratino, for providing data for this work.

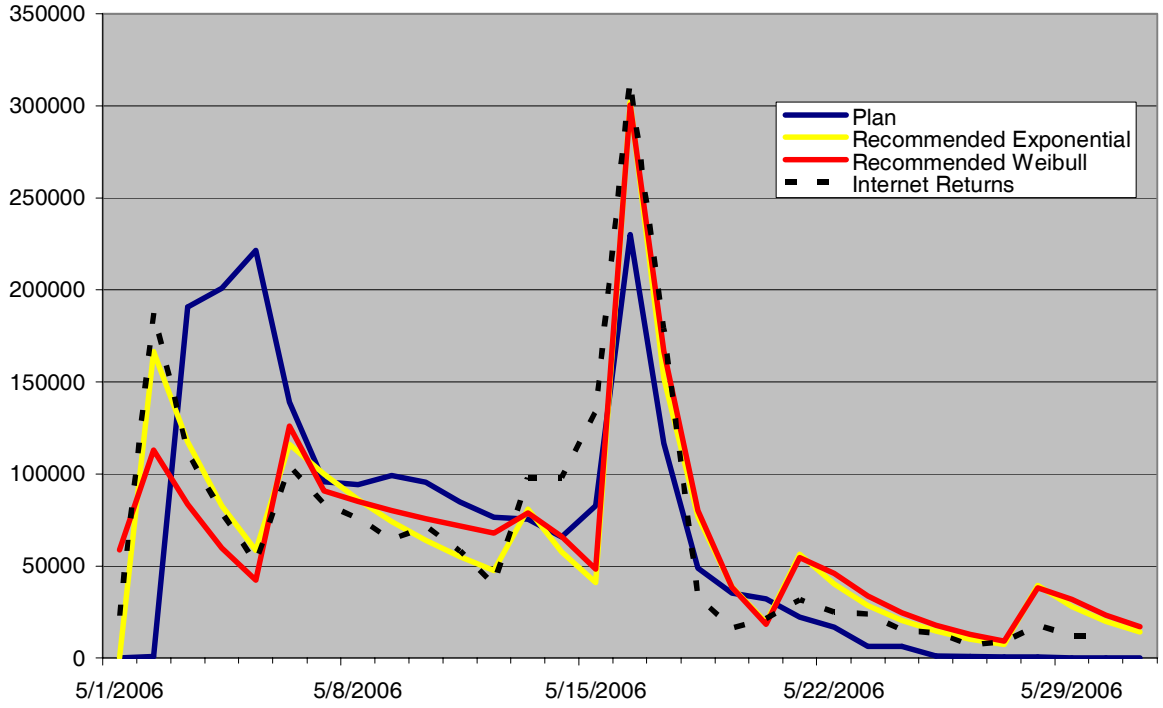



**Figure 5 - Canada 2006 Census Internet Response - Actual vs Recommendations**

## References

Roy, L. and Laroche, D. (2006), "The Internet Response Method: Impact on the Canadian Census of Population Data," 2006 JSM Proceedings, Section on Survey Research Methods [CD-ROM], Alexandria, VA: American Statistical Association: 3622-3628.

Boone, T. (2005), 2005 National Census Test, Decennial Management Division, U.S. Census Bureau, 3/17/05.

Tancreto, J. G. (2006), "An Overview of the 2005 National Census Test," 2006 JSM Proceedings, Section on Survey Research Methods [CD-ROM], Alexandria, VA: American Statistical Association: 3764-3771.

Weibull, W. (1951), "A Statistical Distribution Function of Wide Applicability," Journal of Applied Mechanics, September 1951, 293-297.