# Estimating the Accuracy of User Surveys for Assessing the Impact of HPC Systems

**Melissa Rishel**[1,2], David L. Hart[2], Doug Nychka[2]
mrishel@ucar.edu, dhart@ucar.edu, nychka@ucar.edu
[1]University of Northern Colorado, [2]National Center for Atmospheric Research

## The Problem: How well do surveys perform in compiling the publication impact of shared HPC systems?

NCAR's Computational & Information Systems Laboratory (CISL) operates supercomputing, data storage and archive systems that support more than a thousand users each year. These users pursue science objectives in the areas of climate, atmospheric sciences, and related areas. Each year CISL conducts a survey of current and recent HPC system users to collect publications information to help describe the scientific impact of CISL resources.

In this project, we attempted to answer two primary questions:

Q1: How well does CISL's annual survey capture the actual number of publications published by the HPC users, based on the research of sampled users?

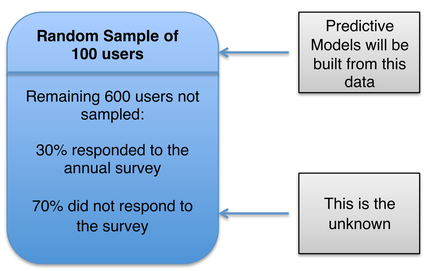
Q2: What are the error bounds on the total number of publications in the previous fiscal year, as measured from sampled users?

**Yellowstone** at the NCAR-Wyoming Supercomputing Center

## The Purpose

The purpose of this project was to determine if the annual survey has a reasonable degree of accuracy in capturing the desired information, impact of our systems.

**700 users of HPC Systems received the CISL annual survey for fiscal year 2014**




## Methods

The predictive models were built using the Poisson regression for modeling count data since our response variable was the number of publications published in the fiscal year per user. The logarithm of the count of publications is linked to the linear function of the explanatory variables thus ensuring we get a predicted value of zero or greater.

A generalized linear model was fit using R as follows for expected count of publications:

Model 1: Users who responded to survey

$$log(u1) = \beta 0 + \beta 1\, x1 + \beta 2\, x2 + \beta 3\, x3 + \beta 4\, x4 + B5\, x5$$

Model 2: Users who did not respond to the survey

$$log(u2) = \beta 0 + \beta 1\, x1 + \beta 2\, x2 + \beta 3\, x3 + \beta 4\, x4$$

Here, u1 and u2 are expected number of publications for the user for each model
*B*1, *B*2, *B*3, *B*4, *B*5 are the covariates for modeling.

Table 1: Totals for sampled users Training Models

| Model Users Responded to Annual Survey | Total Core Hours Used: HPC | DAV | HPSS | Glade | Total Number of Publications found during research | Total Number of Publications Noted on annual Survey |
|---|---|---|---|---|---|---|
| Model 1 Yes: 26 users | 2,417,019.36 | 13,043.14.00 | 67.21 | 15.28 | 28 | 35 |
| Model 2 No: 74 users | 26,354,916.39 | 14,079.42 | 271.36 | 0 | 30 | NA |
| **TOTAL: 100 users** | **28,771,935.75** | **27,122.56** | **338.57** | **15.28** | **58** | **35** |

Table 2: Total for users not included in sample

| Users not included in sample | Total Core Hours Used: HPC | DAV | HPSS | Glade | Total Number of Publications Noted on annual Survey |
|---|---|---|---|---|---|
| Users who responded to survey | 344,403,99.00 | 265,432.00 | 8,268.39 | 848.25 | 282 |
| Users who did not respond to survey | 122,736,103.00 | 120,316.32 | 2,211.17 | 859.60 | NA |
| **TOTAL: 600 users** | **467,140,093** | **385,748.32** | **10,479.56** | **1,707.85** | **282** |

*HPC: Yellowstone high performance computing resources, DAV: Geyser & Caldera analysis & visualization systems, Glade: centralized file systems & data storage, HPSS: data archive

Working through the data, we found several outliers which influenced the ability of our training models to accurately predict the new data.

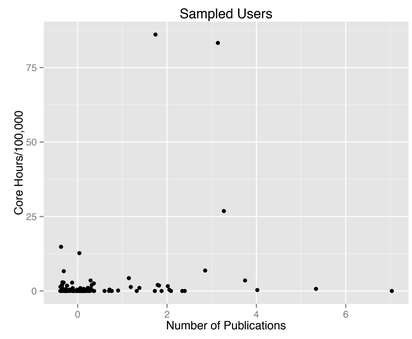



Figure 1: Users from training data

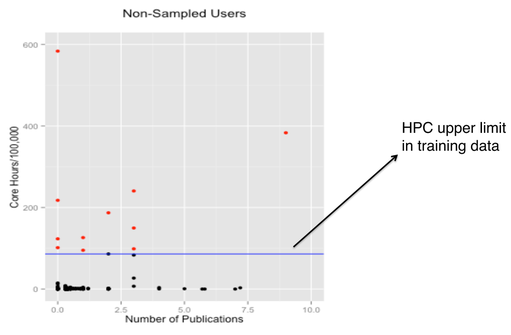



Figure 2: Publications predicted for users with users exceeding upper limit

There were some users who exceeded the upper limit of our training data of HPC usage, these users had to be omitted from the prediction data to not over inflate the number of predicted publications and standard errors.

## Results

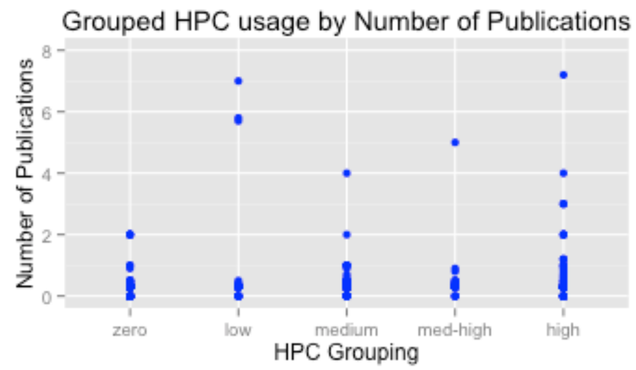



Figure 3: Plot of HPC Grouping by Number of Publications, Predicted and Researched

In Figure 3, we can see that when users were grouped according to usage categories, most of the publications per category cluster below 2. This indicates that high usage may not be a great predictor in the total number of publications published.

Table 3: Total of Number of Publications with Standard Errors from remaining 512 users

| Source | Number of Publications | Standard Error |
|---|---|---|
| Project Research | 58 | 0 |
| Prediction: Model 1 | 73 | 11 |
| Prediction: Model 2 | 137 | 12 |
| **Totals** | **268** | **33** |

A1: The total number of predicted publications is less than the total amount from the user surveys. This amount could be adjusted taking into account the users removed from the data.

The top 14 HPC users removed from the prediction data reported 22 publications and these users account for more than 250,000,000 core hours.

A2: We would expected to have an error bound of +/- 33 publications based on our training model for the prediction data.

**Based on these models, it appears the annual survey is doing a good job capturing the number of publications for the users.**

Future considerations:

It is evident from the high HPC users that high usage does not equate a high number of publications. Rather it is the impact a single paper has on the scientific community. This is more difficult to quantify.

Stratified sampling in the future, would ensure a proper representation of each group of HPC users and help with creating a better prediction model.

Repeating this project for previous year's data would help determine the accuracy of the error bounds.

**Acknowledgements**

This work is supported by NSF AGS-0753581, which provides support for NCAR. The study's authors gratefully acknowledge the many users of Yellowstone (ark:/85065/d7wd3xhc) who cooperated in this study.