ABSTRACT
Each year, the Computational & Information Systems Laboratory (CISL) conducts a survey of its current and recent user community to gather a number of metrics about the scientific impact and outcomes from the use of CISL’s high-performance computing systems, particularly peer-reviewed publications. However, with a modest response rate and reliance on self-reporting by users, the accuracy of the survey is uncertain as is the degree of that uncertainty. To quantify this uncertainty, CISL undertook a project that attempted to provide statistically supported limits on the accuracy and precision of the survey approach. We discovered limitations related to the range of users’ HPC usage in our modeling phase, and several methods were attempted to adjust the model to fit the usage data. The resulting statistical models leverage data about the HPC usage associated with survey invitees to quantify the degree to which the survey undercounts the relevant publications. A qualitative assessment of the collected publications aligns with the statistical models, reiterates the challenges associated with acknowledgment for use of HPC resources, and suggests ways to improve the survey results further.

CCS Concepts
• General and reference–Metrics 
• Social and professional topics–Management of computing and information systems

Keywords
Science impact; supercomputers; user publications.

1. INTRODUCTION
The Computational & Information Systems Laboratory (CISL) at the National Center for Atmospheric Research (NCAR) 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, supported by hundreds of National Science Foundation (NSF) grants or in pursuit of theses and dissertations. The value of the CISL high-end computing environment to NCAR management and to the National Science Foundation (NSF), NCAR’s sponsor, can be described in part by anecdotal examples of user research; however, a more encompassing assessment requires a quantitative understanding of the scientific outcomes from the users, which in turn provides insight into the ultimate “impact” of the high-end computing facility. To quantify these scientific outcomes, CISL conducts an annual survey of current and recent HPC system users to collect data on four relevant metrics: the number of peer-reviewed publications, the number of other citable works, the number of graduate students using the system, and the number of theses and dissertations produced. These metrics, particularly when associated with resource usage data and the users’ institutional affiliations, provide some insight into the scientific impact of CISL resources.

However, surveys are imperfect collection methods. In a typical year, CISL’s annual publications survey has a response rate of approximately 25%. Because CISL’s annual survey is aimed at collecting several data sets that could be quantified and assembled with 100% accuracy, if we had unlimited access to perfect information, the survey provides few clues as to how accurately it produces the desired metrics. To quantify this uncertainty, CISL undertook a project that attempted to provide statistically supported answers two primary questions: (1) How well does CISL’s annual survey capture the actual number of peer-reviewed publications published by the HPC users in the previous fiscal year? That is, how accurately does the collected total match the “true” number? (2) Can we put error bounds on the precision of the number of publications collected relative to the “true” number?

2. RELATED WORK
Interest in the scientific impact of “research infrastructures” is not limited to shared computing facilities. Many studies over the past several decades have pursued the goals of collecting and identifying relevant publications as well as producing quantitative metrics [9]. Because citation practices with respect to research infrastructures are ill-defined and widely varying, Mayernik, et al., noted a range of challenges related to collecting relevant publications as well as the variety of manual, automatic, and approximate approaches that have been developed and applied. In particular, experience at NCAR and from other projects (e.g., [5]) has shown that open-ended self-reporting of publications by users does not always produce good results. Self-reporting as part of another mandatory process—such as the allocation request and review process used by most shared HPC facilities—may improve the collection results, but “gaps” in the process may still limit the effectiveness.

In contrast to exhaustive manual searches—such as the best practices recommended by the International Astronomical Union
for telescope bibliographies [6]—surveys are relatively easy to conduct and, in theory, collect publications and other metrics from those persons who are in the best position to accurately identify relevant publications—the users themselves. In contrast to efforts to automate publication searches (e.g., [11]), a survey does not depend on web site idiosyncrasies or journal licensing to collect the relevant citations.

The effectiveness of self-reporting via surveys has been studied extensively, particularly in the areas of health and wellness (see, e.g., [12]). And while findings based on self-reports in surveys may have some limitations, those findings are not inherently inaccurate or without value [3]. If the purpose of the CISL survey were to gauge the opinions or characteristics of the user community, we would probably consider 25% a strong response rate and be satisfied that the results represented, at least roughly, the rest of the user community. By contrast, we wanted to assess in this study how well the survey worked in assembling a well-defined collection of publications and other metrics.

3. CISL PUBLICATIONS SURVEY

Each year, in roughly January or February, CISL has sent out a survey to canvass its user community and measure the impact of CISL’s high-end computing environment. The survey itself is brief, with a few short-answer questions about the user, their computing project identifier and supporting NSF grants, four free-text questions for gathering the key impact information, and a few rating questions about their use of and satisfaction with the resource environment. Responding to most questions is optional. Respondents are given approximately one month to respond, with three reminder emails sent to non-respondents after the initial invitation.

3.1 Invitees and Response Rate

The survey is sent to individuals who were identified as the lead user on projects that had been open at some point in the prior two years. In this case, a “project” is an entity in our allocations and accounting system; we also extracted the relevant resource usage by the project in the two-year period. Most projects (about two-thirds) had some recorded resource use in the period; others may have used CISL resources prior to that period. We used a two-year period due to the delays that can exist between a researcher using a resource and the eventual publication of that work in a journal, conference, or other venue.

The response rate for these annual surveys using these invitation criteria has typically ranged from 23% to 30% since 2008. (In the one exceptional case, we expanded the invitation list to include all users on projects, instead of just the lead user, which doubled the number of survey invitations and returned the fewest total responses, for an 8% response rate.) The number of collected publications (peer-reviewed, other, and theses or dissertations) ranged from 286 to 442, with peer-reviewed publication counts ranging from 116 to 281. The survey asked specifically for publications that fell within the prior fiscal year. Varying degrees of data “cleaning” were applied to remove duplicates and exclude publications outside of the fiscal year period.

3.2 Survey Response Validation

We based our research project on the recent CISL annual survey sent out in January 2015 to collect publications from fiscal year 2014—October 1, 2013, to September 30, 2014. The survey was sent to 750 invitees and garnered 222 responses, for a 30% response rate. To make the process manageable and to support the statistical approaches, we first selected a simple random sample (n=100) of the survey invitees on which to base our analyses. The sample included both users who responded to the annual survey and those who did not respond. Consistent with the survey overall, the random sample had 26 survey respondents and 74 non-respondents. In addition, the distribution pattern of HPC system usage in the sample was similar to that of the overall population—that is, a small fraction of users account for most of the usage with a long tail of users with relatively little or no usage. This highly skewed, asymmetrical distribution for usage is in fact the normal distribution pattern for HPC systems and should be accounted for in the statistical models [4].

Validating the results then proceeded in two phases. In the first phase, we needed to establish the correctness of the responses for the sample. For each invitee in the sample, an exhaustive manual search for publications and other citable works was conducted via Internet search sites, Thomson Reuters Web of Science, Google Scholar, and an internal NCAR library system. We accumulated a list of publications that we could then compare with what was reported on the surveys. Each user in the sample, therefore, fell into one of four possible categories:

1. **Respondent Match:** The user responded to the survey, and the response matched our search results. Twenty users responded to the annual survey with reported data that matched our search; these users had 9 publications published during the previous fiscal year. Most of these publications did not mention use of any NCAR models or data sets.

2. **Respondent Non-match:** The user responded to the survey, but the response did not match our search. Five users responded to the survey but their response did not match our search; these responses included five extra publications. Most of the publications that did not match were due to the fact that they were published outside of the considered time period.

3. **Non-respondent Match:** The user did not respond to the survey, and our search turned up nothing, that is, the user had nothing to report. Most of the sampled users (57) fell into this category. Over half of these users had zero or negligible usage. The remaining users recorded some use of the HPC system, and we would expect them to publish in the future.

4. **Non-respondent Non-match:** The user did not respond, but our search found one or more relevant publications. Seventeen users did not respond to the annual survey and our search showed relevant publications. We found 30 publications published from these users; eight of the users referenced models or data usage from NCAR.

To further confirm our searches, we contacted every user in the sample and requested brief interviews to confirm the findings from our search of publications and other citable works. We also used the interview to solicit feedback that could help improve the survey process. Most users were contacted two additional times for a request to participate in the interview. We conducted 52 interviews of the 100 solicited; 37 of these 52 users who responded to the interview request had not responded to the publications survey. Based on the interviews, one member of the sample was excluded from the results, essentially for misunderstanding the original survey and providing a very large amount of incorrect data. Thus, a sample of 99 was used for the modeling, still sufficient for the statistical methods. The remaining 15 users we interviewed had responded to the annual survey. Most of the relevant publications found in our publications research project came from users who did not respond to the annual survey but did respond to our interview requests.
3.3 Statistical Models and Methods
In the second phase, we evaluated the users’ HPC use and publication counts to determine the characteristics of the statistical distribution. From that distribution, we then used standard statistical methods to develop models for extrapolating the accuracy of the total survey results from the better-known accuracy within the random sample. From the interviews, we placed each user into one of the four match and non-match categories. We then used the matching information to build our training models. If the users responded to the annual survey, they were in Model 1, and if the users had not responded, they were in Model 2.

In Table 1, we see the different covariates considered for our models. The annual survey captured an array of data on each user; however, we limited our modeling data to the actual usage of CISL systems, the reported number of publications (for Model 1), and the total number of publications found during our research.

Table 1: System use and publication counts for sampled users

<table>
<thead>
<tr>
<th>Did user respond to survey?</th>
<th>Yes 26 users</th>
<th>No 74 users</th>
<th>Total 99 users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yellowstone core-hours used</td>
<td>2,417,019</td>
<td>26,354,918</td>
<td>28,771,936</td>
</tr>
<tr>
<td>Analysis &amp; visualization cluster used</td>
<td>13,043</td>
<td>14,079</td>
<td>27,123</td>
</tr>
<tr>
<td>HPSS archival storage (TB)</td>
<td>67.2</td>
<td>271</td>
<td>339</td>
</tr>
<tr>
<td>GLADE disk storage (TB)</td>
<td>15.3</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>Publications found during research</td>
<td>28</td>
<td>30</td>
<td>58</td>
</tr>
<tr>
<td>Publications noted in annual survey</td>
<td>35</td>
<td>n/a</td>
<td>35</td>
</tr>
</tbody>
</table>

Figure 1 visualizes the spread of the HPC usage, confirming that most users have modest usage for this time period, while a few users have heavy usage. Similarly, we looked at the distribution of the number of publications. Table 2 shows similar distributions for the total number of publications reported in three categories: all survey respondents, the survey respondents in our random sample, and all users in our random sample. These data also helped to determine the best distribution for modeling the survey results. In all cases, we see a clustering around 0–2 publications, which is expected as some users may have used the systems but not yet published or may not have used the systems in the reporting period, but some users do report higher numbers of peer-reviewed published publications. However, as already discussed, we found that the validity of some users’ reported publications was inaccurate.

From this information, we considered different distributions to use for modeling. Since we are dealing with count data, we explored the Poisson distribution [8]. This distribution allows us to model data with a unimodal trend—in this case, a clustering near zero. The few users with a high number of either reported or found publications give a positive skew to the distribution.
3.4 Statistical Model Building

We used the R statistical computing software to perform the analysis and develop a generalized linear model. We built two full models with all the covariates for each and checked for the significant covariates. The covariates of significance for Model 1 (survey respondents) were HPC usage and the reported number of publications. In Model 2 (non-respondents), we used all the covariates for modeling. After building the two models, we ran goodness-of-fit tests to confirm that the data roughly followed a Poisson distribution. A special property of the Poisson distribution is that the mean is equal to the variance, so one way to assess whether the data fits the Poisson distribution is to test the goodness of fit by looking at the ratio of the deviance of the model to the degrees of freedom of the residuals. Table 3 shows the calculated ratio of the deviance to degrees of freedom for both models. In both cases, the mean is approximately equal to the variance within an acceptable tolerance. (The literature suggests “close to one” as an acceptable tolerance [8], but a more precise test is beyond the scope of this project.)

Table 3. Goodness-of-Fit Tests

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>Degrees of Freedom for Residuals</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Users who responded</td>
<td>26.79</td>
<td>22</td>
<td>1.22</td>
</tr>
<tr>
<td>2: Users who did not respond</td>
<td>85.40</td>
<td>72</td>
<td>1.13</td>
</tr>
</tbody>
</table>

After building our training models, we analyzed the rest of the survey invitees, using the two models to predict the number of publications for each user. When doing this, we discovered an issue with our sample. While the general distribution pattern of HPC usage in the sampled users reflected the HPC usage pattern of the overall survey population, the random sample did not include some of the heaviest users of the HPC system. Therefore, our threshold for the maximum HPC usage inferred from the training data—8.6 million core-hours—fell substantially below the actual peak HPC usage in the survey population (for example, two projects had more than 30 million core-hours), and thus our regression model could not make predictions for users with HPC use beyond this threshold.

Figure 3 highlights the distribution of non-sampled users and their usage, showing that a number of users had usage above the threshold limit. Model 1 does not perform accurately beyond this amount and inflates the number of publications per user as it tries to fit beyond our regression model; we discovered a similar issue with Model 2. Therefore, we adjusted the models’ limits, restricting predictions to users whose usage fell within seven times the standard error of the HPC usage plus the median HPC usage of the sample modeling data. (The median HPC usage for the training data was 3,645 core hours and the standard error was 122,315.) We appreciate that excluding “power users” from the pool of users we are statistically modeling is much less than ideal. For the initial investigation, we used the number of publications reported in the survey by power users as the estimate for the true value. However, we note that power users did have a substantially higher response rate compared to other users, with 42 of 67 (63%) of these users responding. We then made predictions for the remaining survey invitees, the number of publications per user and, from this prediction, calculated the error bounds of our survey data.

Table 4 compares the results from the CISL survey, the predicted results from the modeling effort based on the researched sample of users, and a validation check by running the predictive model on all survey invitees. The validation check shows strong agreement with the prediction plus sampled users. Both predictions argue that the total number of publications “in the wild” is more than the 318 reported in the annual survey. However, this would be surprising since the models account for more users than the ones who responded to the survey. Furthermore, from the model we estimate the error bound in the predictions to be ±31 publications.

3.5 Model variations

After determining the sampling issue related to the wide range of HPC usage, we tried two approaches to work with the data. First, we included a post-stratification covariate that grouped each user into usage categories as shown in Figure 4. This approach was based on the statistical characteristics of the HPC usage distribution (standard errors plus a multiple of the median of HPC usage). While this new covariate helped with predicting a slightly lower total number of publications of 413, the standard errors increased unacceptably to ±47%.

Next, we tried a different post-stratification covariate of HPC usage based on CISL’s allocation policies and practices. In particular, “low” usage was defined as near the limit CISL sets for most graduate student and postdoctoral research allocations. “Medium” usage was defined near the limit for small, NSF grant-supported allocations. “Medium-high” and “high” usage both fell into the range of panel-reviewed allocation requests. In the end, however, this categorization was not sufficient to enable the models to perform. The number of publications was again close to 400, but the total standard errors increased further.
4. DISCUSSION OF RESULTS

Overall, the statistical methods employed suggest that the survey does a good job at capturing the true number of publications from users who do respond to the survey; the methods also provide a way to estimate the possible underreporting of publications due to non-respondents. In the most straightforward approach, the models predict the survey is underreporting as many as 40% of user publications. However, the project raised a number of issues that need to be examined in future work. For one, the manual research effort did find that, for some users who had reported a high number of publications, 20% (7 of 35) were not actually within the publications sought by the survey (i.e., outside of the fiscal year period or not directly supported by CISL resources). From this set of respondents, one might expect that the models should predict a lower number of publications from among the rest of the survey respondents due to “over-enthusiastic” reporting.

The more challenging issue related to the reliance on HPC usage as the primary covariate and the resulting limitations on the statistical models. Figure 4, for example, shows that even when users are grouped according to usage levels, the reported publication counts still cluster around 0–2 for all usage levels. Similarly, the models tended to break down when confronted by “power users” with very high levels of usage. In our initial attempts to work with the existing data, we elected to remove the heavy HPC users from prediction, although we recognize that this is a significant limitation of the current models. Future work will explore alternatives that will encompass all users. For example, taking a stratified random sample of the users within each category of usage may help the models make better predictions for each usage level. That is, we would sample 100 users after first grouping users into one of CISL’s categories (zero to high usage) and then select a fixed percentage randomly from each grouping.

These data and modeling challenges are consistent with the findings in other studies that larger usage does not necessarily translate into more publications. The top 14 HPC users removed from the prediction data reported 21 publications, while accounting for more than 250 million core-hours. That is, roughly half of the total core-hours for all respondents resulted in 7% of reported publications. Studies of other HPC user environments[2][13] and even of Hubble Space Telescope users[1][10] have echoed the lack of a clear relationship between large-scale use and publication quantities.

Thus, while HPC usage may be a useful indicator, it does not have a simple linear relationship with publication outputs. No use and perhaps close-to-zero use may indicate no publications—though the time lag associated with the analysis and publication process may complicate this assumption if the usage was from some prior period. But above some “critical threshold” of usage, any amount of usage may simply indicate that a user accomplished their scientific objectives. For example, graduate student may publish one or two articles from their dissertation work (which NCAR policy typically restricts to more modest level). But a “power user” may conduct a similar number of computational experiments but with models at higher resolution or in larger ensembles, thus using more of the system, while still producing a relatively small number of publications.

Future work may examine the utility of other possible covariates, not included in this study, such as the number of users working on a given project, or simply the number of projects being led by a given survey invitee. The number of users may affect publication output because more users actively involved in using the systems may indicate that more potential authors were pursuing several separate lines of inquiry. Similarly, at NCAR, an allocated project is typically linked to a single NSF award, or less commonly to a specific line of inquiry within an NSF award. Thus, more projects may translate to a greater expectation of publications.

5. QUALITATIVE ASSESSMENT

In addition to the statistical modeling work, we undertook a more qualitative examination of the survey and the publications to help us validate the statistical assumptions, to better understand how well the results might apply to prior or future surveys, and to see what the sampled publications might say about the full set of publications. For example, early in the project, we examined the response rate more closely, using information about each invitee’s resource usage. By excluding all invitees below a somewhat arbitrary but still fairly low minimum level of usage—in our case,
acquiring allocations or the user environment on the HPC systems. A few comments were related to acknowledging support from CISL and noted common challenges, including a lack of awareness. One user suggested that this requirement be included in award letters (something CISL has long done), and another did not understand the value of mentioning CISL or its HPC systems specifically, rather than a general blanket acknowledgment of NCAR. We have implemented one additional suggestion to remind users more frequently of the importance of acknowledging support from CISL resources.

6. CONCLUSIONS AND FUTURE WORK

Based on the manual investigation of sampled users and the statistical models developed, CISL’s annual survey appears to do a reasonably good job capturing the number of publications from the HPC users who respond to the survey, particularly when considering the relatively low level of time and effort required to conduct the survey. The manual investigation and resulting statistical models also provide a quantitative method for measuring the undercounting of publications by the survey, as well as bounds on the degree of accuracy in the survey results. Several improvements are being investigated for the models, including the use of improved stratified sampling to better represent each group of HPC users in the highly skewed usage distribution, as well as use of alternative covariates. Such refinements should help improve the prediction model. Additionally, with the excess of zeroes in the number of publications among invitees, another modeling distribution could be investigated—the Zero Inflated Poisson Regression, ZIP [8]. This ZIP model is able to handle the excess of zeroes and a Poisson regression model. Combining both a stratified random sample with a Zero Inflated Poisson Regression model may best handle data of this type.

The corresponding qualitative analysis of the sampled publications suggests that, while imperfect, survey respondents appear to do a reasonable job of following the instructions of the survey. The interviews with users also suggest that more personalized emails rather than more generic emails, such as used in the annual survey, may better capture a response from some users. Further constructive criticism received by contacting users in the random sample indicated that the accuracy of the survey results could likely be improved by two changes, which were implemented for the FY2015 survey. First, we have excluded CISL data services from the support services of interest, instead leaving only the rating question that assess the percentage of citations from CISL and noted common challenges, including a lack of awareness. One user suggested that this requirement be included in award letters (something CISL has long done), and another did not understand the value of mentioning CISL or its HPC systems specifically, rather than a general blanket acknowledgment of NCAR. We have implemented one additional suggestion to remind users more frequently of the importance of acknowledging support from CISL resources.

We also compared the survey results against input provided as part of CISL’s university allocations process. CISL accepts requests from university researchers for large-scale allocations twice per year. As part of these requests, the submitters are asked to include information about publications stemming from their work with prior allocations on CISL’s resources. Because the spring 2015 allocation request opportunity closely followed CISL’s FY14 publications survey, we extracted publication lists from these allocation requests and compared them to the results from the survey. If the survey was doing a good job at collecting publications, it should have captured many of the FY14 publications that these submitters included in their allocation requests. The allocation requests included 23 submitters who included one or more publications from previous work; a total of 96 unique publications were cited. All of them (or the lead on the prior project they had worked on) received a publication survey invitation, and all but two (91%) responded to the survey. Of the 96 publications, 31 appeared in the survey results. Upon close examination, only two of the 65 not reported in the survey would have been legitimate survey responses; the vast majority of the others were not published in FY14. Thus the accuracy of the survey in capturing publications identified by the allocations process was 93%, far more accurate for these relatively larger-scale users than the error bounds calculated for the statistical models.

We also wanted to examine characteristics of the publications from the sampled users to shed some light on the full set of publications collected by the survey. With respect to the fiscal year boundary, only seven publications could not be construed as being published within the requested time period, suggesting the time boundary may produce a 10% error in the “correct” set of publications. This ZIP model is able to handle the excess of zeroes and a Poisson regression model. Combining both a stratified random sample with a Zero Inflated Poisson Regression model may best handle data of this type.

The articles were also examined for references to NCAR, CISL, Yellowstone, or other NCAR-related components (models, software, data, and so on). About 32% definitively acknowledged CISL computing, albeit in a wide variety of ways. As a national laboratory providing a wide range of services to the climate and atmospheric science community, NCAR often presents a citation and acknowledgment challenge to authors. Reiterating the findings of many prior analyses of citing and acknowledging research infrastructures [9], authors variously and idiosyncratically acknowledged NCAR, CISL, the Yellowstone system, other NCAR-supported components, or some combination thereof. With the exception of the recurring “blanket” statement that NCAR is supported by NSF, the number of acknowledgment variations seemed to track the number of groups in the survey sample.

Finally, the interviews conducted with users in the sample did not offer much substantive feedback for improving the survey or the survey process. More comments discussed questions about

---

we set a minimum of 50 core-hours of HPC system use—we found that the response rate for the remaining invitees was considerably higher, approaching 40%. We made the same simple analysis of the response rate in an earlier year’s survey (508 invitations, 123 responses). After eliminating invitees below the same minimum level of usage, the response rate for that survey climbed to 35%. This simple analysis of the “meaningful” response rate argues that the survey results are somewhat closer to “truth” than they might first appear and that the results from the FY14 survey may apply reasonably well to other years.

“truth” that the response rate for the remaining invitees was

...
statistical models that rely too directly on HPC usage levels tend to inflate the number of publications expected from “power users” on the system. The most significant challenge remains compliance and consistency with CISL’s guidelines for acknowledging or citing support.

Further studies of similar surveys used by other shared user facilities and other types of research infrastructures may be warranted to solidify these findings. However, these initial results indicate that relatively simple surveys can be a cost-effective and efficient tool for compiling user impact bibliographies, significantly reducing the effort associated with exhaustive manual searches and avoiding investment in specialized tools for automated searches.

7. ACKNOWLEDGMENTS

This work is supported by NSF AGS-0753581, which provides support for NCAR, and NSF ACI-1448480. The authors gratefully acknowledge the many users of NCAR’s Yellowstone system (ark:/85065/d7wd3xhc) who cooperated in this study.

8. REFERENCES


