Using Statistics to Manage GIS Growth in a State Agency

Twenty-Seventh ESRI International Users Conference Proceedings, San Diego, CA

Richard C. Daniels and Jay W. Callar

Abstract

In small government agencies or offices the maximum user base for GIS can often be estimated prior to deployment of the technology. In larger agencies, or agencies that are distributed over a wide geographic area, the GIS manager may not be aware of all GIS initiatives underway. This makes scaling of the GIS environment difficult. A required step in addressing this problem is the development of a GIS usage tracking methodology. It is necessary to track both the number of GIS installations as well as to measure actual usage of the technology. Once a sufficiently long period of record has been compiled, usually twelve months, you will be able to develop usage projections, plan for growth, and justify funding requests. By planning in this way you will be able to encourage the use of GIS in your agency without being hampered by foreseeable resource constraints.

Introduction

In small offices or agencies the maximum potential user base for GIS can often be estimated prior to deployment of the technology. In larger agencies, or medium size agencies that are distributed over large geographic regions, this is often not the case since the GIS manager may not be aware of all GIS initiatives that are being undertaken -since the decision to use GIS for a given project is often driven more by the technical background of the project manager than by the capabilities of the available software solutions.

The Washington State Department of Transportation (WSDOT) is a large State Agency with approximately 7,800 employees, of these 7,500 have desktop or laptop computers. As one would expect for an agency of this type, it is geographical disperse across the State. To manage this dispersed work force the agency is divided into six Regions and the Washington State Ferry System. Thus WSDOT epitomizes the most challenging case for a GIS Manager, a large agency made up of geographically dispersed semi-autonomies departments.

GIS at WSDOT

GIS was first endorsed for use by the WSDOT Executive Committee in 1993. In 1994 this endorsement lead to the selection of ESRI’s ArcView 3.x software as the standard desktop GIS solution for the agency and ArcInfo as the solution for professional cartographers. However, this endorsement came with several caveats, chief among these were:

- GIS is a tool – users control its use,
- GIS funded from current budget levels – no new funding,
- GIS applications will be office level, not corporate, and that
- GIS will be incrementally implemented within the agency (GIS Support Team 2006).
In 1995 the WSDOT Information Technology Committee (ITC) adopted a stronger vision statement for GIS, “GIS is the foundation on which future information needs will be met” that encouraged the continued growth of GIS at WSDOT. By 1997-1998 the first GIS Data Administrator and GIS Programmer were hired within the MIS department (now the Office of Information Technology). A GIS Manager position was created in 1999 and a dedicated GIS Application Development Team was formed in 2006.

Notwithstanding the across the board growth in GIS usage at WSDOT (e.g., MacDonald 2004); GIS is still negatively influenced by the caveats imposed in 1993 by the WSDOT Executive Committee. Specifically, the GIS community has had difficulty obtaining funding to support the increased demand for GIS in the agency, and until recently, has not been consulted during the development of large corporate IT solutions.

**Monitoring License Usage**

Given the static nature of base funding for GIS at most government agencies, the ability to answer the following questions on-demand is of great importance (Tomlinson 2003).

1. What data is needed?
2. What hardware is required?
3. What and how much software is needed?
4. Where will the hardware be positioned within the organization?

To be able to answer questions 2, 3, and 4 with any certainty the responsible GIS Manager will need to develop, or have access to, a system for tracking and measuring GIS usage. Though often seen as unglamorous, the tracking and measuring of usage through time is a key component in successfully managing a GIS environment.

It is necessary to track both the number of GIS installations within an agency as well as to develop ways to measure the actual usage of the technology (e.g., one of our Regions installs ArcGIS on all new computers prior to delivering them to the end user thus some installations may never be used). If a concurrent licensing model is used, measurement of license usage can be logged with some sort of license management software, such as FLEXnet Manager (Macrovision 2007), or a custom solution using system event logs. If you are using a single use or node locked license model you will have to develop other ways to track usage, for example by deploying a VBA script for ESRI’s ArcMap that would log the startup of the software to a database. If you are using web mapping solution you would want to capture the identity or number of unique users who are accessing your site. In either case, it is imperative that this information be collected and analyzed on a scheduled basis to identify usage trends and, if necessary, for billing purposes.

**Usage Tracking at WSDOT**

Currently WSDOT is using a custom solution to obtain the information required to calculate usage trends. Our methodology revolves around the use of a batch job that has been scheduled to run on our ESRI concurrent use license server. The batch job executes the FLEXlm lmstat command, used to monitor network licensing activities, to obtain the list of currently checked out licenses (Globetrotter 2004). The results of each
run are appended to a log file. The log file is copied each month to an archive and the original log file truncated as needed. The frequency at which this job runs is user defined, currently we run the process every thirty minutes. The reoccurrence intervals of this process can be set to any interval; however periods of one minute or less would be difficult to implement since you would start running into situations where the new job was starting prior to the completion of the previous job. Potentially impacting both the performance of the server and causing issues with the license manager itself. If real time usage statistics are required, vs. 5 minute, 30 minute, etc. snapshots you would need to use a purpose built usage tracking system.

The code shown in Table 1 is what is run at WSDOT to monitor our license activity. At this time we store the user id, machine name, and the license check out date/time. Two versions of the code are running at this time, one for tracking the usage of ArcInfo and the other for ArcView. When ArcEditor is added to our software suite in 2007 a third job will be scheduled. Note that this same code may be modified to track usage of ArcGIS extensions such as 3D Analyst and Spatial Analyst.

Table 1. Batch job code for tracking ArcInfo usage using the license manager statistics (lmstat) command provided with the ESRI license manager.

| REM ESRI License Usage Utility for Arc/Info |
| REM Jay W. Callar |
| REM Office of Information Technology |
| REM Washington State Department of Transportation |
| REM Olympia, WA 98504-7430 |
| REM This bat file is scheduled to run every 30 minutes on the license server, but could be scheduled more or less often. |
| REM This version of the batch job is setup for Arc/Info, to run for ArcEditor or ArcView you would modify the lmstat command parameters (e.g., lmstat -f viewer). |
| REM Create variable "count" and set it to zero. Note the /a specifies a numeric expression to be evaluated. |
| REM Run the lmutil command which is part of FLEXLM. The results of the lmutil command is a list of all the currently checked out licenses..tokens=1,2,8* capture the 1st, 2nd and the 8th variable. The 8* means take all variables after the 8th and combine them into one. |
| REM The first and second values are the userid and machine name, the 8th value is the start date and time when the license was checked out. |
| REM For every current user of ArcGIS we will run the following code. |

::LOOP
REM For every license that has been checked out we call the LOOP.
for /f "tokens=1,2,8*" %%i in ('lmutil lmstat -f arc/info -c "\myServerName\c$\program files\esri\license\arcgis9x\37104232.lic" ^| findstr MyLicenseServer') do call :LOOP %i %j %k %l goto :EOF
REM first increment the count variable.
set /a count=%count%+1

REM Create new variable.
set result=

REM The following for loop checks to see if the license has been
REM checked out since the last time the script ran.
REM If the license is new then it won't exist in the arcinfo.log file
REM and needs to be added.
for /f "tokens=1" %%x in ('findstr /M /C:"%1 %2 %3 %4 %5 %6 %7 %8"
arcinfo.log') do set result=%%x
if NOT defined result echo %1 %2 %3 %4 %5 %6 %7 %8 %count% >>
arcinfo.log
goto :EOF
:EOF
REM Exit Job

The log files generated by these batch jobs are suitable for use as input to most
statistical analysis packages (e.g., SAS, SPSS) and common desktop tools, such as
Microsoft Excel. If interest warranted the additional programming time, it would be
relatively straightforward to develop a data transfer and load process to import the log files
into a SQL database, thus exposing the information to a wide range of data reporting and
manipulation tools.

Data Sampling Intervals

Once the usage data is collected individual, group, or agency behavior can be
examined. When conducting these analyses the level at which you bin or aggregate the
data is of importance. Common aggregation levels used for this type of data are hourly,
daily, weekly, monthly, or annually. The selection of the most appropriate level to use is
driven by the purpose of your analysis. In our case we are trying to use this data for trend
analysis to model current and future GIS usage. As such extremely short aggregation
levels (hourly or daily) or long periods (annual) are unsuitable for our purposes (see
Appendix for additional details). Both the weekly and monthly aggregation level may be
used for forecasting purposes. However the weekly data would need to be normalized by
the number of work days per week to account for holidays. To reduce computation
overhead and to simplify explaining the data to management we have chosen to use the
monthly aggregation level in our forecasting.

Usage Mapping & Analysis

Several complex statistical methods for analyzing usage data are available (e.g.,
T-Test, ANOVA, etc.); however experience has shown that simple correlation analysis
and graphing methods provide sufficient information for the job. In addition, these
methods produce results that are understandable by a large audience –when meeting with
agency executives you do not want to be giving a statistics lecture when you should be
discussing funding requirements.

In our case we have focused on the generation of charts and maps that show the
variation in usage over time and attempt to relate obvious inflection points in the data
record with known events (Figure 1). For example, we upgraded all personal computers
in our agency from Windows NT to Windows XP in October-November 2005; simultaneously we converted from ArcGIS 9.0 to 9.1. Not surprisingly a measurable reduction in the number of unique GIS users occurred during this same period. It is important that you are able to identify and explain these inflection points to ensure your statistics are seen as credible by management.

![Figure 1. Number of unique users of ArcGIS Desktop per month.](image)

Figure 1 shows a high level view of GIS usage within WSDOT and helps explain large variations in month to month usage that occurred due to changes in the agencies information technology infrastructure. By tracking unique users, in addition to total usage, we have obtained a better understanding of the number of individuals using the technology within our agency.

Another way to look at such data is geographically. Since we know each users name and their machine we can group the users by their assigned office or Region (Figure 2). This method can be useful in identifying Regions that need additional assistance in implementing GIS. In our case we found that an assumption that we had been operating under for several years was incorrect. We had assumed that the rural Regions, or Regions with fewer employees, would have lower GIS usage. However, when usage was mapped as a percentage of total employees assigned to each area we found the rural/suburban Regions actually had higher usage rates than the Olympic or Northwest Region –Regions containing large urban centers (i.e., Seattle and Tacoma).
Another way to look at our monthly data is with bar charts (Figure 3). These simple charts enable you to quickly identify those offices or Regions who are using the most GIS resources. Though academically interesting, the real power of this type of graph is that it provides a strong basis for requesting funding and support from the high-usage offices when major technology upgrades are needed, usually every two to four years (McCormack 2004). In addition they help identify user groups that the central GIS support group may not be aware of. For example in Figure 3 the relatively large size of the headquarters ‘other’ group indicates the presence of an underserved group or office that has not been recognized in the past.
Figure 3. Unique GIS users by Office or Region for the month of August 2006.

The proceeding three Figures provided examples of common products that can be created to illustrate the usage of GIS technology within your agency. The graph shown for Figure 1 provides a good way to look at usage over the entire period of record or over a one or two year time spans. Figures 2 and 3 demonstrate how data can be decomposed to obtain additional information regarding who your users are, what software they are using, and where they are located. Note that any of these Figures could have been done for a specific ESRI product (ArcInfo, ArcEditor, or ArcView) or extension (3D Analyst, Spatial Analyst, etc.); in this case the Figures show total ArcGIS Desktop usage.

**GIS Technology and the S-Curve**

GIS is a type of information technology. As such its acceptance and growth within an agency is influenced by many of the same issues that affect other software technologies. Specifically, when first introduced a technology must meet an unfulfilled need to be retained within an organization. If it does, the technology is accepted and its usage will slowly increase as users learn the technology. As the technology matures the user experience becomes of greater importance (i.e., the GUI). Rapid growth of the technology can be expected if the ‘next’ release of the technology continues to meet the ever growing expectations of the users. At some point in time a transition will occur were the technology satisfies basic user requirements and further increases in the functionality or complexity of the technology is of little interest to most users. At this point the technology would be seen as mature and usage increases would begin to level off (Newman 1998). This process can be represented mathematically by an S-curve or a cumulative normal distribution as shown in Figure 4. The only way for the technology to expand its user base once it is mature is for a technology paradigm shift to be introduced into the equation (e.g., ESRI’s introduction of ArcGIS Server) or for a fundamental change in how the technology is used in the organization to occur (e.g., GIS starts to be used to maintain the inventory of highway features -not just to map them).
The inevitable decline of a technology (e.g., replacement of ArcView 3.x with ArcGIS) occurs once the technology has reached maturity and can no longer offer new features or support for new devices. Two common types of decline are seen, a slow decline that occurs when a still functional technology is replace by a newer but complimentary system (e.g., phone → wireless phone → cell phone → camera phone) or a rapid decline. Rapid declines are often brought about by business decisions or regulatory changes. Examples include ESRI’s decision in the late 1990’s to stop supporting UNIX as a desktop operating system, Y2K (Farwell 1999), or the upcoming discontinuation of the analog TV signal (Futch et al. 2001).

Being aware of the S-curve and its relationship to the acceptance and growth of GIS within your agency is important. By examining your usage information (e.g., Figure 1) and comparing it to an ideal S-curve, you can estimate where on the curve your organization falls. Knowing this will help your organization select the most appropriate growth scenario (low, moderate, or high) when forecasting future hardware and software requirements.

**Forecasting**

The capability to forecast software and hardware needs is important for the success of GIS within any agency, department, or corporation. This ability is of even more importance for government organizations since unanticipated software or hardware needs may not be addressable till the next budget cycle -in the case of Washington State, with its biennial budget cycle, that could be as far as two years out.

There are several available methods for forecasting future GIS usage based on historic usage data. Chief among these are linear, nonlinear, and weighted least square regression and the newer locally weighted polynomial regression (LOESS) method – which equates to a weighted moving average when using a zero degree polynomial (NIST 2006). Each of these methods requires the availability of historical data to derive a solution.

A good rule of thumb is that you should not project forward in time longer than your period of record. In addition, given the technology lifecycle for vendor software of
18-36 months (Tomlinson 2003), forecasting usage of a particular technology more than two years into the future will have little value. Thus you should be able to make useful projections with as little as one year of record and be able to begin seeing trends in your data with as little as six months of record.

The simplest of the forecasting methods to use and explain is the linear regression method. With this method we can model the existing data with a linear equation and then use that equation to projection into the future. If you are aware of new user groups that will be brought on-board within your planning horizon you will need to adjust your forecast accordingly. Notwithstanding, starting from a quantifiable forecasting method and modifying it based on known future changes in your environment is a good place to start when a SWAG\(^1\) is necessary.

![Figure 5. Comparison of several non-linear regression equations against a standard S-Curve.](image)

Other nonlinear forecasting models are available; these include logarithmic, polynomial, power, or exponential equations (Figure 5). Similar to the linear method, a nonlinear equation can be derived for these curves for forecasting purposes. The exponential and polynomial methods are of special interest since they are capable of replicating portions of the S-Curve shown in Figure 4. If you suspect that the technology you are tracking is ‘mature’ in your organization, and you have a long period of record, you should consider using a 2\(^{nd}\) or 4\(^{th}\) order polynomial in one of your forecasting scenarios since they allow for both increasing and, upon reaching an inflection point, decreasing usage over time -as well as changes in the rate of growth.
**Growth Scenarios**

Since WSDOT is on the increasing side of the S-curve we initially chose to use linear and exponential models for deriving our growth scenarios. In 2007 we began using a polynomial method for our high growth scenario since we are now further along the S-curve and need to begin looking for an inflection point in our growth rate. In addition, in some cases we have chosen to derive linear projections based only on the last twelve months of record to avoid issues related with the starting tail of the technology S-Curve.

![Figure 6](image)

**Figure 6.** An example showing low (blue), moderate (red), and high (black) growth scenarios forecasted out to December 2006. Scenarios were created using 2002-2005 data. Actual usage levels for 2006 are shown in purple.

Depending on the statistical sophistication of your audience you should develop several growth scenarios for forecasting and then select the most likely scenario based on your agencies situation. At WSDOT we generally develop a low, moderate, and high growth scenario to estimating the number of unique individuals we should plan to provide service for. Figure 6 shows three growth scenarios derived for WSDOT, these scenarios were obtained using linear, exponential, and polynomial methods, respectively.

In addition to the growth scenarios, we also derive license requirements based on past usage patterns. To do this we calculate the average and high license usage level; where usage is expressed as a percentage of unique users logged on at one time. For example, in August of 2005 we had 230 unique users login during the month. On average 7% of these individuals were logged on simultaneously, the maximum number of users logged on at any one time was 35, or 15% -note at WSDOT our percentages are based on the average of the averages and the maximum of the maximums over the entire year, resulting in higher ‘average’ and ‘high’ planning values. The planning values are important as they allow you to convert the forecasted number of unique users to the
number of licenses required. For example, to serve the maximum usage for the month of August 2005 we needed 35 ArcView concurrent use licenses—vs. the 68 owned by WSDOT at the time.

Based on the growth scenarios and planning usage levels selected, a matrix of possible GIS growth scenarios should be built (Table 2). From this matrix the most likely forecast can be select and used to justify additional software acquisitions and for planning purposes. The decision to plan for the average or high usage level when determining future license requirements is dependent on your organizations tolerance for ‘license unavailable’ messages and your available funding. In our case we selected the moderate growth scenario under the high usage level as our planning horizon. Based on that scenario we determined that we would need a minimum of 49 ArcView concurrent use licenses on-hand in 2006. Since we already had 68 licenses available we were able to delay the purchase of additional licenses until 2007.

**Table 2. Example of a scenario matrix for WSDOT showing the estimated number of ArcGIS Desktop unique users and future license requirements for December 2006 -based on usage data through December 2005.**

<table>
<thead>
<tr>
<th></th>
<th>Low (Linear)</th>
<th>Moderate (Exponential)</th>
<th>High (4th Order Polynomial)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of Unique Users</strong></td>
<td>261 Users</td>
<td>325 Users</td>
<td>658 Users</td>
</tr>
<tr>
<td><strong>Planning Usage Levels</strong></td>
<td>Average 7%</td>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>High 15 %</td>
<td>40</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>99</td>
</tr>
</tbody>
</table>

WSDOT has been using the forecasting methods described here for the last several years and have found the one year forecasts to be highly accurate. Our 2005 and 2006 moderate growth projections were found to agree within 10% of actual ArcGIS Desktop usage rates.

Figure 7 provides an example of the current (i.e., 2007) growth curves being used by WSDOT. The three curves, linear, exponential, and 4th order polynomial provide our low, moderate, and high growth scenarios respectively. Actual growth rates for first quarter 2007 are shown in purple on Figure 7 and are closely tracking the moderate growth scenario.
Summary

If your organization maintains and collects GIS usage statistics you will be able to estimate future growth trends with as little as twelve months of data. The statistical methods described here produce results that are understood by most agency executives, thus maximizing the amount of time available for discussing GIS strategy and funding issues when meeting with them. By projecting and planning for growth in this way you will be able to support and encourage the use of GIS within your organization. The failure to keep appropriate records may result in GIS Management being blind sided by growth and being forced to limiting the addition of new users to the existing system—a position that we do not want to be in since we are supposed to be championing the technology in our organization.

One new challenge that is facing many GIS Managers is the increased penetration of GIS technology into the software market. This has resulted in GIS technology being intergraded into many web based business solutions. As a result, many users of these applications are unaware that GIS is a key component of them. This may make it difficult to obtain support for continued or increased funding of GIS within an agency—unless statistics similar to those described here are presented to your agencies decision makers on a regular basis.

1SWAG – Scientific Wild Ass Guess. A term used by technical teams when establishing high level estimates for large projects.
Appendix – Aggregation Levels

The selection of the appropriate aggregation level to use when modeling and conducting statistical analysis is driven by the goal of the analysis. In our case we are trying to develop simple models to forecast future usage of a software technology (i.e., GIS). The aggregation levels available to you will be partly data driven and partly subjective. In our case we combined our 30 minute data snapshots into recognizable groups for analysis -hourly, daily, weekly, monthly, and annually.

Mapping the hourly or daily usage of a technology is of interest. However the growth parameter that we are looking for would be swamped by unrelated processes, such as employee behavior and the work schedules maintained at the given work site. Figures A-1 and A-2 demonstrate this for the hourly and daily aggregation levels.

One common use of hourly data is to look for changes in employee behavior that may impact your environment. For example if you are working in a license limited situation employees may try to work around the ‘license unavailable’ problem by logging in when they arrive at work and holding onto the licenses over the lunch break. If you don’t see a dip in usage (Figure A-1) over the lunch hour, your organization may be experiencing this problem. When discovered this mentality should be discouraged since it artificially inflates the number of licenses required by your organization.

Figure A-1. Hourly usage levels of a software technology. Note the lunch break occurring around the noon hour and the fall off in usage as the end of the work day approaches.
The weekly aggregation level is the first level where we can begin focusing on and identifying usage trends of a technology. However to be useful for forecasting purposes you will need to normalize the weekly information based on the number of work days per week. For example during the 4th of July, Thanksgiving, Christmas, and New Years weeks you should expect a measurable drop in usage since these are high vacation periods for most employees. In this example (Figure A-3) the raw number of unique users is shown in purple and normalized user levels, based on the number of work days in each week, are in blue. In this example the drop off in usage during the Columbus and Thanksgiving vacations in week 42 and 47, respectively, are plainly visible.
The monthly aggregation level is the most intuitive level to conduct trend analysis with. The monthly level is a large enough aggregation unit to minimize the impacts of variations in employee daily and weekly work schedules. This level of aggregation will still pick up major changes in workload (e.g., the annual beginning of the State Legislative session and Christmas break) but these variations do not swamp the signal that we are trying to model—the growth of GIS usage in the agency. It is for this reason that WSDOT has chosen to utilize the monthly aggregation level for its forecasting.

Note that the annual aggregation level, though of interest for historical comparison purposes, is of limited use for forecasting. The 18-36 month technology lifecycle of vendor software limits the number of available data points to only two to four events, not enough for most budgeting and planning purposes (Tomlinson 2003, McCormack 2004).

Disclaimer
This report was prepared as an account of work sponsored by an agency of the State of Washington. Neither the Washington State Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the State of Washington or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the State of Washington or any agency thereof.
References


GIS Support Team. 2006. GIS History & Milestones at WSDOT. White Paper, GIS Support Team, Washington State Department of Transportation, Olympia, WA.


Newman, David. 1998. The invisible computer: why good products can fail, the personal computer is so complex, and information appliances are the solution. MIT Press, Cambridge, MA.


Author Information

Richard C. Daniels, GISP
GIS Product Support
Office of Information Technology
Washington State Department of Transportation
P.O. Box 47430
Olympia, WA 98504-7430
Phone: (360) 705-7654
FAX: (360) 705-6817
danielri@wsdot.wa.gov

Jay W. Callar
GIS Server Administrator
Office of Information Technology
Washington State Department of Transportation
P.O. Box 47430
Phone: (360) 705-7675
FAX: (360) 705-6817
callarj@wsdot.wa.gov