

BIOGRAPHICAL INFORMATION

Dan Shannon
Engineering Manager
TELUS Communications Inc.

Specific Responsibilities

Joined BC Tel in 1981 as a Drafting & GIS Technician. Dan is currently responsible for TELUS' Outside Plant design team in the Central and North Okanagan as well as the Thompson / Nicola regions of British Columbia's Interior South Region.

Past Experience

Dan has been involved with Telecom GIS since 1988 as a user, programmer and GIS Manager. Additionally he has field and management experience in survey, GPS and Conduit Engineering. Dan was a member of the Core Project Team for BC TEL during their migration from their Synercom based AM/FM system to Verizon's ICGS platform. More Recently Dan was the Project Manager for TELUS' Design On Line Project which included migration and conversion components, consolidating their ICGS data on to IMAGE, TELUS' FRAMME implementation.

Educational Information

Drafting Technology Certificate – College of New Caledonia (1980)
Certificate in Front Line Supervision – BC Open University (1996)
Knowledge Management – Royal Roads University – Current

Professional Memberships

GITA – Board Member

KEY INDICATORS IN GIS PROJECT DEVELOPMENT AND EVALUATION

Dan Shannon
TELUS Communications Inc.
2002 Enterprise Way
Kelowna, BC, V1Y 9S9

ABSTRACT

Whether it's business case development, RFPs, bids or contracts there are two key components to quantitative analysis: Getting the right numbers, and getting the numbers right. This presentation will provide methods for developing the right figures, and will suggest practical approaches that can be used by Project Managers and Business Analysts for communicating numbers in a meaningful way to customers, senior management and other decision makers.

ESTIMATING

Estimating. We all do it. Not just Project Management, but operations groups, IT departments and vendors.

We do it for project planning, budgeting and workload forecasting.

Estimating goes to the core of the big questions we ask in project planning and business case development:

?? What will happen if we do this; if we go down this path?

?? What will it cost?

?? How long will it take?

Estimating goes to the core of projects simply because we're looking into the future, and we can't measure data we don't have, so it has to be estimated.

So it seems obvious: We estimate. But if this is an obvious truth, it's one that often gets lost in the course of a project. Possibly it's due to continued focus on the next step, the next milestone, or the next deadline. Perhaps it's a reluctance to rebuild a project plan around new estimates. Whatever the reason, initial estimates are often left un-refined.

Now, even if everybody gets through the project okay, vendors are paid, the work gets done, we still face the infamous PIR. Now, certainly in years past the PIR was a bit more like planning for an earthquake. Obligations were acknowledged, but the rigor and actual execution of a PIR remained a rare occurrence. But to our great delight we are in a world of scarce capital and increased fiscal scrutiny. If for no other reason than you may need money again someday for your next "How I'll Save the Company with GIT" project, you'll want to make sure estimates are as accurate as possible as early as possible in the '*Arc of Executive Awareness*'.

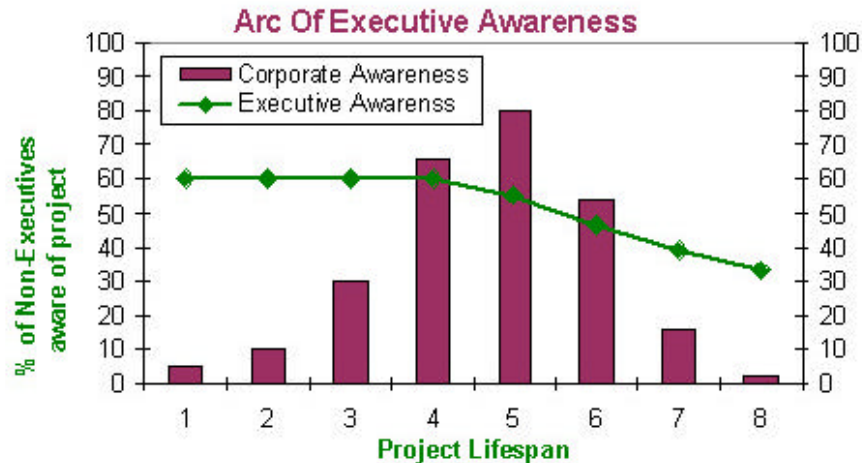


Chart 1 – Arc of Executive Awareness

This phrase simply means that as a project gathers momentum, there is a brief period where it is above the threshold of Executive Awareness, when a VP is consciously aware of a project, and somehow senses its importance to him personally. That is, it could affect his variable pay. The physics of dealing with executive hearing dictate that the first numbers quoted are remembered forever while qualifying words like ‘preliminary’ or ‘tentative’ simply pass silently through the air, unheard. Project Sponsors and therefore you, the Project Manager, will be held to these expectations.

Refining Estimates

So it’s important to get the best possible estimate right up front. And there are some techniques that will help you do that. But no matter how brilliant, or cleverly conceived, don’t fall in love with your first Microsoft Masterpiece. Early assumptions must be verified. Limited sample sets collected ahead of a deadline need to be ‘fleshed out’ and ‘early returns’ must be factored in.

You’re going to be faced with the uncomfortable task of amending parts of your project plan when data upon which your plan, business case, RFP, or bid depends. You have one viable option: Take your medicine. Make the changes and present them to your VP, your steering committee, and your Project Sponsor. Follow your change control process. All manner of unpleasant consequences will follow, all of which are preferable to the alternative. Perhaps the most appealing temptation to be resisted is letting your contingency absorb the discrepancies.

More surprises will be coming, and when was the last time any of us got a call informing us that things are going much better than planned? Data has turned out to be cleaner than expected? Certainly, some of these kinds of calls come in from time to time, but when you think back to your last project what is the overwhelming feeling that washes over you? Giddy joy? Calm serenity? No, I’m betting it’s something more along the lines of ‘Roller Coaster Nausea’, shortness of breath or an alarming tightness in the chest (ah, good times). So keep your contingency intact.

Furthermore, working to adjusted estimates helps to minimize problems downstream that would more likely grow into problems that even more relentlessly eat into contingencies.

Source Data - What to Measure

Estimating for a business case, an RFP or for work planning? Ideally, the units of measure would be identical. However, it is unlikely that the key measures will be consistent across the lifespan of a project.

In the early stages of a project, the focus is likely to be on more general figures, as a broad assessment of the work to be done influences early broad gauge estimates. As more accurate estimates are required, more detailed analysis begins, as does the search for more specific elements to be used in estimating.

Once vendors are engaged, the matter is complicated by the very likely chance that the units of measure the vendors prefer to use to estimate billing, work schedules and volumes will be different from the ones used to this point by the customer's project management team. Reconciling the two may be an easy adjustment, or may present a more difficult roadblock to be overcome before proceeding to the next key milestones.

There is more likely to be an impetus to change 'data lanes', so to speak, encountered earlier on than vendor engagement.

The data you have that is complete is unlikely to be the data you need. Further, some of the best indicators for estimating can be an attribute or item for which there is not a full data set. Fortunately partial data can usually be estimated quite well. And, it is important that you do, because you will be producing further estimates based on data this data; data that's already been estimated.

The good news is most corporations have some component of your network that's considered quite important. Therefore it is tracked, or at least enumerated, pretty well. It's quite likely the data may reside outside of your GIS, but that might not be a big hindrance. There will likely be some kind of attribute in your spreadsheet that will help distinguish items, allowing them to be broken into groups for estimating, whether it is a field in your databases, or a column in your manual ledgers.

Let's take a look at some of the possible kinds of data that may be available for quantifying work.

Maps or Plats

Maps or plats are usually pretty easy to quantify. There'll be an index or listing available. They are almost always broken out by serving area, or into some geographic region. They are also pretty certain to be part of the delivery package, so they have immediate relevance to both the client and the vendors as a meaningful unit of measure.

There are some key limitations however. While maps give a pretty good estimate of the amount of geography covered, they don't tell you the amount of work involved. Plant density varies, as do map scales. Sampling from maps provides an estimate of facilities, but it is clearly unwise to assume there is a direct relation across all maps or regions. Still, the physical plat will yield some good information. Some categorization will help. And there are some characteristics of the maps that will prove helpful. Categorization based on terrain or latitude, for example, while not an obvious first consideration, may help distinguish work sets by specific geography and climate, which in turn can have a real impact on the provisioning methods used, and thus the mix of network components. Another categorization could be the operating regions that originally created the records. Groupings like this will keep regional anomalies grouped together, anomalies that can have a real impact on project difficulty and approach not only in data preparation and later during quality assurance, but also in how vendors approach work assignment.

Work Grouping

Valuable estimates need to be developed at the work unit, or package level. It's great if you have good quantity estimates for your total project, but that will not be much help in determining work load planning when splitting work out across work teams or in scheduling discreet milestones. Estimates at a total project level are no help in gauging incremental progress. There will be no way of verifying if those quantity estimates were correct until all the work is back.

One example of this would be utility poles. Let's say we know how many poles our company has, and we have determined poles to be a key indicator for project work volumes. If there is no pole count for each work package sent away, how do you know how big each package is? How do you know if you've got half your data back if you don't know how many you've sent out? Certainly you'll get a pole count as you get the data back, but how would you know if that count matched what you'd sent away? Well, of course you wouldn't. Consequently, work will most commonly be categorized and 'packaged' by serving areas, by municipality or tax district, or possibly by route. So you'll need work volume estimates at these levels, which match, or more likely, determine, what constitutes a work package. And, it's pretty likely you'll have some key field in your data that determines where at least one of your key indicators fall within these work packages: For example, a ledger of transformers may have a column which tracks the municipality or the serving area.

Categorization

Categorization will help keep estimates from varying too wildly across work groups. Most utility companies will find that the type and amount of facility used will vary in some predictable ways across different serving areas. By putting work groups into some identified categories, estimates can be adjusted for each category.

Table 1 shows an example of inferred information in terms of conduit items. Using actual figures from converted data, estimates are built into the ‘Projected Conduit’ column, in this case used for feature count estimates for scheduling and billing. Categorization will help keep estimates from varying too wildly across work groups. Most utility companies will find that the type and amount of facility used will vary in some predictable ways across different serving areas.

WC	STATUS	GIS	CUSTOMER LINES	FACETS SIZE	Projected Conduit	Current Conduit
ABFD	GIS	TRUE	58025	589 > 5000	66779	4452
AGSZ	GIS	TRUE	4301	153 1000 to 5000	324	324
AKLK	GIS	TRUE	204	400 to 500	8	8
ALBA	GIS	TRUE	934	9500 to 1000	6	6
ALBN	GIS	TRUE	18668	95 > 5000	3326	3326
ALCK	GIS	TRUE	528	98 500 to 1000	17	17
ALPN	GIS	TRUE	39414	119 > 5000	2661	676
AMHT	GIS	TRUE	38027	176 > 5000	2567	1036
ARGV	GIS	TRUE	14977	507 > 5000	1034	1034
ARMS	GIS	TRUE	5041	161 > 5000	979	979
ASCF	IN PROGRESS	FALSE	1431	113 1000 to 5000	92	
ASPK	IN PROGRESS	FALSE	2333	148 1000 to 5000	150	
AVLA	GIS	TRUE	61	161 0 to 500	0	
AYNS	MANUAL	FALSE	622	16 500 to 1000	8	
BALF	IN PROGRESS	FALSE	1214	38 1000 to 5000	78	
BARR	IN PROGRESS	FALSE	1904	196 1000 to 5000	122	
BCLK	GIS	TRUE	1144	120 1000 to 5000	73	3

Table 1 – Sample of Partial and Categorized Data

By breaking work groups into categories based on numbers of customer lines, different estimating formulas can be applied. In the above case, for example, conduit turned out to be relatively more prevalent in the larger category offices, and so relationships between known group attributes and estimated attributes can be ‘fine tuned’ to be more suitable for the group in question.

If a linear relationship is established between a key element and other items being estimated then applying that relationship across dissimilar groups may give an overall result that works. But, as covered before, there are some very practical reasons for wanting those estimates to be somewhat accurate down to a work-unit, or ‘package’ level.

This doesn’t mean that completely different metrics need be developed, only that more appropriate constants be used in each category for the same formulas. In offices of a certain size, for example, a serving density can be determined by dividing by the number of maps. Table 2 shows 4 theoretical offices. More insight can be gleaned by looking at both number of customers and number of plats, than by evaluating based on any one alone.

Office	Plats	Customers	Density
North	100	500	5
East	100	5000	50
South	10	50	5
West	10	500	50

Table 2 – Example of Similarity Scoring

In this example, looking at plats alone would indicate similarity between the North and East Office. Looking solely at the number of customers the North office would seem be more akin to the West office. However, by using the two attributes together, and getting a density figure, the North office appears to be more similar to the South office. Formulas that fit the South office would also likely fit the North office best.

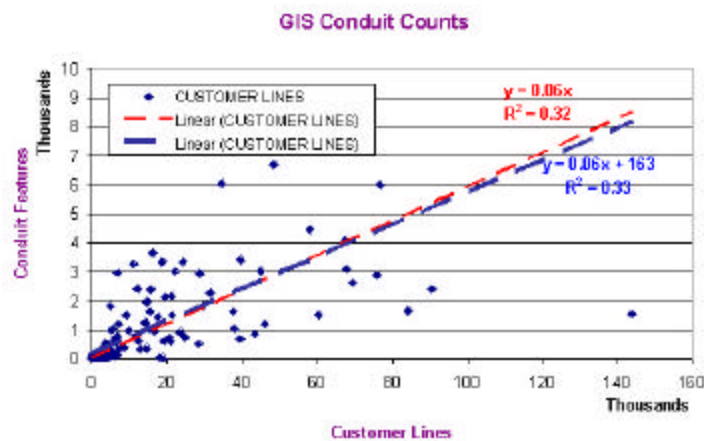


Chart 2 – Example of Low Correlation Data

Chart 2 provides an example of the full conduit data set shown in part in table 1. The first method of plotting the relationship of the Conduit Features (dependent) to the Customer Lines (independent) is to use the relationship expressed as the average number of customer lines per conduit element. The red dashed line in Chart 2 is plotted on this line. Expressed in a formula predicting the unknown (in this example, the unknown conduit elements) for work packages without the information available, the formula would be $y = x * 0.06$. The correlation on this data, or the R-value, is 0.32, or expressed informally “The number of conduit features is about 32% dependent, or explainable, by the number of corresponding customer lines in each work package.” And clearly, the ‘clouding’ of the data well off the intercept shows the limitations of using this independent data as a predictor of the dependent data.

The next way of adding some refinement to this formula is to incorporate an intercept. This is shown also on Chart 2, with the blue dashed line showing an intercept of 163. If a predictive formula was used working with this line, it would read $y = 0.06 * X + 163$ which yields a slightly better correlation, with an R-value of 0.33, or 33% influence.

We can begin to see a limitation. On smaller data sets we could end up with a prediction of a negative amount. Just as misleading, in this example, any work sets, no matter how

small, would calculate out to predict at least 163 conduit elements. Of course one could simply modify with a minimum value of 0, but that would be using a blunt instrument so to speak. And besides, and R-value of .33 isn't exactly going to inspire confidence in your result. So this data, while not to be disregarded, clearly isn't your best indicator, but may be more helpful in estimating a particular data group.

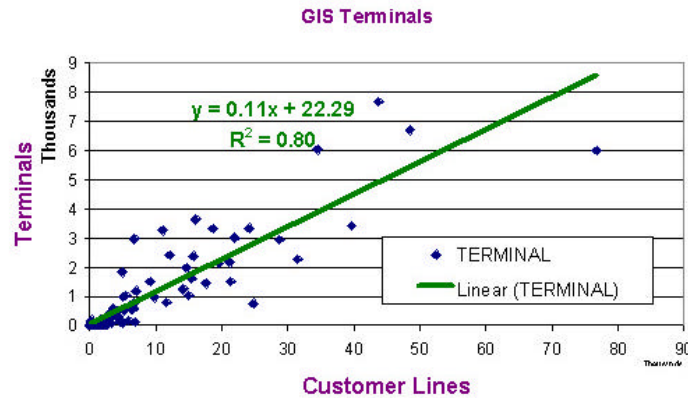


Chart 3 – A Better Predictive Relationship

Chart 3 shows data that seems to relate more closely. In other words, what we know should more accurately provide us with an estimate for what we don't know.

When you find a data type that fits better, where a tighter clustering shows a more consistent relationship, then you know you've got a better key indicator. As mentioned, this key indicator may not be of particular interest in and of itself, but if the R-value is high, it can be used to estimate missing data for an item that may be a key indicator.

More useful, however, is the use of multiple references. With no one data set being ideal, there is a much better result to be had by applying a slope/intersect formula to the dependent data from more than one well related set of independent data. Using the R-Values as relative weightings, the amount of influence each independent data set has upon the estimate can be balanced.

SIMILARITY SCORES

Similarity scores are a related but somewhat different way of looking at relationships with data. The concept comes from the world of Rotisserie League Baseball, or baseball pools. Bill James¹ began an exhaustive study of baseball statistics in the late 1970s and early 1980s. Baseball of course is a statistician's dream, with so much information recorded and kept, including figures on dozens of facets of a player's performance.

One of the things that Bill James looked for among all this raw data was a way to predict future performance based on past performance. He wanted to forecast. As we have been discussing here, one of the first challenges is to find the most meaningful data. Mr. James did find 'attributes' that had a strong correlation to future performance. For example, while a team's batting average has some correlation to the number of runs they

will score, but On Base Percentage multiplied by Slugging Percentage (OPP*SLG) is much more closely correlated to Runs scored.

He went a step further; enhancing the prediction of performance based on the statistical similarity of one player to others. Similarity scores compare a number of relevant statistics, and weight them based on importance. Once a similarity is established, one can see how well one players career matches, year over year, with another who is a good match.

So, how does this relate to estimating for GIS? Well, the data in our GIS systems are as complex and varied as the networks they describe. Therefore, no one single method or relationship can describe your data, or provide you with all the information you need.

Different data sets, if categorized well, can be broken out into groups that among the members of each group share a high similarity score. Once you determine the elements which best describe the data set, you can then build formulas that work well, or ‘fit’ most of the data sets in that grouping. In other words, you will get predictive or ‘independent’ data with a high R-value. There will be a high correlation to the dependent data you’re trying to estimate.

	B	C	D	E
2				
3	Region		Independent Data	
		Customers	Poles	Transformers
4	Pilot A	243	56	4.9
5	Pilot B	156	38	2.8
6	Pilot C	148	44	4.2
7	Pilot D	394	195	6.2
8	Target Data A	567	308	8.2
9	Target Data B	146	67	3.9
10	Correlation on Known Data		0.80	0.71
11	Slope		0.0111	0.0159
12	Intercept		1.9182	3.1996
E8:	$((C8 * C\$11 + C\$12) * C\$10) + ((D8 * D\$11 + D\$12) * D\$10) / (D\$10 + C\$10)$			

Table 3 – Example of Using Multiple Data Sets For Feature Estimates

In Table 3 there is an example of estimating transformer counts based on the correlation of two other ‘knowns’; Customers & Poles. The ‘weight’ each of the ‘knowns’ is given is based on the R-squared value. This is a simple example of applied similarity scores.

You won’t find data as well catalogued as for baseball players, but if you dig, and if you fill in partial data sets with careful inference, there will be enough different ‘independent’ attributes from which to work out the best estimates possible.

Show Your Work.

In the interest leaving an audit trail for yourself, and for those who need to work with and build upon your estimates, Show your work.

Oh, I don't mean actually try and *show* it to anyone. Two things you need to know:

- 1) No one actually wants to see your work
- 2) Be ready though, in case someone actually wants to see your work.

For the most part, you'll want to record how you got where you're going so that later you can remind yourself what adjustments you made and why. If results are outside of expectations, you will have a better feel for where you need to adjust.

And, if something in the project goes off-metric, you can bet the Manager of Blame Assignment will suddenly be overcome with an urge to wade into your project in great detail, for at least as long as it takes to be satisfied that you've done your calculations, and can defend them.

CHALLENGE THE LOGIC

And finally, when you're immersed in all the numbers, apply some common sense. A relationship between two sets of data doesn't necessarily mean there's a connection. An example of this comes from the years during the polio epidemics in North America during the Fifties.

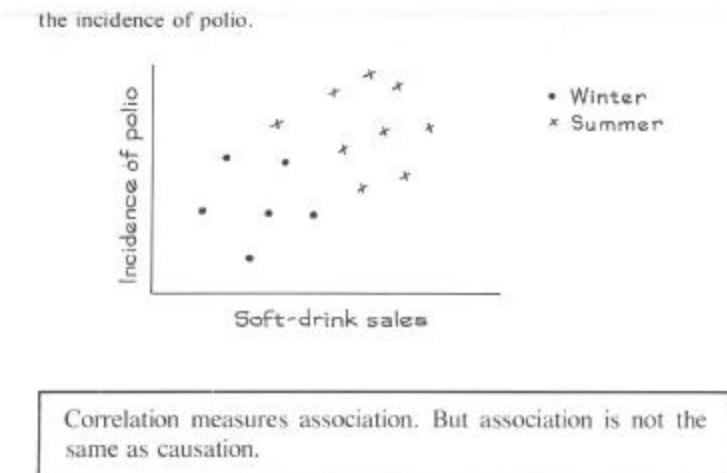


Chart 4 – Sample of Correlation Without Causation²

A very strong correlation was noted between soft drink sales and incidents of Polio when plotted by week. It was too strong a correlation to be a coincidence. But did soft drinks cause Polio? Of course not. Rather, both sets of data were dependent on the real independent data set, the seasons. Time of year played a role in influencing each of these two unrelated events.

Bibliography

1. Freedman, Pisani, Purves "Statistics", hardcover, 1978, W.W. Norton & Co.
2. James, Bill "[The Politics of Glory](#)", hardcover, 1994, Macmillan, 452 pages.