BAM Attachment 4 – Statistical Baseline Modeling

Introduction/Overview

The purpose of this project was to develop predictive models for estimating the presence and speed of xDSL coverage throughout the United States.

A multi-phase approach was taken. Each project phase was designed to inform later phases as well as prepare data at the appropriate scale for the final set of predictive models. Time constraints required that all relevant data be identified by the end of Phase I and be acquired by the end of Phase II. Models developed in Phases I and II guided the conceptual development of the final (Phase III) model.

During the first two Phases, data were identified, collected, assembled, checked for reasonableness and consistency, and statistically characterized. Essential preliminary computations were made, including finding all mutual intersections of wire center[1] polygons and Census blocks, deemed "fragments," and calculating geographic characteristics of these fragments (such as their areas).

Indications of xDSL availability at all speed thresholds were computed throughout Alabama and were obtained for Pennsylvania (at 1.5 Mbps nominal download speed) and Minnesota (at 768 Kbps nominal download speed).[2] Collectively these comprise the "estimation dataset" from which the model coefficients are ultimately estimated. One-half of the estimation data, selected by taking one-half the wire centers at random, were used to explore the data and to fit and assess the models, reserving the remaining estimation data (the "hold-out" data) to check model quality. Final models were fit using all the estimation data.

Model estimation proceeded in stages, closely paralleling the procedure advocated in Hosmer & Lemeshow (2000) chapters 4-5. After determining how best to express the variables (in many cases by using their logarithms), initial models were estimated at target speeds ranging from 384 Kbps to 6.0 Mbps using both forward and backward stepwise logistic regression. Data were weighted by the fragment area relative to the area of its parent block: in effect, this weighting conducts the analysis at the conceptual level of the Census block, while accommodating the complication that some blocks cross wire center boundaries. Due to the large amount of data available—the smallest data set used still had more than 30,000 fragments — stepwise regression rarely rejected a variable as insignificant. However, certain variables were collinear with others and other groups of variables (such as the distribution of non-family income) were so close to being collinear that including them created unstable models, as evidenced by widely varying coefficients from one speed to the next. Such groups were systematically dropped or coalesced into fewer variables until the models appeared stable.

[1] Wire center boundaries were obtained from TeleAtlas, September 2009

[2] Broadband availability datasets were examined from other states, including New York, Maine, California, and Wyoming. These were found not to be usable for this project due to differences in the quality of the data, the procedures used to collect them, their comparability, their lack of specificity (to xDSL alone), and their geographic scale.

The base model was augmented by introducing pairwise interactions, again using stepwise logistic regression. Almost all interactions with [hascable] were significant and had sizeable coefficients. This variable indicates cable coverage in a block[3]. Therefore, it was decided to create two families of models: one for blocks within the cable coverage areas and one for blocks outside the cable coverage areas. Both models use the same variables but are separately estimated. Each model is appropriate only for the kinds of blocks used for its estimation. Because the no-cable blocks will be closely associated with unserved and underserved areas, attention focused on optimizing the quality of the no-cable version of the model for each speed.

The result of these efforts is a set of twelve complementary models: a no-cable and with-cable model for each of six speed thresholds: 384K, 768K, 1.5M, 3.0M, 4.0M, and 6.0M.[4] Each pair of these models computes a single linear combination of the variables within each block fragment to produce a "logit." This is a number typically between -10 and 20. It can be interpreted as the strength of evidence for availability of xDSL within a fragment, with larger values corresponding to a greater likelihood of availability. Three additional steps were needed to produce the final results. First, for each speed a binary determination of xDSL availability—yes or no—is made by comparing the fragment logit to a *model-specific* threshold, which typically is a number between -0.1 and 1.0. Fragments with a logit exceeding the threshold are predicted to have xDSL available. Second, a Census block is considered to have xDSL available whenever xDSL is predicted in any of its fragments. . Third, to assure consistency among the model results at the various speeds, the predictions are adjusted where necessary so that prediction of xDSL at speed *x* implies prediction of xDSL availability at all speeds less than *x*, too.[5]

Prediction for the hold-out data indicated these models are typically 80% to 90% accurate within populated blocks. (Including unpopulated blocks would increase the apparent accuracy rate.) That is reasonable accuracy, but it will still produce estimation errors nationwide. Errors are of two types: false positives, where xDSL is predicted at a speed but is not available at that speed, and false negatives, where xDSL is predicted not to be available at a speed that is really available in a block. The importance of an error is proportional to the number of occupied housing units it affects, [hu_occ]. For example, a false positive in a block with ten occupied housing units does not fully compensate for a false negative in a block with 20 housing units: the latter is twice as important. The thresholds were chosen to make the errors exactly balance out on average: the numbers of *housing units* within the false positive blocks equal the numbers of housing units within the false negative blocks for each of the twelve models. These counts are based on Alabama data, which are common to all the models. Assuming that the proportions of errors of each type occurring in Alabama are typical of what will occur nationwide, this method of choosing thresholds gives predictions whose false positive errors should numerically balance its false negative errors when weighted by housing unit. Thus, statistical summaries of predicted xDSL availability expressed in terms of total housing units should be, at least to this first approximation, unbiased.

[3] This variable was derived from an analysis of MediaPrints September 2009.

[4] As 384K reflects a speed below minimum current reporting standards, these models were later dropped from the final results.

[5] This adjustment is based on the logits independently calculated by each model within each Census block. The final speed assigned to a block is the maximum speed for which all lower speed models indicate xDSL availability. For example, a block that shows availability at 6 Mbps, no availability at 4 Mbps, and availability at all speeds less than and equal to 3 Mbps would be assigned a speed of 3 Mbps. Such apparent inconsistencies, although rare, can occur because the models at each speed are estimated independently of one another.

The model coefficients and thresholds were developed on a combined GIS-statistical platform built on ArcGIS 9.3.1 and Stata 8.2 SE and then ported to a SQL Server platform for application to the national dataset. To accomplish the port without error, field definitions, model coefficients, and thresholds were tabulated to double precision in a computer-readable format and processed into a SQL stored procedure. The original Stata predictions were compared to the SQL Server predictions for 20% of the Alabama data to check that agreement was achieved within expected floating point roundoff error.

Dependent variables

Pennsylvania and Minnesota

A binary indicator of broadband availability within Pennsylvania was obtained from a statewide, Census block level analysis of broadband availability at a nominal download speed of 1.5 Mbps, classified by technology (xDSL, cable, *etc*.). A binary indicator of broadband availability within Minnesota was obtained from Connect Minnesota. The nominal speed threshold is 784 Kbps.[6]

Alabama

In Alabama, xDSL speeds were derived from calculations of wireline loop distances to the nearest DSLAM or xDSL-enabled Central Office (CO) located within the parent wire center. The distances are those along the network of roads and other likely wire loop rights of way. Conversion from loop lengths to download speeds was accomplished by interpolating published attenuation curves for ADSL-2 and VDSL technologies (Converge Network Digest, http://www.convergedigest.com/bp/bp1.asp?ID=15&ctgy=). See Table I, which also shows the interpolated distances for ADSL-2, ADSL-2+, and VDSL at speeds from 384 Kbps to 6.0 Mbps to demonstrate that the assumed technology has only a small effect on the speed-distance relationship.

The distances were then summarized to obtain the *smallest* distance found within each fragment within 25 meters of a road or right of way. Thus, the speed associated with a fragment corresponds to the *largest* speed that can be delivered within 25 meters of the roads within that fragment. (xDSL is therefore presumed unavailable within any fragment having no roads or rights of way.)

Table I Speed-distance conversion

	Speed, Mbps:	0.384	0.768	1.5	3	4	6
Meters	VDSL	7,038	5,990	4,979	3,932	3,497	**2,884**
	ADSL 2+	5,168	4,661	4,116	3,470	3,167	2,688
	ADSL 2	**5,475**	**4,923**	**4,323**	**3,598**	**3,250**	2,684
Kft	VDSL	23.089	19.654	16.335	12.899	11.473	**9.463**
	ADSL 2+	16.956	15.290	13.502	11.383	10.389	8.819
	ADSL 2	**17.963**	**16.151**	**14.184**	**11.805**	**10.664**	8.807

The bold figures indicate the distances used for each speed.

[6] Minnesota data were supplied from Connected Nation. Pennsylvania data were supplied by the Department of Community & Economic Development, Technology Investment Office, State of Pennsylvania

Independent Variables

An initial "data dictionary" of possible independent variables was constructed from a comprehensive list of variables found to be significantly associated with xDSL availability in eleven published, (generally peer-reviewed) papers. Most of these studies were performed with data obtained at geographic levels ranging from zip codes to entire countries, and therefore included many strong ecological correlations that may not apply to block-level data. Thus, it was expected that only some of these variables would eventually be useful, but that almost any useful variable would appear somewhere in this list.

We supplemented the data dictionary with additional variables, primarily measuring economic characteristics of wire centers, that could be associated with entry-to-market decisions made by DSL providers. Some variables, such as population and income, were summarized at coarser levels of geographic resolution (often the wire center) in order to produce some indicators of conditions within the wider spatial neighborhood.

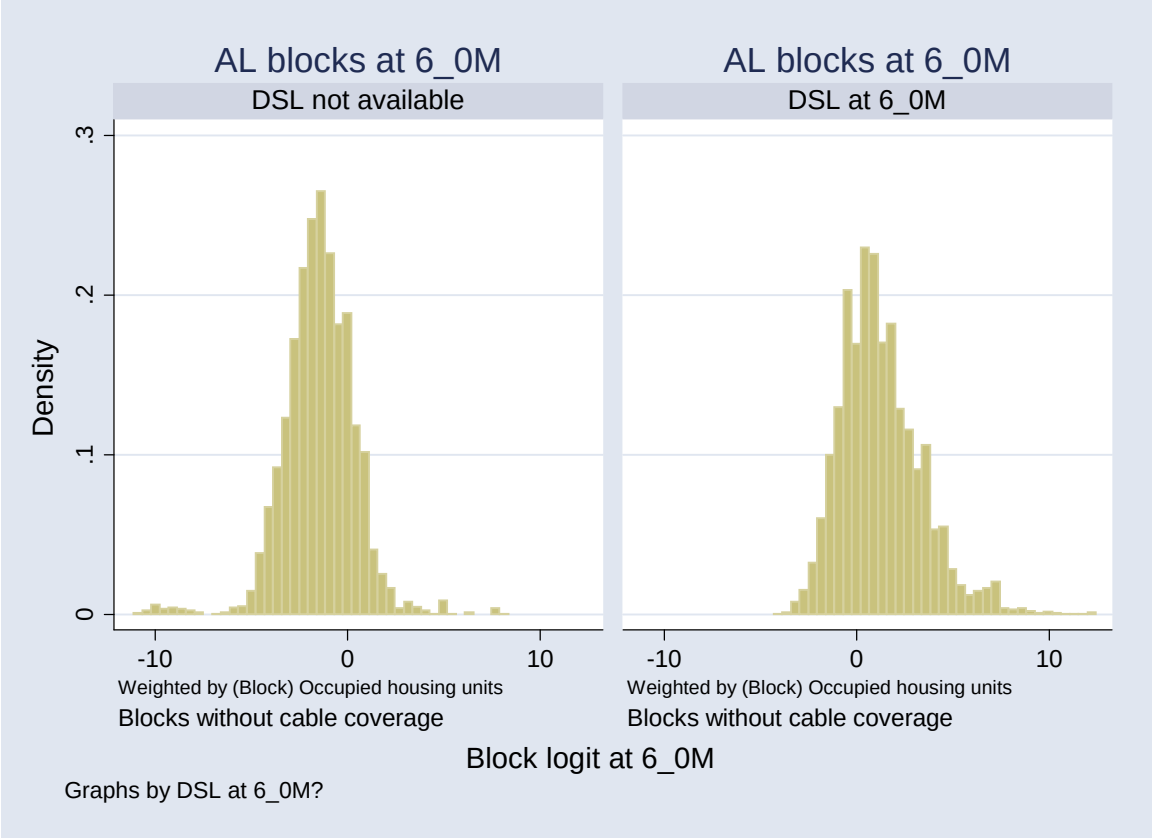
Results

Unlike many regression models, interest here lies in the accuracy of the predictions rather than the accuracy of the model coefficients or their interpretation. Prediction of a binary attribute like xDSL availability (at a specified speed) is *discrimination*: the model must divide all cases (here, Census blocks) into two categories. Perfect prediction occurs when no case is put into the wrong category.

As explained above, the model combines values of the independent variables to produce a numerical value (the logit) for each block. If the logits for all blocks truly with xDSL are numerically separated from the logits for all blocks truly without xDSL, then perfect discrimination is possible by comparing the logits to any value lying between the two groups of logits, a "discrimination threshold." Because the dependent variables, numerous as they are, are not perfectly correlated with xDSL presence, overlap occurs: the logits for some blocks with xDSL will overlap with typical values for blocks without xDSL, and *vice versa*.

Figure 1 illustrates a difficult case: the no-cable model for 6 Mbps service. This model was developed entirely from Alabama fragment data, which were the only ones with information for this speed. The histogram on the left summarizes the block logits for blocks without xDSL at 6 Mbps or better, *weighted by the number of occupied housing units per block*. It covers mostly negative logits, but its right tail—representing about 30,000 occupied households—extends into positive territory. The histogram on the right similarly summarizes the logits for blocks with xDSL, representing 133,000 occupied households. It covers mostly positive logits, but its left tail—about 33,000 occupied households—extends into negative territory. The overlap of the histograms near zero indicates perfect discrimination is not achieved. Evidently, about as many false positive errors are made (as counted in terms of occupied households) as false negative errors when the discrimination threshold is set near zero. The threshold used in this case was actually -0.062, causing approximately 31,500 households to be false positives and an equal number to be false negatives.

Figure 1 Performance of a logistic model



In this fashion discrimination thresholds were estimated for each model (by speed and cable availability). The error rates were typically better (*i.e.*, lower) than observed in this example, which exhibits only 80% accuracy overall.

Discussion

It is noteworthy that the model's discrimination accuracy is substantially higher, at 85%, when measured as the proportion of Census blocks correctly predicted (*i.e.*, when not weighted by numbers of occupied households). This comes about for many reasons relating to the fact that discrimination is most difficult at the boundaries of xDSL coverage areas where population densities are still relatively high (compared to rural areas). Households situated just outside a boundary are likely to share many characteristics of those just within the boundary. In such locations, only 50% accuracy can reasonably be expected. When spatial correlation is high, these bands of low-accuracy discrimination can be wide, significantly degrading predictive performance. This phenomenon is almost impossible to eliminate with a statistical model. The attempt to divide households by equal numbers into the false positive and false negative groups is motivated by the hope that such errors will tend to balance each other in follow-on summaries and analyses based on relatively large regions, such as entire states. Thus, although at the scale of the individual Census block these accuracies might be 80% to 90%, at the larger scale of a state or the nation we expect accuracy to be higher.

It may be tempting to look at other gauges of model fit, such as the "pseudo R-squared" values emitted by the statistical software. These are numbers between 0 and 100% that might be interpreted like the familiar squared correlation coefficient of ordinary regression, R-squared. Indeed, pseudo R-squared was useful in model selection—and it improved tremendously from Phase I through Phase III—but it indicates little about the accuracy of any particular prediction. The reason is that it mixes a number of factors, not all of which are relevant. For example, if a few unusual blocks (such as the unpopulated blocks) are included in the estimation dataset, they will greatly increase the pseudo R-squared but likely not add anything (and maybe take away something) from the model accuracy.[7] Another gauge is the standard error of prediction. So many records were available for estimation (at least one per block, which translates to tens or hundreds of thousands of records per model) that the standard errors of prediction would likely look small no matter what. Moreover, there is no way to combine these standard errors when "rolling up" the logits from the fragments to the blocks.

The plan to divide the data into estimation and prediction datasets worked well, but in the end it may not have captured enough of the potential for variation. At this point it is not known how xDSL availability varies geographically over areas greater than one state or within states with characteristics substantially different from those used for model estimation. It has also been impossible to identify a reliable relationship between xDSL availability and some potentially important explanatory variables, such as indicators of the financial deployment decisions of the carrier, because those variables did not exhibit sufficient variability within the estimation datasets. As more estimation data, consistent in production methods and vintage, are obtained it would be a useful next step to re-analyze and adjust the models based upon data that are more representative of conditions across the United States and its territories.

These uncertainties are mitigated by several factors. In particular, follow-on analyses are protected by the fact that this model is used to estimate the speeds only in a minority of blocks; namely, those where cable coverage is not independently indicated.

[7] This is related to the better known phenomenon with R-squared: if in any scatterplot a single additional point is placed beyond the end of the scatterplot, then R-squared can become close to 1.0 because the data look almost linear when the picture is zoomed out far enough to see both the original points and the new point.