PHOENIX
CENTER
FOR ADVANCED LEGAL & ECONOMIC PUBLIC POLICY STUDIES
www.phoenix-center.org

PHOENIX CENTER POLICY BULLETIN NO. 27

George S. Ford, PhD

March 2011

CHALLENGES IN USING THE NATIONAL BROADBAND MAP'S DATA

Abstract: As part of the American Recovery and Reinvestment Act of 2009, Congress set forth an ambitious goal of establishing a definitive National Broadband Map. Understandably, researchers are now anxious to use the data received to quantify the extent of broadband availability and to explain the relationships between availability and socio-economic factors. While the National Telecommunications and Information Administration should be lauded for its inaugural effort to complete this Herculean task, it is nonetheless important to recognize that the mapping data currently has many known and yet-to-be-discovered defects. These errors include, but are not limited to, measurement errors and sample selection, both of which can cause severe problems with empirical analyses. In light of the known defects in the data and the lack of a robust data verification process, the analyst must proceed with caution and considerable modesty. Indeed, until a data verification process is put into place and the known defects remedied, all statistical and econometric results using the National Broadband Map data should be viewed with skepticism as such work permits no strong causal and policy-relevant conclusions. Over time, as some of the measurement error and sample selection problems are resolved, econometric models may offer more robust conclusions. As always, empirical research using the Map data should begin with a sound, rational framework explaining observed outcomes.

PHOENIX CENTER FOR ADVANCED LEGAL & ECONOMIC PUBLIC POLICY STUDIES
5335 Wisconsin Avenue, NW, Suite 440
Washington, D.C. 20015
Tel: (+1) (202) 274-0235 ● Fax: (+1) (202) 318-4909
www.phoenix-center.org

I. Introduction

The American Recovery and Reinvestment Act of 2009 mandated the development and maintenance of a comprehensive nationwide inventory map of existing broadband service capability and availability in the United States.[1] The National Broadband Map ("NBM") is generated and managed by the National Telecommunications and Information Administration ("NTIA") with some assistance from the Federal Communications Commission ("FCC"). The difficult work of collecting the underlying data is handled by state-level organizations or their designees in voluntary collaboration with broadband service providers. Data collection procedures, computation algorithms, and provider participation vary by state and designee. The state data is forwarded to the NTIA for translation into the national map. This NBM was first published on February 17, 2011, meeting the statutory deadline.[2] Many states have operating state-level maps.

As observed by the NTIA, one primary use of the data is to "compare broadband availability among geographic areas and across demographic groups, which can inform policies to support private sector investments in deploying broadband."[3] For this task, it is important to keep in mind that the NBM is a work-in-progress with many known and yet-to-be-discovered defects. Some of these problems arise simply from the scale of the undertaking, which is massive and complex, and may be remedied over time. Others problems, however, may persist. The "Beta Release" nature of the NBM data and systematic problems with the data collection effort pose significant concerns for the use of this data in empirical analysis, though such problems may wane with time. Today, the data collection and generation process introduces measurement errors and, more severely, selection bias, which is systematic in nature. In this BULLETIN, I demonstrate that these problems may have severe consequences for econometric analysis, providing invalid estimates of casual relationships.

Also, based on the pattern of historical research using similar data, I discuss the shortcomings of commonly applied empirical techniques for public policy analysis. I also discuss the difficulty of using the NBM data to assess some more extreme interpretations of the

---

1 Pub. Law 111-5 (2009) (available at: http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=111_cong_bills&docid=f:h1enr.pdf). About $200 million in federal funding was allocated to the creation of the map.

2 http://www.broadbandmap.gov.

3 http://www.ntia.doc.gov/press/2011/NationalBroadbandMap_02172011.html. To facilitate this analysis, the underlying data supporting the map is also publicly-available (available at: http://www.broadbandmap.gov/data-download).

"Digital Divide."[4] My intent is to discuss and illustrate these problems without the heavy use of mathematics and theory so that the analysis is accessible to a wider audience. I rely primarily on a Monte Carlo analysis, a technique which is little more than a fancy numerical example. The topics covered include: (1) measurement error in the variables, including aggregation issues; (2) selection bias; (3) bivariate analysis and omitted variables; and (4) the assessment of demographic determinants of the "Digital Divide."

My recommendations are straightforward: First, the apparent defects in the first release of the data are so serious that, at this time, use of this data to inform important public policies, such as subsidized broadband deployment or assessment of gaps in availability among important groups, is not prudent. Second, prior to the use of the NBM data for policy-relevant research, the data should undergo a considerable validation effort by federal and state participants, as well as independent researchers. The presence and likely consequences of short-term and long-term defects in the data must be carefully studied, ideally providing some quantification of the likely magnitude of such problems. Hopefully, future releases of the NBM data will be more accurate and thus more useful. Third, any research conducted for public-policy purposes should apply valid statistical techniques driven by sensible underlying frameworks of analysis.

II. Background and Monte Carlo Simulation

In its first release, the NTIA claims that the NBM data indicates that about 5 to 10 percent of American households lack access to broadband "at speeds that support a basic set of applications, including downloading Web pages, photos and video, and using simple video conferencing."[5] A primary use of the NBM data will be, as observed by the NTIA, to compare and explain the differences in "broadband availability among geographic areas and across demographic groups."[6] Evaluating availability is but one potential use of the data. Past research efforts on broadband Internet service suggest the data will also be used to assess the effects of broadband (whether availability, speeds, competition, and other measures computed using the NBM data) on economic outcomes.[7] The goal of such analysis, of course, is to get a

---

[4] The term "Digital Divide" can mean different things to different people, but typically connotes differential access to or adoption of modern communications technologies, in particular high-speed Internet services, among various socio-economic groups.

[5] http://www.ntia.doc.gov/press/2011/NationalBroadbandMap_02172011.html.

[6] http://www.ntia.doc.gov/press/2011/NationalBroadbandMap_02172011.html.

[7] *See, e.g.*, S.G. Ford and G.S. Ford, *Internet Use and Depression Among the Elderly*, PHOENIX CENTER POLICY PAPER NO. 38 (October 2009) (available at: http://www.phoenix-center.org/pcpp/PCPP38Final.pdf); T. R. Beard, G.S. Ford and R.P. Saba, *Internet Use and Job Search*, PHOENIX CENTER POLICY PAPER NO. 39 (January 2010) (available at: http://www.phoenix-center.org/pcpp/PCPP39Final.pdf); S. Gillet, W. Lehr, M. Sirbu, *Measuring the Economic*

(Footnote Continued….)

reliable (or "unbiased" and "efficient") estimate of any such relationship, and to assign a causal meaning to such associations. The satisfaction of these goals requires more than the use of econometric methods; it requires the use of appropriate econometric methods.[8]

Following the NTIA lead, say that our primary interest is in the geographic and demographic determinants of availability. Let y be an outcome of interest (e.g., broadband availability). In general, we expect any relevant outcome, y, to be influenced by a number of factors x. For example, broadband availability, which is determined by the flow of profits expected from subscriptions, will be driven by supply- and demand-side factors such as income, population density, geographic terrain, and so forth. The researcher is interested in the expected response of some outcome, y, to changes in some set of relevant factors. Assuming there are five such factors (x_1, x_2, x_3, x_4, x_5) the multivariate regression equation used to estimate the relationships of interest is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + e, \qquad (1)$$

where e is the econometric disturbance term which accounts for the fact that the there are influences on the outcome y not accounted for by the five x's. With data on y and all the x's, it is possible to estimate the coefficients β_0, β_1, β_2, β_3, β_4, and β_5 using appropriate econometric methods. Note that the change in y due to a change in x_1 is β_1, which answers the question "how does y change given a change in x_1 with x_2, x_3, x_4, and x_5 held constant?" In a simple linear model, if $\beta_1 = 1.0$, then a one-unit change in x_1 leads to a one-unit change in y, other things constant.

Impact of broadband deployment, FINAL REPORT, NATIONAL TECHNICAL ASSISTANCE, TRAINING, RESEARCH AND EVALUATION PROJECT 99-07-13829, Economic Development Administration, US Department of Commerce (2006); L. Holt and M. Jamison, *Broadband and Contributions to Economic Growth: Lessons from the US Experience,* 33 TELECOMMUNICATIONS POLICY 575-581 (2009).

[8] There are many documented cases of the unskillful implementation of empirical methods for policy research. *See, e.g.,* G.S. Ford, *Whoops! Berkman Study Shows "Open Access" Reduces Broadband Consumption,* PHOENIX CENTER PERSPECTIVE NO. 09-05 (November 12, 2009) (available at: http://www.phoenix-center.org/perspectives/Perspective09-05Final.pdf); G.S. Ford, *Finding the Bottom: A Review of Free Press's Analysis of Network Neutrality and Investment,* PHOENIX CENTER PERSPECTIVE NO. 09-04 (October 29, 2009) (available at: http://www.phoenix-center.org/perspectives/Perspective09-04Final.pdf); G.S. Ford, *Econometric Analysis of Broadband Subscriptions: A Note on Specification,* PHOENIX CENTER PERSPECTIVE NO. 09-02 (May 12, 2009) (available at: http://www.phoenix-center.org/perspectives/Perspective09-02Final.pdf); G.S. Ford, *Fabricating a Broadband Crisis? More Evidence on the Misleading Inferences from OECD Rankings*, PHOENIX CENTER PERSPECTIVE NO. 10-05: (July 7, 2010) (available at: http://www.phoenix-center.org/perspectives/Perspective10-05Final.pdf); G.S. Ford and L.J. Spiwak, *Substantial Profits in the Broadband Ecosystem: A Look at the Evidence,* PHOENIX CENTER PERSPECTIVE NO. 10-04 (April 22, 2010) (available at: http://www.phoenix-center.org/perspectives/Perspective10-04Final.pdf).

A. *Unbiased Estimates*

The goal, of course, is to get an accurate estimate of the "true" coefficients of interests. If the estimated coefficients, typically labeled $\hat{\beta}$ or "beta hat," can be expected to equal the true value of the unknown parameter β, then the estimated coefficient is said to be an *unbiased* estimate.[9] A biased estimate, therefore, is not expected to reflect the true relationship between variables of interest. Importantly, bias is of a degree; that is, a coefficient may be "a little biased" or "a lot biased." Bias is undesirable, so if bias is expected, then some attention needs to be paid to its potential size. In some cases, but not all, it is possible to obtain an estimate of the size, or least the direction, of bias. An unbiased estimate of the β coefficients requires that a number of technical conditions be met; I refer the reader to any econometrics text for a full discussion.[10]

B. *Efficiency*

It should be kept in mind that regression analysis provides an estimate of the coefficients, and estimates have distributions, i.e., they are random variables. While the output of regression analysis is summarized by a point estimate, this point estimate is the average of statistically plausible values. In some cases, the range of potential values is "tight" around the point estimate. For example, if $\beta_1 = 1$, then a "tight" estimate may have a range of, say, 0.99 to 1.01. Alternately, a "wide" interval may be something like -0.20 to 2.20, where the range includes zero. A "tight" range typically produces a large t-statistic which is then interpreted to be "statistically significant." If, as in the latter case, the range includes zero, then the t-statistic will be small and the coefficient estimate will not be "statistically different from zero."[11] Theoretically, a coefficient has an ideal range, and if the econometric model matches that range, the estimate is said to be "efficient." If data or model errors lead to the estimate of wider or narrower range, then there is a loss of efficiency. Both bias and inefficiency are problems encountered in the analysis below. In some cases, researchers will trade off one for the other, but such advanced topics are beyond the scope of this BULLETIN.

---

[9] In some cases, bias may exist in small samples but not in large samples. Large sample unbiasedness is referred to as consistency. An inconsistent estimate is biased even in large samples. *See, e.g.,* D. Gujarati, BASIC ECONOMETRICS (1995), pp. 182-4.

[10] For example, excellent introductory texts include D. Gujarati *id.*; J. Wooldrige, INTRODUCTORY ECONOMETRICS (2001); R. Hill, W. Griffiths, and G. Judge, UNDERGRADUATE ECONOMETRICS (2001). More advanced texts include J. Wooldridge, CROSS SECTION AND PANEL DATA ECONOMETRICS (2002) and W. Greene, ECONOMETRIC ANALYSIS (2008).

[11] Typically, the hypothesis being tested in that β = 0, so "statistically significant" means statistically different from zero with high probability.

C. *Monte Carlo Setup*

While bias arising from the properties of the NBM data can be demonstrated theoretically, very few interested persons would be able to follow the arguments. Instead, I employ a simple Monte Carlo analysis that avoids most technical details, thereby making the analysis more accessible.[12] (Consequently, my discussion here is somewhat superficial. I refer the reader to any standard econometrics text for a more thorough treatment.) Monte Carlo analysis is a very simple but useful tool. In effect, a data set is generated with known properties (that is, we know the exact values of the β coefficients and the properties of the disturbance e). Since the true data generating process is known, it is possible to measure directly any bias resulting from data problems and modeling choices, since such problems can be incorporated into the Monte Carlo simulation.

To begin, a data set with known properties is needed. The data generation process for the Monte Carlo analysis used in this demonstration is as follows. An outcome y is determined by five variables x_1, x_2, x_3, x_4, and x_5, along with some random determinant e. The x's and e are simply made up; they are created using random number generators. The variable x_1 is drawn randomly from a chi-square distribution with 10 degrees of freedom; consequently, x_1 has only positive values and is skewed right (with a mean of 10). This variable is intended to proxy population or household count (which are positive and typically skewed right) and will be used in any weighted averaging conducted later. Other covariates x_2, x_3, x_4, and x_5 are all drawn randomly from a uniform distribution, so they have equal probability of being any value between 0 and 1. Most demographic variables are correlated across geography. For example, areas with higher average proportions of college graduates are also likely to have higher average incomes. Therefore, the variables x_2, x_3, x_4, and x_5 are assumed to be correlated to varying degrees.[13] The NBM has very large samples, so I generate a dataset of 10,000 observations. The outcome y is computed by assigned values to the β_1 coefficients from Equation (1), so that

$$y = -0.7 + 0.01x_1 + 1.5x_2 - 0.3x_3 + 1.0x_4 - 1.0x_5 + e\,, \tag{2}$$

---

[12] On Monte Carlo analysis, *see, e.g.,* C. Mooney, MONTE CARLO SIMULATION (1997). Many econometric textbooks also contain discussions of Monte Carlo analysis.

[13] Correlation is generated as follows: (1) let u be a random draw from the uniform distribution; (2) let $x_i = 0.3x_j + 0.7u$. The weighted sum is then re-scaled to fit on the unit interval. This process generates correlation coefficients with x_2 of about $\rho = 0.40$ for x_3, $\rho = 0.12$ for x_4, and $\rho = 0.03$ for x_5.

where e is a normally distributed random variable.[14] The outcome variable y is a continuous, normally-distributed variable, so the regressions are estimated, unless otherwise noted, by the technique termed Ordinary Least Squares.[15]

In a Monte Carlo analysis, the model is typically estimated many times with new data, either some or all of it generated at each turn. Then, the averages of the estimated coefficients (and other interesting statistics) are computed and compared to the true values (which are known). Bias can be directly computed as the difference between the true value of a coefficient and its estimated value. For example, if I set the disturbance term equal to 0 for all observations and estimate Equation (1), then the output of the regression will exactly match the coefficients in Equation (2). The results are summarized in Tables 1 and 2, Column A (at the end of the document). Table 1 contains the estimated coefficients, while Table 2 contains the percent deviation of the estimate from its true value [i.e., $(\hat{\beta}-\beta)/\beta$]. As expected, by setting all e = 0, the model fits the data perfectly and all the coefficients are *exactly* those listed above. The R^2 of the regression is 1.00, which means that the prediction of y using the estimated coefficients exactly matches the true values of y. This exercise merely confirms that least squares regression is capable of computing the true model in the absence of any influences outside the x's.

Alternately, if the disturbance e is assigned values from a normally-distributed random process, then the x's are not the sole determinants of y. Nevertheless, if I estimate Equation (2), then all the coefficients remain unbiased estimates of the true values. (See Tables 1 and 2, Column B.) This occurs because e is a random draw. The average percentage difference from estimated values and the true values is less than 1%. Note, however, that the R^2 falls below 1.00 (to about 0.86), so that now the prediction of y does not exactly match the true y;[16] the difference is simply e. Therefore, a correctly estimated model renders unbiased coefficient estimates if the data and model are valid. If the model is poorly specified, the data is measured with error, or the assumptions of regression analysis are otherwise violated, then regression may not render unbiased estimates. Using the Monte Carlo analysis, I can measure the bias by comparing the observed estimates to those in the "true" case (Table 1, Columns A and B).

III. Measurement Error

The presence of measurement error in the NBM data simply means the data includes noisy measures of things like speed, availability, and, by computation, the number of providers in a

---

14 The disturbance is generated as a normal random variable scaled by 0.2 ($e \sim N(0, 0.2)$,

15 Gujarati, *supra* n. 9, Ch. 3.

16 The R^2 can be modified by adjusting the scale of e.

5335 Wisconsin Avenue, NW, Suite 440
Washington, D.C. 20015
Tel: (+1) (202) 274-0235 ● Fax: (+1) (202) 318-4909
www.phoenix-center.org

geographic area. Obviously, errors in measurement occur only for things that are measured; errors related to things not measured are another issue (i.e., Selection Bias), discussed later.

There are many reasons to suspect that some of the variables in the NBM data are measured with error. The NTIA, for example, observed that "about 5 to 10 percent" of Americans do not have access, with such a wide range indicating that a more precise quantification of non-availability was not possible. Also, the NBM and most state maps permit users to submit corrections to the map, directly indicating the expectation of numerous errors. Systematic sources of error, some of which may persist, include, but are not limited to: (1) the data is collected at the state level by different entities using different methods; (2) the data is not native to a Census Block format, thus requiring translation from the raw to the final data; (3) the quality of service provider responses may systematically vary by service provider and by state; (4) Census Blocks are arbitrary boundaries and do not necessarily coincide with an economic market; and (5) the use of theoretical "coverage circles" may under- or over-state actual coverage. I suspect other researchers—particularly those closest to the collection of the raw data—can provide a number of additions to this list of measurement errors. Furthermore, while demographic variables from the Census Bureau are typically assumed to be low-error values, Census 2010 data is not yet available. As such, demographic data used in conjunction with the NBM are from Census 2000 and, thus, about ten years out of date.[17] Consequently, for now, it must be assumed that all available demographic data is measured with error.

As a specific anecdote, my home broadband connection, provided by Charter cable, typically runs at an impressive 30-40 Mbps down and 2 Mbps up. AT&T also offers DSL at my location. Yet, the NBM shows no wired service at my home. For Phoenix Center President Lawrence J. Spiwak, the NBM shows he has the option to buy Verizon's FIOS. Yet, while he could throw a rock from his home to Verizon's FIOS distribution plant, the service is not available at his house. It is unlikely that our experiences are unique.

---

[17] Even if the demographic data is error-free, it may not be possible to find a demographic measure of the exact variable required by the theory. In such cases, an error-free variable is used as a proxy for another variable, and thus the variable is mis-measured.

Table 1. Measurement Error in Speed

City	Download Speed Reported by NBM
Birmingham, AL	768 Kbps - 1.5 Mbps
Nashville, TN	3 – 6 Mbps
Milwaukee, WI	768 Kbps - 1.5 Mbps
Boston, MA	3 – 6 Mbps

Another example of measurement error is provided in Table 1. In the table, the download speed for Verizon Wireless in four major cities is provided.[18] Verizon's mobile broadband service offering is technically identical in these four cities, yet the NBM indicates sizeable differences. These differences are the result of variations in state data collection efforts. Which speed value is more correct is beyond the scope of this BULLETIN; the point is simply that the data is measured with error.

Measurement error is unquestionably present in the current version of the NBM data, so it is important to understand the possible consequences.[19] Formally, measurement error is defined as

$$y = y^* + v \qquad \text{or} \qquad x = x^* + v, \tag{4}$$

where y is the observed value of a variable of interest, y^* is the true value of that variable, and v is an error in measurement (and similarly for x).[20] The key question for measurement error is its impact on the estimated coefficients of a regression model. Next, I will consider the effects of errors in the measurement of the outcome y, and then in the measurement of the explanatory variables x using the Monte Carlo simulation.

A. *Measurement Error in the Dependent Variable*

When the dependent variable is measured with error, the consequences may be somewhat trivial or possibly profound. First, consider the case where the error in the dependent variable y is random and uncorrelated with the x's or e from Equation (1). For this case, we generate v as a normally distributed random number and add it to the y generated using Equation (2). In this simplest of cases, the random addition to y can be algebraically shown as a part of the random disturbance term e.[21] None of the assumptions that lead to unbiased estimates are violated, so

---

[18] The address provided the NBM was that of a U.S. Court House in each city.

[19] Indeed, the NBM and state-level maps all provide links to correct errors.

[20] This is the standard definition of measurement error. *See, e.g.*, Wooldridge (2002), *supra* n. 10, pp. 70-76.

[21] The model is $y^* + v = \beta x + e$. By algebra, we have $y^* = \beta x + (e - v)$.

5335 Wisconsin Avenue, NW, Suite 440
Washington, D.C. 20015
Tel: (+1) (202) 274-0235 • Fax: (+1) (202) 318-4909
www.phoenix-center.org

the coefficient estimates are unbiased estimates of the true values. The results are provided in Tables 1 and 2, Column C. In Table 3, I provide the percent difference in the estimated standard errors of the model with reference to their efficient levels (found in Table 3, Column B). Note that the measurement error reduces the efficiency of the estimates, meaning larger standard errors (about 40% larger) and a smaller R^2 (falling to 0.75). (See Table 1 and Table 3, comparing Columns B and C.) The impact on efficiency and R^2 will depend on how substantial the measurement error becomes.

Second, consider the case where the measurement error v is correlated with a covariate, in this case x_3. For example, it seems plausible that the speed and availability measures in the NBM data could vary by collection methods and procedures across states (see Table 1 above), and demographics certainly vary by state. The results of interest are provided in Tables 1 and 2, Column D. As expected, the coefficients are biased (particularly β_3). Tables 1 and 3 show a considerable reduction in both R^2 and efficiency.

B. *Measurement Error in the Explanatory Variables*

The NBM data may be used to measure outcomes or, alternately, used to explain outcomes (e.g., employment, wages, growth). Measurement error in the explanatory variables x is typically more problematic than error in the outcome y. Consider the case where the measurement error is limited to a single covariate x_3, which is correlated with x_2 and x_4, and to a lesser extent x_5. The results are provided in Table 1 and 2, Column E. As expected, the coefficient β_3 is biased (off 51%), as is β_2 (off 3%) and β_4 (off 4%). In fact, measurement error in the explanatory variables may bias all coefficients. Measurement error may be a greater concern when using the NBM data as an explanatory variable.

C. *Choice of Aggregation Level*

One source of measurement error in the explanatory variables arises from the choice of aggregation level. The purpose of the NBM was to provide availability data at a highly disaggregated geographic unit. As such, data is available at the Census Block level, of which there are approximately 8.2 million in the United States. While useful for highly-detailed maps, Census Bureau demographic data available at the block level is limited. Data on income, economic activity, and so forth is available only at higher levels of aggregation. The fine level provided by block-level data is no doubt attractive, but assume a proper model of availability requires the inclusion of variables measured only at higher levels of aggregation. An interesting question, therefore, is whether it is proper to extrapolate more aggregated data to less aggregated units, thereby exploiting the finest level of detail in the NBM data. In other words, can Block Group data be extrapolated to all Census Blocks within that group?

The extrapolation of group-level data to the block level can be viewed as a measurement error problem (the x's are wrong). Obviously, applying a block-group level average to each of its component blocks causes mis-measurement, since demographic factors may vary considerably across blocks within a given group. This mis-measurement is always problematic, but more so if the extrapolated variable is correlated with any of the other variables, and in almost all cases of demographic variables, there is correlation. Certainly, income is correlated with race, population density, age, and so forth. Table 1 summarizes the correlation coefficients for a number of demographic variables measured at the Census Block Group level (Census 2000) including percent of white population, percent of black population, percent of Hispanic population, percent of male population, percent of population age 22 to 29 years, percent of population 65 years or older, median age of the population, average household size, and population density (population per square mile).[22] The table shows very high correlations across many of the demographic factors.

Table 1. Correlation of Demographics

Variable	White	Black	Hispanic	Males	Age 22-29	Age 65+	Med. Age	HH Size	Pop. Density
White	1.00								
Black	-0.74	1.00							
Hispanic	-0.25	0.10	1.00						
Males	0.13	0.18	0.42	1.00					
Age 22-29	-0.03	0.35	0.49	0.76	1.00				
Age 65+	0.23	0.20	0.20	0.46	0.53	1.00			
Median Age	0.09	-0.42	-0.55	-0.76	-0.90	-0.45	1.00		
HH Size	-0.11	-0.36	-0.42	-0.79	-0.87	-0.77	0.89	1.00	
Pop. Density	-0.31	0.12	0.23	-0.07	0.07	-0.10	-0.04	0.05	1.00

To demonstrate the problem using the Monte Carlo approach, I divide the full sample of 10,000 block groups into 250 block groups (40 blocks per group, which is about average for the country). Assume variables x_1, x_2, and x_3 are measured at the Census Block level, but x_4 and x_5 are measured only at the Block Group level. Let x_i' represent the data at the group level computed as an x_1-weighted average of x_i. (In data generation, the x_1 variable is created to represent population or household count. The variable has only positive values and is skewed right.) Using this average, I extrapolate the data to the Census Block.[23] Practically, this approach assumes that, for example, household income measured at the Block Group level is equal in all Census Blocks within that group, which is unlikely to be true, thus causing measurement error.

---

22 Aggregate data may exhibit higher correlation than micro-level data.

23 Say there are 2 Blocks in a Block Group, one with an x of 2 and the other 4. Population in the blocks (measured by x_1 in the simulation) are 10 and 30. The variable x' is $(2 \cdot 10 + 4 \cdot 30)/(10+30) = 3.5$.

First, I estimate the model at the lowest level of aggregation (10,000 observations) replacing x_i with x_i'. As shown in Table 1 and 2, Column F, the coefficients β_4 and β_5 are severely biased, but β_3 also departs from its true value due to correlation.

Second, I move up the Block Group level and estimate the model again (using 250 observations), where all variables are an x_1-weighted average (including y'). Tables 1 and 2, Column G show that the estimated coefficients are all very close to their true values. As long as the β coefficients are identical across Census Blocks, then aggregating up to the Block Group does not exclude the potential for unbiased estimates.[24]

IV. Selection Bias is Inherent in the NBM Data

Selection bias poses significant problems for econometric analysis. Selection bias occurs when there is a systematic error due to a non-random sampling of a population. Ideally, sampling should be random. If it is not, then more sophisticated techniques may be required to address the non-random nature of the sample. In some cases, however, nothing useful can be done. In my opinion, such bias is unfortunately systematic in the NBM data today, particularly with regard to availability, and, consequently, the count of providers in a given area. Selection bias is different than measurement error, since this bias is, in part, a consequence of things not measured at all.

Data for the NBM is collected at the state level and requires compliance by the service providers in each state. Compliance is not complete and varies by state. In one state, for example, I know that only about 80 percent of providers submitted data. In the NBM algorithm, either a Census Block can be assigned a participating provider, or it cannot.[25] As a consequence, the NBM data only indicates that a Census Block has service by at least one provider. It does not directly indicate that a Census Block does not have broadband service. Rather, a Census Block either "has service" or the Census Block is missing from the data. By assumption, and assumption only, a "no report" from missing data is assumed to mean "no availability," but this assumption need not be valid. There are, in fact, four possibilities for a Census Block: (1) data is provided for all providers in the area; (2) data is provided for some, but not all, providers in the area; (3) data is not provided by any provider serving that area; and (4) there is no service in the area. The practice of assigning the first two possibilities as "having service" and the latter two as "not having service" is illegitimate. Likewise, counting providers

---

24 *See*, *e.g.*, A. Zellner, *An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias*, 57 JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION 348-368 (1962). One approach is to use mixed linear models in such cases.

25 In some cases, a Census Block has no data. In others, the Census Block may have data for some providers, but exclude data from non-participating providers.

in a given area as a measure of competition ignores the second and third possibilities in the list by assuming complete participation.

Selection bias occurs when non-participation in the sample is systematic. Unfortunately, at present, the selection bias is systematic. This can be illustrated simply. Let p be the probability a provider, serving an area inclusive of block b, refuses to participate in the program. If this provider is the only provider, then the probability we do not observe data for block b is p. If there are two providers, alternately, then the probability we do not observe data for block b is p^2. Generally, the probability is p^N, where N is the number of providers in the block. Thus, availability data (and the count of providers) has an error of a systematic nature. Blocks with few providers are more likely to have a "no report" than blocks with many providers. Since unserved markets have demographic and economic characteristics similar to markets that will be served by few providers, the bias is particularly acute for markets like those that will be genuinely unserved. Plainly, concluding that a "no report" means "no availability" is inaccurate and systematically so. This flaw is inherent in the nature of the NBM data. Over time, detailed validation efforts at the state level, as well as public participation, may reduce the selection bias.

Another indicator that the selection problem is systematic is that, while the non-participant rate may be, say, 20 percent of providers, the population implicated by non-participation is likely to be less than 20 percent. Consequently, there is systematic bias in that non-participants serve smaller markets than participants, and market size may be correlated with other demographics such as population density, education levels, racial composition, and so forth.

Systematic selection bias is devastating, making the measurement of causal effects very difficult. Theoretically, selection bias implies that measurement error is correlated with the disturbance term of the regression (e is Equation 2). Its effects can be evaluated using the Monte Carlo analysis by letting the measurement error, defined as before, be correlated with e. In particular, we adopt a sample inclusion rule that is a function of x_1, x_5, e, and an additional random disturbance term. (The simulation excludes 20% of the observations from the sample based on the selection rule.)

As shown in Table 1 and 2, Column H, the coefficient estimates are biased when the sample is non-randomly selected. All the coefficients are biased. Since selection bias is inherent to the NBM data, the empirical methods employed to analyze the data must address sample selection. The lack of a clear distinction between "no report" and "no service" renders most of the standard approaches for handling selection useless. It will be interesting to see how this selection bias is addressed by researchers, and if it is addressed at all. The use of household responses to correct for selection errors may improve the data substantially over time. The effect of household participation is fertile ground for research.

V. Bivariate Regression

If history is a useful indicator, then there will be a great deal of descriptive analysis using the NBM data. By descriptive analysis I mean simple statistical tests of means differences, correlations, and bivariate regression analysis. A bivariate regression is a regression that includes only *one* of potentially many explanatory factors x. Importantly, while these approaches may serve some particular legitimate purpose of the researcher, they do not permit causal conclusions and may be very misleading in that regard. The problem can be shown using the Monte Carlo analysis.

As shown in Equation (1), the outcome y is determined by five factors x_1, x_2, x_3, x_4, and x_5. A bivariate regression, in contrast, evaluates the outcome y with respect to only one of these factors, say x_3. Of course, if y is based on all five factors (x_1, x_2, x_3, x_4, x_5), then y cannot be legitimately modeled as being based on only one (x_3). To demonstrate the problem, I estimate a regression of y on x_3 alone. (*See* Tables 1 and 2, Column I.) While the true relationship between y and x_3 is negative (by Eq. 2), the estimated effect in the bivariate regression is positive.[26] The coefficient is biased to a significant degree, leading the researcher to conclude that the effect works in the opposite direction of the true relationship. The implication is obvious: if an outcome has many determinants, then the truth about the relationship between the two variables cannot be determined by a regression that includes only one of these determinants. To model a multivariate relationship as a bivariate relationship leads to omitted variables bias, so that the estimated relationships are biased. Bivariate regression is not much different than the use of correlation coefficients or simple means differences, so the expectation is that these methods are likewise biased estimates of the causal relationship of interest. For example, the correlation coefficient between y and x_3 average about 0.30 in the simulated data (though the true relationship is negative).

Bivariate regression, and the more general problem of omitted variables, need not create biased estimates of relationships. As with measurement error, the effect of omitted variables depends, in part, on the correlations among the variables. Recall that x_1 is, by construction, uncorrelated with the other determinants of y. If we estimate $E(y|x_1)$, then the coefficient is unbiased. As shown above, however, if we estimate $E(y|x_3)$, the estimated relationship is nothing like the true relationship. Given that almost all demographic variables are correlated with each other, the bivariate regression should never be used to assign a causal interpretation to the estimate.

---

[26] The coefficient is statistically different from zero.

VI. Modeling Broadband Availability

Much of the statistical analysis using the NBM data, as recognized by NTIA, will focus on comparing broadband availability among geographic areas and across demographic groups. Some of this analysis will be descriptive in nature, while some will attempt to assign causal linkages to relationships of interest. Either approach should be rooted in a meaningful economic framework establishing the determinants of availability across geographic areas.

In most cases, broadband is provided by private firms, and the standard assumption for the decisions of such entities is profit maximization. Put simply, broadband will be provided where it is profitable to do so. As noted in the *National Broadband Plan's* TECHNICAL PAPER NO. 1,

> Private capital will only be available to fund investments in broadband networks where it is possible to earn returns in excess of the cost of capital. In short, only profitable networks will attract the investment required. Cost, while a significant driver of profitability, is not sufficient to measure the attractiveness of a given build; rather, the best measure of profitability is the net present value (NPV) of a build.[27]

Availability, therefore, is determined by both the supply- and demand-side conditions of a geographic market that influence profits. Notably, a Census Block is not an economic market, but we set aside that issue for the present analysis, although this fact could be very important.[28] The geographic markets of a communications firm may be established by a license agreement with state regulators, so that a given Census Block may be in the licensed area of a particular firm, and not open to any and all broadband service providers.

Adhering to the FCC's view of the economics of availability, which matches the standard economic view, a market will be served if the following condition is satisfied:

$$(R - C) - F \geq 0, \tag{4}$$

where R is the expected flow (present value) of revenues, C is the (present value) expected flow of operating costs, and F is the (present value) fixed cost of deploying the network (the left hand side being the NPV of serving the market). The flow of revenues R is equal to $P \cdot Q$, where P is

---

[27] *The Broadband Availability Gap*, OBI TECHNICAL PAPER NO. 1, Federal Communications Commission (March 2010) (available at: http://download.broadband.gov/plan/the-broadband-availability-gap-obi-technical-paper-no-1.pdf) (hereinafter "*Gap Report*") at p. 1.

[28] As noted above, the Census Block does not comport with the provider's deployment decision.

price and Q is the quantity of service sold. The flow of cost C is based on $c \cdot Q$, where c is the operational cost per-unit multiplied by the output sold Q. We assume F includes the fixed maintenance and operational costs over time. We could Equation (4) as,

$$(P-c)Q-F \geq 0\,. \tag{5}$$

Any analysis of availability must start with this condition (and perhaps other considerations), and an econometric model must include each relevant factor directly, or the determinants of each relevant factor if the direct measures are not available, to sensibly replicate the real-world business decision to serve a given geographic unit.[29] For example, price may be determined by factors x, unit costs by factors z, quantity by factors w, and fixed costs by k, where the determinants x, z, w, and k may have common elements. Thus, when one asks the question about the availability outcome (y) in an econometric model, the general model is $y = f(x, z, w, k)$. Excluding any of these factors leads to omitted variables bias, which leads to biased estimates of the coefficients. As the explanatory variables are proxies for the direct measures, the researcher must consider the problem of measurement error and its implications.

Since availability is driven by profits, any meaningful statistical assessment of the presence or absence of a Digital Divide, typically defined as a lack of availability (and in some cases adoption), must account for the economic determinants of profits. Without question, geographic areas characterized by low incomes will have less broadband availability, since it is not in the interest of firms to spend millions to deploy network to persons and households unlikely to purchase services and thereby generate revenues sufficient to cover such investments. A sound economic framework for availability is particularly important in cases where a researcher aims to invoke the notion of "redlining." Redlining is "the practice of restricting or denying access to services in a spatially defined area" based on socio-economic factors (typically race) rather than simply good business practices (i.e., profits).[30] Since profits

---

[29] On proxy variables, *see* Wooldridge (2002), *supra* n. 10, at 63-7.

[30] E. Cohen-Cole, THE NEW PALGRAVE DICTIONARY OF ECONOMICS, Online Edition, 2010 (Eds. by S. Durlauf and L. Blume) (available at: http://www.dictionaryofeconomics.com/article?id=pde2010_R000280)) ("Redlining is the … the term arose from urban activists in Chicago in the 1960s in response to the literal practice by banks of drawing red lines on local maps to demarcate minority areas to which lending should be curtailed. Until the Fair Housing Act of 1968, this practice was legal and commonly used as a way to minimise the real or perceived risk of lending in these areas. Because minority areas are correlated socio-economic risk factors, they also tend to be correlated with financial risk. Current research has attempted to identify whether lenders differentiate supply of credit due exclusively to financial risk or due to racial factors"); BLACKS LAW DICTIONARY at 1279 (6th ed. 1990) (a pattern of discrimination in which financial institutions refuse to make mortgage loans, regardless of credit record of the applicant, on properties in specified areas because of alleged deteriorating conditions); W. Gruben, J. Neurberger, and R. Schmidt, *Imperfect Information and the Community Reinvestment Act,* 1990 ECONOMIC REVIEW 25-46 (1990) ("The arguments have centered around 'redlining,' a practice whereby a financial institution indiscriminately limits loans for the purchase of property in certain "undesirable" neighborhoods within its market area. According to this practice, lenders are

(Footnote Continued….)

are often correlated with such socio-economic factors, an econometric model of redlining must distinguish between the profit-driven decisions of firms and those driven strictly by socio-economic discrimination. Redlining is defined not as differential treatment, but differential treatment not attributable to the differential profits.[31] In other words, redlining exists when firms forgo profits in order to discriminate against particular socio-economic groups or factors.

To elaborate, say we have the profit condition for broadband availability for some geographic area,

$$(P-c)Q(r)-F \geq 0 \tag{6}$$

where r is a socio-economic factor of interest (e.g., income). It is well established that households with higher income are more likely to subscribe to broadband service (i.e., broadband service is a normal good), so that Q is larger in areas with higher average incomes, and thus profits are higher since profits rise as Q rises. Recent evidence of this relationship is provided in the NTIA's report entitled, *Exploring the Digital Nation*.[32] In that study, income was found to be a potent determinant of subscription, with higher income leading to higher rates of adoption, even with many other relevant factors held constant. Geographic areas with a low r have lower expected profitability and thus lower broadband deployment. If we observe lower deployment to areas with small r, therefore, then we cannot be surprised as "safe and sound operations" of a business would dictate a lower deployment rate. As such, observing a negative relationship between income and availability does not imply redlining. The presence of redlining requires deployment to be lower than "safe and sound operations" would prescribe, so that the availability decision is based on socio-economic discrimination (that, by implication,

---

alleged to deny loan applications for purchases in those geographic areas, regardless of the credit worthiness of the individual borrower. … A financial institution is said to redline if it indiscriminately denies loans for the purchase of property in certain 'undesirable' neighborhoods within its market area").

[31] *See, e.g.,* 12 C.F.R. § 25.21 (the Community Reinvestment Act does "not require a bank to make loans or investments or to provide services that are inconsistent with safe and sound operations. To the contrary, the [Office of the Comptroller of the Currency] anticipates banks can meet the standards of this part with safe and sound loans, investments, and services on which the banks expect to make a profit. Banks are permitted and encouraged to develop and apply flexible underwriting standards for loans that benefit low- or moderate-income geographies or individuals, only if consistent with safe and sound operations").

[32] *Exploring the Digital Nation: Home Broadband Internet Adoption in the United States*, NTIA (November 2010) (available at: http://www.ntia.doc.gov/reports/2010/ESA_NTIA_US_Broadband_Adoption_Report_11082010.pdf), at 47-8 ("Persons with high incomes, those who are younger, Asians and Whites, the more highly-educated, married couples, and the employed tend to have higher rates of broadband use at home (at i)").

reduces profits). In other words, redlining occurs when a broadband service provider forfeits profits in order to indulge its preference for discrimination.[33]

Socio-economic factors relevant to redlining arguments typically include income, race, and urban-rural distinctions. Identifying redlining based on these factors is complicated by the fact that each has a substantial direct and indirect influence on profits. As noted, broadband is a normal good, so profits are expected to be lower in low-income areas. Likewise, research indicates that Blacks and Hispanics have a lower demand for broadband, even holding other things like income and education constant. As observed by the NTIA,

> The gaps between Whites and Blacks registered at 10 percentage points and between Whites and Hispanics at 14 percentage points, even after controlling for household characteristics. A similar analysis found the urban-rural gap to be 7 percentage points.[34]

Since race is also correlated with income and education, indirect influences are also present. Rural markets are more costly to serve, so less availability is expected in rural markets based on lower profits.[35] An econometric analysis of redlining, therefore, requires the researcher to develop a model that permits recovery of an "enhancement" factor (i.e., a discrimination coefficient), where the socio-economic factor of interest is given more credit than it deserves and reduces the profits of the broadband service provider.[36] Since the socio-economic factors of interest are so fundamentally bound up in the profit function, and are likely relevant to sample selection (as discussed above), identifying redlining, *if possible*, will require highly-sophisticated econometric modeling. Econometric analysis, by the NTIA for example, suggests that the NBM data will reveal less broadband availability in areas with lower incomes, higher proportions of Blacks and Hispanics, and lower population densities (rural areas). Certainly, it is insufficient simply to assert that redlining exists because areas with a lower income, a particular racial mix, or low population density, have less broadband availability.

Observing socio-economic relations to availability may be very important for public policy. Universal service programs for telecommunications, for example, target lower income households for subsidized telephone services. Extending such programs to broadband service

---

33 This idea of redlining (or discrimination) is based on G. Becker, THE ECONOMICS OF DISCRIMINATION (1971). *See also* F. Blau, and L. Kahn, *Gender Differences in Pay*, 14 JOURNAL OF ECONOMIC PERSPECTIVES 75–99 (2000).

34 *Exploring the Digital Nation*, *supra* n. 32, at i-ii.

35 *Id. See also GAP Report*, *supra* n. 27.

36 Becker, *supra* n. 33.

is being considered by the FCC, as are subsidy programs to increase deployment in more rural areas of the country.[37] Broadband providers have likewise participated in numerous programs to encourage broadband adoption by lower income households and racial groups with historically low demand for such services. Similarly, major broadband service providers offer discounted broadband plans targeted at low income households.[38] Both wireline and mobile broadband providers participate in programs that encourage adoption and deployment to low income households.[39] AT&T provided $100 million to One Economy for providing free Internet service to 50,000 low income households by 2010.[40] Comcast has also worked with One Economy to expand broadband adoption and availability in low income communities.[41] Connected Nation, funded in part and otherwise supported by broadband providers, promotes broadband adoption and deployment in lower income areas.[42] In light of these efforts, it is incredulous to argue that broadband providers are systematically discriminating against low income households. While social programs to target such communities may be valid, their lack of broadband service need not be the consequence of discriminatory treatment based solely on socio-economic factors.

VII. Conclusions

As part of the American Recovery and Reinvestment Act of 2009, Congress set forth an ambitious goal of establishing a definitive National Broadband Map. Understandably, researchers are now anxious to use the data received to quantify the extent of broadband availability and to explain the relationships between availability and socio-economic factors. While the National Telecommunications and Information Administration should be lauded for

---

37 *See, e.g., In the Matter of Connect America Fund; A National Broadband Plan for Our Future; Establishing Just and Reasonable Rates for Local Exchange Carriers; High-Cost Universal Service Support, Developing an Unified Intercarrier Compensation Regime; Federal-State Joint Board on Universal Service; Lifeline and Link-Up,* FCC 11-13, NOTICE OF PROPOSED RULEMAKING AND FURTHER NOTICE OF PROPOSED RULEMAKING, __ FCC Rcd __ (rel. Feb. 9, 2011).

38 *See, e.g., In re Applications Filed by Qwest Communications Int'l and CenturyTel for Consent to Transfer Control,* FCC 11-47, MEMORANDUM AND ORDER, __ FCC Rcd __ (rel. March 18, 2011) at ¶¶ 35-37; *In the Matter of Applications of Comcast Corporation, General Electric Company and NBC Universal, Inc. for Consent to Assign Licenses and Transfer Control of Licensees,* FCC 11-4, MEMORANDUM OPINION AND ORDER, __ FCC Rcd __ (rel. Jan. 20 2011) at ¶ 233.

39 *See, e.g., AT&T and Cleveland Housing Network Connect Low-Income Residents*, News Release (October 4, 2007) (available at: http://www.att.com/gen/press-room?pid=4800&cdvn=news&newsarticleid=24491); *Leading Wireless Providers Team Up to Educate the Wireless Generation*, New Release (October 7, 2009) (available at: http://www.att.com/gen/press-room?pid=4800&cdvn=news&newsarticleid=27212).

40 *AT&T Announces $100 Million "AT&T AccessAll" Signature Program*, News Release (June 14, 2006) (available at: http://www.prdomain.com/companies/A/AT&T/newsreleases/200661533436.htm).

41 http://www.comcast.com/Corporate/About/InTheCommunity/Literacy/One_Economy.html.

42 http://www.connectednation.com/community_programs.

its inaugural effort to complete this Herculean task, it is nonetheless important to recognize that the mapping data currently has many known and yet-to-be-discovered defects. These errors include, but are not limited to, measurement errors and sample selection, both of which can cause severe problems with empirical analyses. In light of the known defects in the data and the lack of a robust data verification process, the analyst must proceed with caution and considerable modesty. Indeed, until a data verification process is put into place and the known defects remedied, all statistical and econometric results using the National Broadband Map data should be viewed with skepticism as such work permits no strong causal and policy-relevant conclusions. Over time, as some of the measurement error and sample selection problems are resolved, econometric models may offer more robust conclusions. As always, empirical research using the Map data should begin with a sound, rational framework explaining observed outcomes.

Table 1. Estimated Coefficients

	[A]	[B]	[C]	[D]	[E]	[F]	[G]	[H]	[I]
β_0	-0.70	-0.70	-0.70	12.17	-0.81	-0.81	-0.70	-0.64	-0.35
β_1	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00
β_2	1.50	1.50	1.50	1.41	1.46	1.50	1.50	1.50	0.00
β_3	-0.30	-0.30	-0.30	-3.77	-0.15	-0.09	-0.30	-0.30	0.68
β_4	1.00	1.00	1.00	1.37	0.96	0.82	1.00	1.01	0.00
β_5	-1.00	-1.00	-1.00	-1.05	-1.00	-0.81	-1.01	-1.23	0.00
Obs.	10,000	10,000	10,000	10,000	10,000	10,000	250	10,000	10,000
R^2	1.00	0.86	0.75	0.19	0.85	0.64	0.86	0.88	0.08

Table 2. Percent Bias in Estimated Coefficients

	[A]	[B]	[C]	[D]	[E]	[F]	[G]	[H]	[I]
β_0	0%	-0.1%	…	…	…	…	…	…	…
β_1	0%	-0.1%	-0.3%	17.0%	-0.1%	0.1%	0.7%	9.9%	…
β_2	0%	0.0%	0.0%	-5.7%	-2.8%	0.0%	0.1%	0.0%	…
β_3	0%	0.0%	0.1%	1156%	-50.9%	-70.3%	-1.0%	0.1%	-326%
β_4	0%	0.0%	0.1%	37.2%	-4.2%	-18.4%	0.1%	1.0%	…
β_5	0%	0.0%	0.1%	4.5%	0.1%	-19.1%	0.7%	23.1%	…

Table 3. Percent Bias in Standard Errors of Coefficients

	[A]	[B]	[C]	[D]	[E]	[F]	[G]	[H]	[I]
β_0	0%	0.0%	41.4%	3076.0%	2.0%	31.1%	516.8%	6.3%	…
β_1	0%	0.0%	41.4%	3068.8%	2.2%	60.4%	414.0%	5.0%	…
β_2	0%	0.0%	41.4%	3056.5%	-1.6%	60.4%	545.3%	6.2%	…
β_3	0%	0.0%	41.4%	3052.7%	-28.1%	53.5%	546.8%	6.1%	…
β_4	0%	0.0%	41.4%	3056.6%	0.1%	4465.4%	545.5%	4.6%	…
β_5	0%	0.0%	41.4%	3043.4%	2.2%	4658.4%	544.8%	21.6%	