Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

ON DISCLOSURE PROTECTION FOR NON-TRADITIONAL STATISTICAL OUTPUTS: KERNEL DENSITY ESTIMATORS[1]

David R. Merrell[2] and Arnold P. Reznek[3]

ABSTRACT

In this paper, we discuss a specific component of a research agenda aimed at disclosure protections for "non-traditional" statistical outputs. We argue that these outputs present different disclosure risks than normally faced and hence may require new thinking on the issue. Specifically, we argue that kernel density estimators, while powerful (high quality) descriptions of cross-sections, pose potential disclosure risks that depend materially on the selection of bandwidth. We illustrate these risks using a unique, non-confidential dataset on the statistical universe of coal mines and present potential solutions. Finally, we discuss current practices at the U.S. Census Bureau's Center for Economic Studies for performing disclosure analysis on kernel density estimators.

KEY WORDS: Kernel Density Estimator; Disclosure Protection; Non-Traditional Statistical Outputs

1. INTRODUCTION

By law, government statistical agencies must protect the confidentiality of the microdata provided by respondents to surveys and censuses. Agencies' traditional data products are aggregates of the microdata—usually tables of frequency counts or totals, or public use microdata files. For these products, agencies employ a variety of confidentiality protection (statistical disclosure limitation) methods, including restriction of output (e.g., cell suppression and re-aggregation) and perturbation (e.g., rounding of output, or random perturbation of either the underlying microdata or the output).

Researchers at the U.S. Census Bureau's Center for Economic Studies (CES) and its Research Data Centers (RDCs) commonly generate what is (for the Census Bureau, at least) "non-traditional" statistical output.[4] This output typically includes parameter estimates and related statistics from linear and non-linear regression models, and sometimes includes output from semi-parametric and non-parametric estimation

[1] This paper reports the results of research and analysis undertaken by U.S. Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official U.S. Census Bureau publications. This report is released to inform interested parties in ongoing research and to encourage discussion of work in progress. We would like to thank conference participants at the 2001 Statistics Canada Symposium in Ottawa, Canada as well as seminar participants at the Center for Economic Studies in Washington, DC for very helpful comments and suggestions. We also would like to thank Timothy Dunne for a number of insightful conversations and suggestions.
[2] David R. Merrell, Center for Economic Studies, U.S. Census Bureau and Department of Economics, UCLA, California Census Research Data Center, 4250 Public Policy Building, Box 951484, University of California Los Angeles, Los Angeles, CA 90095-1484 (dmerrell@ccrdc.ucla.edu)
[3] Arnold P. Reznek, Center for Economic Studies, U.S. Census Bureau, Washington Plaza II, Washington, DC 20233 (rezne001@ces.census.gov)
[4] For more information about CES and the RDCs, see the CES web site at http://www.ces.census.gov. See also, Cooper, Merrell, Nucci, and Reznek (1998) and Reznek, Cooper, and Jensen (1997).

models, or from simulation models.[5] We believe that current disclosure limitation methods may not always be appropriate for these outputs. However, the literature provides little guidance, despite its inclusion as a worthy research topic in a U.S. Federal Committee on Statistical Methodology report (FCSM, 1994, p. 87). This situation makes it difficult to balance the need to release meaningful research results while maintaining the confidentiality of respondent information. More guidance is needed since many agencies in and outside the U.S. have established or are considering establishing RDCs or similar restricted access sites.

This paper presents part of a larger work in progress. In this program of research, we are attempting to describe the risks of disclosure surrounding linear and non-linear regressions, and to evaluate the applicability of existing confidentiality protection methods in a regression context. See Federal Committee on Statistical Methodology (1994) for a discussion of existing confidentiality protection methods, and see Merrell and Reznek (2001) for a discussion of the directions our research is taking.

We also are investigating non-parametric estimators of probability densities, i.e. Kernel density estimators, focusing on how disclosure risk may be affected by the selection of a bandwidth, the selection of a density support, and the presence of outlying observations. We also are assessing the effectiveness of existing disclosure rules are in maintaining confidentiality when applied to fully non-parametric estimates of density functions. We will report on this strand of research in a future paper.

In this paper, we concentrate on the case of kernel density estimators. Though we only present examples pertaining to establishment data, observations could be on households, persons, or even firms. We believe that our arguments pertain to either economic or demographic types of data. In either event, the problem faced is a common one: we wish to allow the production and dissemination of high quality statistical estimates (such as can be obtained from kernel density estimates) while at the same time preventing the disclosure of individual respondent information. We examine potential disclosure risks that these forms of estimators present and also discuss some potential ways to address those risks.[6]

This paper proceeds in the following way. Section 2 defines and describes kernel density estimators. Section 3 discusses the data, illustrates two related disclosure risks and potential ways to address them, and synthesizes the inter-relatedness of these two concerns. Section 4 discusses current policy at the U.S. Census Bureau's Center for Economic Studies and its Research Data Centers when performing disclosure analysis on kernel density estimates. Section 5 concludes.

2. KERNEL DENSITY ESTIMATORS

2.1 Definition

Kernel density estimators allow researchers to describe cross-sectional distributions in very simple, clear ways. These estimators smooth the distributions based on a specified bandwidth and a kernel function. The kernel function places weights on each observation that makes that observation proportional to its

[5] The types of outputs discussed here are, of course, quite standard in statistics and social science research generally. However, they are "non-traditional" in the context of statistical disclosure limitation; the measurement of disclosure risks and statistical disclosure limitation methods are most developed for tabulations and public use files.

[6] To be sure, most uses of kernel density estimators involve examining differences in cross-sections—examining changes in the kernel density estimate. For example, one might be interested in changes in the size distribution of establishments between two, say, entry cohorts of manufacturing plants or coal mines. To examine this, one would need to plot more than one kernel density estimate. A single function such as those presented below normally would be atypical in research output. We use single estimators to present disclosure risks in any one kernel density estimate and implicitly make the case that the risks, potential solutions, and current policies apply no matter how the outputs are presented. Nonetheless, we recognize that most research output would compare two (or more) density estimates.

distance from the center of a given band. Formally, the kernel density estimator is given by the following equation:

$$\hat{f}_K = \frac{1}{nw} \sum_{i=1}^{n} K\left[\frac{x - X_i}{w} \right]$$

where n is the sample size, w is the bandwidth, K is the kernel function, x is an observation, and X_i is the midpoint of a given band. Note that the kernel function, K, can take a number of types: Epanechnikov, rectangular, or Gaussian, inter alia. In short, the kernel density estimate is the sum of weighted deviations of an observation from the midpoint of a band where the weights are assigned by the kernel function.

Mechanically, the kernel function is of little consequence to most researchers,[7] and most software packages will choose a default function. There is, however, considerable interest in the choice of the bandwidth, w. The bandwidth determines how many observations are included in each band and hence has a great amount of influence on the shape of the calculated kernel density. While there are a number of methods to calculate the bandwidth, an "optimal" bandwidth can be calculated as well; see Tapia and Thompson (1978) or Fox (1990) for discussions.

2.2 Potential Disclosure Risks

Kernel density estimators are useful and simple ways of illustrating cross-sections of data in a non-parametric way. In this way, they are of great benefit to users of confidential microdata. At the same time, since they are "non-traditional" statistical estimators, they present different sorts of risks when thinking about protecting the confidentiality of respondent information. This section briefly outlines those risks and in the next section, we illustrate those risks using non-confidential data that have the same properties and structure as some of the data housed at the U.S. Census Bureau's Center for Economic Studies and its Research Data Centers.

Generally, we think that kernel density estimators present potential disclosure problems in two very related ways. First, kernel density estimators present a potential risk when there are a small number of extreme values in the dataset. As will be shown, the calculated support of the kernel density estimator sometimes very closely approximates the true support of the empirical distribution—at one or both ends of the distribution. Second, and somewhat related to the first concern, choosing a small bandwidth (whether chosen optimally or not) tends to make more individual observations directly observable (particularly observations in the tails of the empirical distribution) and also tends to force the calculated support of the kernel density estimator to approximate closely the true support of the empirical distribution.

3. ILLUSTRATING DISCLOSURE RISKS IN KERNEL DENSITY ESTIMATORS

3.1 The Data

Before illustrating the two potential disclosure risks we associate with kernel density estimators, we describe the data used to illustrate them. We use a unique microdata set that contains the statistical universe of coal mining establishments in the U.S. in 1996. These data are collected under the regulatory and oversight authority of the U.S. Mine Safety and Health Administration and contain mine-level responses on quarterly employment, hours worked, and short tons of coal produced—among other things. To be sure, though these data are establishment level, they are not confidential. See Merrell (1999, 2000) and Dunne and Merrell (2001) for detailed descriptions of the data.

[7] For that matter, the kernel function is of little consequence in the performance of disclosure analysis as well. We do, however, mention a case below where the kernel function is a material component of the disclosure protection process. However, from a general standpoint, it is of little consequence when thinking about disclosure analysis.

A main benefit of these data for our purposes is that they have a similar file structure and similar statistical properties as data collected by the U.S. Census Bureau that are confidential. For example, the U.S. Census of Manufactures (a data set that is confidential under U.S. statutes) has establishment-level information on essentially the same variables (and then some) but the distributions of, say, size (measured by employment) are highly skewed. This prevents the creation of a public use dataset on manufacturing and would also present us from illustrating the problems with non-traditional estimators using these sorts of data. On the other hand, the coal mining data have the same sort of skew in the size distributions, but since these data are not confidential under U.S. statutes, we can use them as a substitute for manufacturing data and illustrate our concerns and solutions for non-traditional statistical outputs—like kernel density estimators.

To illustrate potential disclosure problems with kernel density estimators, we calculate the employment size distribution of mines in 1996. Employment at a mine in 1996 is measured as the mean number of workers per year (averaged over quarters). There are 1,328 mines determined as "active" in 1996; again, see Merrell (1999) for details. The largest mine had 840.75 workers, and the smallest mine had 0.5 workers.[8]

Figure 1 shows the kernel density estimate for these mines. To be sure, the kernel function is of the Epanechnikov type, and the band width is calculated as follows:

$$m = [\sqrt{\text{var}(x)}, \frac{IQR(x)}{1.349}]$$

$$w = \frac{0.9m}{n^{1/5}}$$

where x is an observation, IQR(x) is the inter-quartile range of x, and n is the sample size. This is the default method of calculating bandwidth using the Stata econometric package. See "Stata 7, Reference H-P."

Figure 1 presents this kernel density estimate. There are a few things to note. First, and most generally, the kernel density estimate is a very clear (and hence, high quality) description of the employment size distribution of coal mining establishments in 1996. Second, and more specifically, note that the mass of the distribution resides at the lower end of the density. Third, there is a very long tail at the upper end of the kernel density estimate resulting from the presence of a single very large mine in 1996. Finally, within that long tail there are a few larger but less extreme sizes of mines; these are shown by the small "humps" in the long tail of the density estimate. These mines in the upper end of the kernel density estimate present us with a potential disclosure problem associated with this type of non-traditional statistical output.

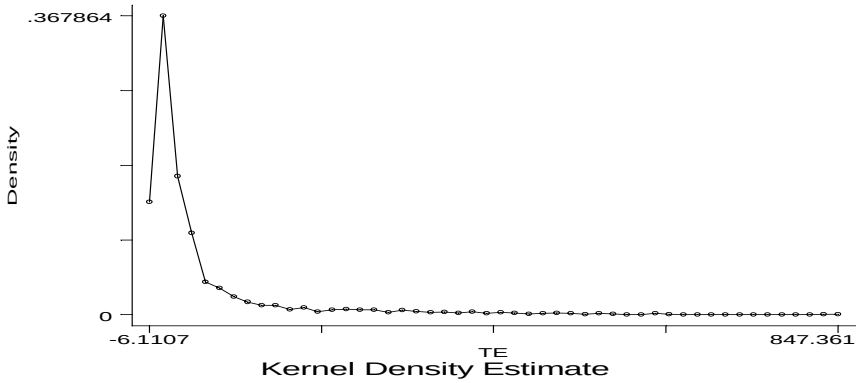

Figure 1. Kernel Density Estimate of Employment Size in 1996 (Default Bandwidth)

[8] Below, these values are referred to as the "true" maximum and minimum values.

3.2 Problems with Extreme Values and Possible Solutions

While it is true that the kernel density estimator provides a very clear, concise, and simple estimate of the distribution of employment size, it does so at the risk of disclosing respondent information—again, representing a tension between providing high quality statistical estimates and the requirement of protecting respondent information. Refer again to Figure 1 and specifically to the extreme values in the upper tail. There is a single mine that is very large relative to most mines, and the kernel density estimate for that mine size is 847.36 workers. Note that the true maximum value (viz., the actual size of the largest mine) is 840.75 workers. These values are extremely close; in fact, the calculated value is within one percent of the true value.

In most scenarios, this would be considered a disclosure violation for two reasons. First, reporting the true maximum value is the same as identifying an individual establishment—rather than an estimate of the size of that establishment. Second, because the kernel density estimate is virtually the same as the true maximum value, typical disclosure protocols would prevent its release.[9]

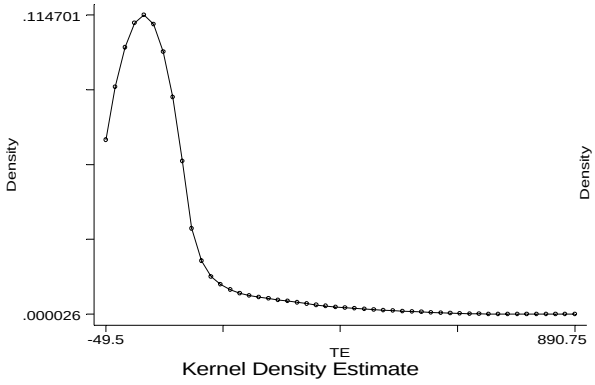

Figure 2a. Kernel Density Estimate, w=50

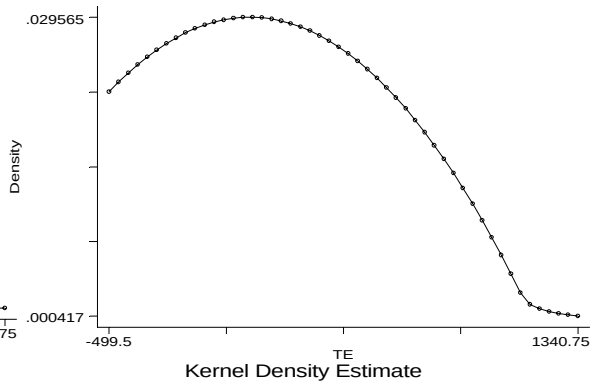

Figure 2b. Kernel Density Estimate, w=500

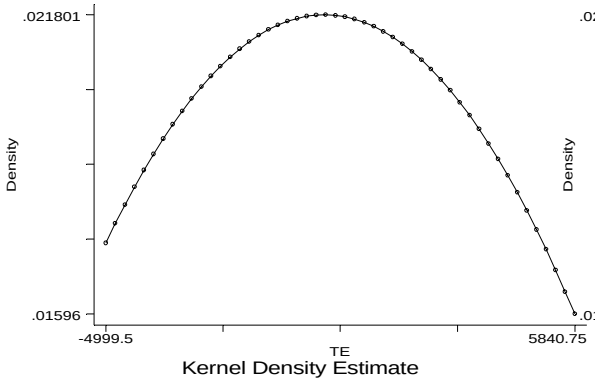

Figure 2c. Kernel Density Estimate, w=5,000

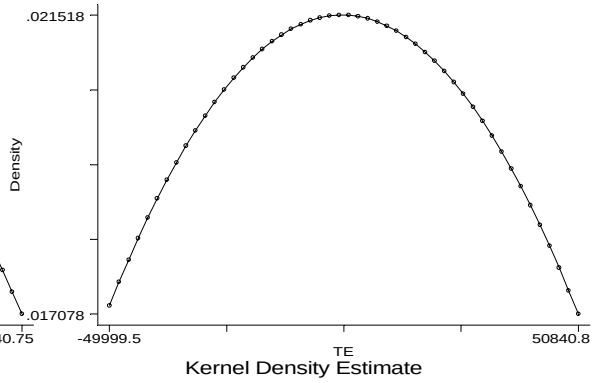

Figure 2d. Kernel Density Estimate, w=50,000

[9] While it is true that there are observable small "spikes" between the mass of the distribution and its extreme values, it is unlikely that those mines would be as easily identified (if at all) as the most extremely valued establishment; whether they are identifiable depends on the context. Hence, they are likely to be of lesser concern relative to the maximum or largest observation—at least in the context of disclosure analysis.

So, how do we resolve the tension between the need for high quality statistical estimates obtained using kernel density estimators with the requirement of maintaining the confidentiality of respondent information? One solution comes from the properties of the kernel density estimate itself. That is, we know that the choice of the value of the bandwidth determines fundamentally how smooth the density estimate will appear. The idea is that choosing a broader bandwidth should place less emphasis on any one individual observation—viz., that the kernel function will place a smaller weight on that extreme observation since it now lies even further from the center of a now larger band. This has the effect of masking the true value of the maximum observation by increasing the range of the calculated support of the density estimate.

Consider Figures 2a, 2b, 2c, and 2d. These figures adjust (manually) the width of the bands used to estimate the kernel density. Figure 2a sets the bandwidth at 50, and for the remaining three figures, the bandwidth is increased one order of magnitude over the previous one. The effects are clear. First, the distribution becomes more and more smooth. Second, and more importantly for present purposes, the calculated support on the distribution increases from 847.36 workers in Figure 1 to 890.75 workers in Figure 2a to 50,840.8 workers in Figure 2d.[10] The result of this exercise is that the true maximum value (recall that this is 840.75 workers) is substantially different than the calculated value as the bandwidth grows larger and larger. When the bandwidth is set to 50, the true value is within six percent of the calculated value (versus one percent with the default bandwidth in Figure 1). As the bandwidth gets larger, the calculated value at the upper extreme diverges from the true value.[11]

So, it would appear from a purely technical standpoint that the solution to dealing with extreme values would be to set a larger bandwidth; this would force a divergence between the true values of the extreme observations and the calculated values obtained from the kernel density estimator. While this is true, it comes at the cost of losing detail (and hence quality) in the estimate. Comparing across Figures 1 and 2a-2d, it is clear that the distribution goes from showing a reasonable amount of detail in the size distribution of coal mines to showing no detail at all—especially when the bandwidth exceeds the total sample size. The power of the kernel density estimator to describe cross-sectional size distributions is eliminated as the bandwidth is set larger and larger.[12]

3.3 Problems with Small Bandwidths and Possible Solutions

The second general concern with kernel density estimators is setting the bandwidth "too small"—which can happen even when choosing the "optimal" bandwidth. While a larger bandwidth tends to smooth the distribution and can "over-smooth" it, the opposite is true as well: smaller bandwidths tend to show more detail or, from a disclosure protection perspective, to "under-smooth" it. The effect is to make more observations individually observable at both the lower and upper extremes of the distribution. Additionally, the calculated support of the kernel density estimate tends to approximate closely the true values of the empirical distribution at the extremes.

---

[10] Clearly, Figures 2c and 2d have little meaning since the bandwidth is larger than the total number of observations available. That is, all observations are used in calculating the kernel density estimate at each point. Hence, all observations will fall within the calculated support in both of these cases.

[11] Though it is less consequential at this point, a similar effect occurs at the lower extreme of the distribution. That is, the calculated value of the lower extreme quickly becomes substantially different from the true value.

[12] Another potential disclosure risk (that we save for discussion until Section 4) pertains to the scales on the horizontal axis. Note that if one subtracts the bandwidth (assuming arguendo that it is known) from the maximum value labeled in the plot, one gets very close to the actual value, and in some cases, one might be able to obtain the exact value of the true maximum. (We thank Randy Becker and Kristen McCue for pointing this out to us.) As will be discussed later in this paper, our policy for dealing with this potential disclosure risk is to suppress the labeling on the horizontal axis on any graphical depiction of the kernel density estimate.

To illustrate, consider Figures 3a, 3b, 3c, and 3d. These figures reduce (manually) the bandwidth from the default in Figure 1 downward toward zero. The effects are clear. First, comparing across all figures, one will note that there is more and more detail revealed in the upper tail of the kernel density estimate. That is, there are more "spikes" in the upper tail—meaning that more larger coal mining establishments are being revealed as the bandwidth declines—not surprisingly.

The second feature to note is the calculated values of the kernel density estimate at both the larger and the smaller extremes. At the upper end, the calculated value of the maximum with the default in Figure 1 is 847.36 workers and is 845.75 workers when w=5, 841.75 workers when w=1, and 841.25 workers when w=0.5. Clearly, as the bandwidth declines, the calculated value approximates the true value more closely.[13]

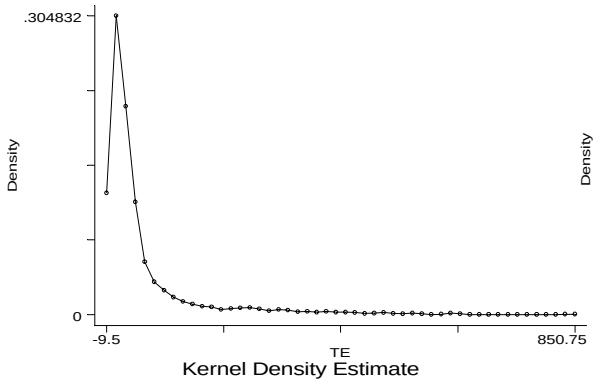

Figure 3a. Kernel Density Estimate, w=10

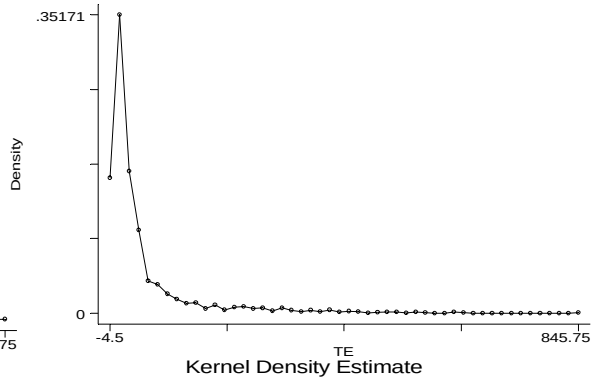

Figure 3b. Kernel Density Estimate, w=5

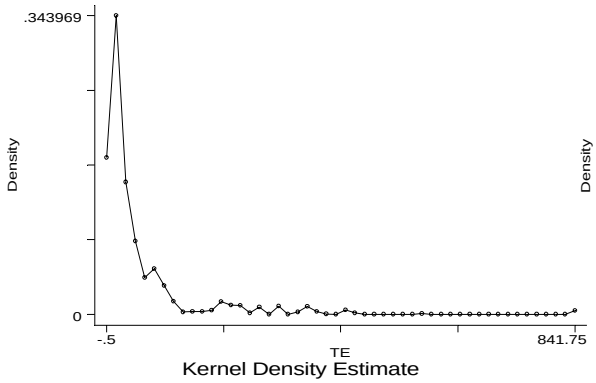

Figure 3c. Kernel Density Estimate, w=1

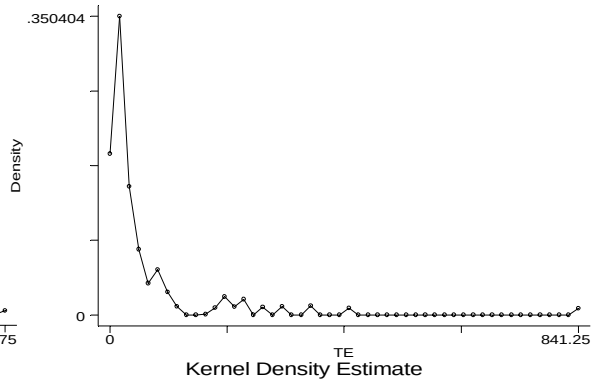

Figure 3d. Kernel Density Estimate, w=0.5

At the lower end of the kernel density estimate, a similar result appears: the calculated minimum value tends toward the actual minimum value. Notice that in Figure 1, the calculated minimum value is –6.11 workers—a value substantially different than the true value of 0.5 workers. However, as the bandwidth is set at smaller and smaller values (cf., Figures 3b-3d), the calculated minimum moves closer to the actual minimum. So, in the case where the bandwidth is set "too small," not only is more detail revealed about more establishments in the upper end of the distribution but also both the minimum and maximum calculated extreme values are more closely approximating the true maximum and minimum values. Again,

[13] Note that when the bandwidth is set to ten (Figure 3a), the upper and lower ends of the kernel density estimate grow larger than in Figure 1. This suggests that the "optimal" bandwidth as calculated above is less than ten—in this example.

the revelation of more detail about establishments in the upper tail and the more close approximation of the calculated support to the true empirical support are likely to be viewed as disclosure violations.

So, while it is the case that larger bandwidths tend to "over-smooth" the kernel densities, smaller bandwidths tend to present potential disclosure issues not only with presenting more information about larger values (in the sense of identifying more detail in the upper tail and the calculated maximum closely approximating the true maximum) but also with presenting more information about smaller values (in the sense of the calculated minimum closely approximating the true minimum). So, it would seem that from a simple perspective, the solution to smaller bandwidths would be to choose larger ones—the same solution as presented in Section 3.2. Again, the cost of doing so is to reduce the descriptive power and hence the quality of the kernel density estimator.

3.4 Synthesis

In both Sections 3.2 and 3.3, we argue that the selection of the bandwidth is a material factor in preventing the disclosure of respondent information when using kernel density estimators. In the case of extreme values, those can be masked by choosing a larger bandwidth, and in the case of small bandwidths, again, extreme values can be masked with choosing a larger bandwidth. The cost in both cases would be to reduce the quality of the estimate since the calculated support of the kernel density increases substantially in the choice of the bandwidth. Additionally, there is no objective way to set a bandwidth that both preserves the quality of the estimate and at the same time mitigates the risk of identifying the information of a given respondent—whether extremely large or extremely small.

One final solution, perhaps, would be to allow estimates of size distributions based on categorical information. That is, one could divide the number observations into categories of employment size and plot the frequency of their occurrence—to generate a simple histogram (a special type of kernel density estimator). Clearly, this solution is the most restrictive of all possibilities since it would proscribe the use of a very simple, yet powerful (viz., high quality) method of describing cross-sections—especially since there are methods (albeit subjective methods) for controlling potential disclosure risks.

4. CURRENT PRACTICES AT THE U.S. CENSUS BUREAU'S CENTER FOR ECONOMIC STUDIES

While we admit that kernel density estimators pose potential disclosure problems, we feel that their ability to present high quality estimates should compel disclosure analysts to devise methods of controlling those potential risks. In this section we discuss current practices at the U.S. Census Bureau's Center for Economic Studies and its Research Data Centers as they pertain to the disclosure analysis associated with kernel density estimators. We feel that these practices, though certainly subjective, balance the need to produce high quality statistical estimates against the concomitant need to maintain the confidentiality of respondent information.

The first step in disclosure analysis on kernel density estimators is that we test whatever sample of establishments or people using established statistical non-disclosure protocols. That is, for establishments, we (generally) examine the number of firms and the predominance of larger firms (based on real sales, employment, or whatever other measure might be available and appropriate). For data on people, we examine the number of respondents in each data cell. In either case (viz., for establishments or people), we compare calculated cell characteristics against a specified minimum; see Merrell and Reznek (2001) for a discussion of these rules. In order for a given kernel density estimate to clear disclosure analysis, these rules must be satisfied entirely.

Once it is determined that the overall sample underlying a given kernel density estimate satisfies established statistical non-disclosure protocols, we typically apply another condition that we think mitigates the concerns outlined in Section 3. That is, as a matter of general practice, we require that researchers suppress the scale on any graphical depiction of the kernel density estimate. By doing that, we effectively

can ignore the problems of bandwidths that are small enough to heighten the risk of identifying either extremely large or small observations as demonstrated above and as discussed in Footnote 12.[14]

Additionally, as a matter of policy, we only allow the release of graphical depictions of kernel density estimators; we never allow the point estimates themselves to be released. Consider the case where the point estimates (viz., the calculated values) were released. In this case, all one would need to know to recover the true values of individual respondent information are the kernel function and the bandwidth. With those two pieces of information, one would be able to reconstruct the true dataset—a clear, first order violation of non-disclosure policy. Further, the ease with which this could be accomplished is slight; all one would need to do is use standard spreadsheet software to recover the true respondent information.[15] A simple examination of the kernel density estimator formula shows this; refer to the equation above.

Finally, we take a rather conservative approach when dealing with non-traditional statistical outputs. In such, we spend a good deal of effort seeking the counsel of a large number of disclosure analysts. We hope that by doing so, we can not only come to some consensus when dealing with these more innovative, high quality statistical techniques but also to make whatever efforts are possible to eliminate the risks of primary and secondary disclosure. When there is some doubt or lack of understanding, we consult the U.S. Census Bureau's Disclosure Review Board for their collective input.

5. CONCLUSION

A recurring dilemma faced by statistical agencies that allow access to confidential microdata is the need to produce and promote the production of high quality statistical estimates while at the same time maintaining strict standards of confidentiality and statistical non-disclosure. As the state of the art in producing such estimates advances, so must our thinking about disclosure analysis and policy. This paper presents one of potentially many other types of estimators that produce both high quality statistical estimates and also new risks of disclosing respondent information.

We present a simple example using a unique dataset that has very similar distributional properties of non-public establishment data like the Census of Manufactures. We show that the risk of disclosing respondent information is materially dependent on the choice of the bandwidth when producing kernel density estimates. If the bandwidth is selected to be "too small" (whether chosen according to some "optimal" condition or selected manually), there is the risk of identifying more individual observations in the high end of kernel density, and there is also a very strong risk of identifying the true maximum and minimum values as the calculated support of the kernel density function tends to approximate closely the true support of the empirical distribution. On other hand, choosing larger and larger bandwidths to mask the identities of small and/or large extreme observations sacrifices the clarity of the statistical technique to produce simple yet meaningful descriptions of cross-sections.

Unfortunately, we have only been able to find subjective solutions to this problem. We apply traditional non-disclosure protocols to samples used to calculate the kernel density estimates, and once those criteria have been satisfied, we then require that scales on graphical depictions of the kernel density estimate be suppressed. In this way, we are able to eliminate the ability of the estimator to identify respondent information, but we do so at the cost of reducing the explanatory power of the estimate. Further, as a matter of policy, we only allow the release of graphical depictions of the kernel density estimator since the point estimates themselves, combined with knowledge about the type of the kernel function and the bandwidth, can be used to recover the true values of individual respondents.

In short, as the ability to produce innovative and high quality estimates takes statistical agencies into the future, we must allow our thinking about disclosure analysis to evolve as well. We must be able to balance the need for high quality, meaningful estimates against the also fundamental need to protect respondent

[14] This particular requirement often meets with considerable resistance from the users of confidential mircodata.
[15] We thank Timothy Dunne for suggesting that we consider this point.

information. The case of kernel density estimates presents an interesting case in the sense that they are simple yet powerful descriptions of cross-sections, but they present new disclosure risks that traditional disclosure limitation techniques may not by themselves be adequate to address. We use this example to illustrate both the dilemma that is faced more and more frequently by statistical agencies that grant restricted access to confidential microdata and also possible solutions to that dilemma.

REFERENCES

Cooper, J.M.R, D.R. Merrell, A.R. Nucci, and A.P. Reznek (1998), "Protecting Confidential Data at Remote Access Sites: Lessons Learned from Research Data Centers," *American Statistical Association Proceedings of the Section on Government Statistics and Section on Social Statistics*, 1998, pp. 91-96.

Dunne, T. and D.R. Merrell (2001), "Gross Employment Flows in U.S. Coal Mining," *Economics Letters*, 2, pp. 217-224.

Federal Committee on Statistical Methodology (1994), "Report on Statistical Disclosure Limitation Methodology," Statistical Policy Working Paper 22, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

Fox, J. (1990), "Describing Univariate Distributions," in J. Fox and J.S. Long (eds.) *Modern Methods of Data Analysis*, Newbury Park, CA: Sage Publications, pp. 58-125.

Merrell, D.R. (1999), "Acquisitions and Productivity in U.S. Coal Mining," Center for Economic Studies Working Paper No. 99-17, Washington, DC: U.S. Census Bureau.

Merrell, D.R. (2000), "Implications of Resource Exhaustion on Exit Patterns in U.S. Coal Mining," mimeo, Los Angeles, CA: UCLA.

Merrell, D.R. and A.P. Reznek (2001), "On Disclosure Protection for Non-Traditional Statistical Outputs," mimeo, Washington, DC: Center for Economic Studies, U.S. Bureau of the Census.

Reznek, A.P., J.M.R. Cooper, and J.B. Jensen (1997), "Increasing Access to Longitudinal Survey Microdata: the Census Bureau's Research Data Center Program," *American Statistical Association Proceedings of the Section on Government Statistics and Section on Social Statistics, 1997*, pp. 243-248.

StataCorp (2001), "Stata 7 Reference H-P," *Stata Statistical Software: Release 7.0*, College Station, TX: Stata Corporation.

Tapia, R.A. and J.R. Thompson (1978), *Nonparametric Probability Density Estimation*, Baltimore, MD: Johns Hopkins University Press.