THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

Two Notes on Relating the Risk of Disclosure for Microdata and Geographic Area Size

No. 134

Brian Greenberg and Laura Voshell
U.S. Census Bureau

December 1990

U.S. Department of Commerce U.S. CENSUS BUREAU

SURVEY OF INCOME AND PROGRAM PARTICIPATION

Working Paper Series

Two Notes On Relating the Risk of Disclosure for Microdata and Geographic Area Size

No 9029 134

Brian Greenberg
Laura Voshell
Bureau of the Census

SIPP Working Paper #134

December 1990

ACKNOWLEDGEMENT

The authors wish to thank some of their colleagues at the Census Bureau for their help during the preparation of this paper. They thank Jay Kim for assistance with obtaining and recoding the data used for this work and Jim Fagan for his valuable discussions on the topics in this paper. They also thank Cynthia Clark, Colleen Sullivan, and Eric Schindler for their comments on an earlier version.

TABLE OF CONTENTS

Page

1. Introduction..... 1

2. Modeling a Reduction in the Size of a Geographic Region..... 3

3. Size of the Geographic Region Versus Percent of Unique Households..... 3
   3.1 Varying the Number of Key Variables..... 4
   3.2 Varying the Categorical Breakdown of a Key Variable..... 4

4. Equivalence Classes, Geographic Detail, and Percent of Population Uniques..... 5
   4.1 New Uniques..... 6
   4.2 Equivalence Class Structure and Overall Entropy..... 6

5. Effects of Enlarging a Geographic Region on Original Uniques..... 7
   5.1 Varying the Number of Variables..... 7
   5.2 Varying the Categorical Breakdown of a Key Variable..... 9
   5.3 Using Overall Entropy to Measure Dispersion..... 9

6. Effects of Decreasing the Size of a Region on Population Uniques.... 10

7. Conclusions..... 10

Appendices

Figures

References

The Geographic Component of Disclosure Risk for Microdata

Brian Greenberg and Laura Voshell

ABSTRACT

National statistical agencies have the responsibility of collecting information about a nation's population and of publicly releasing this information without violating pledges of confidentiality. A statistical agency must consider the geographic detail on microdata files prior to their release. The finer the geographic breakout, the greater the risk that a respondent may be identified. In this paper, we regard the number of population uniques on a file as one component of a measure of risk and then relate this component of risk to identifiable geographic area size. The objective is to support the development of geographic area cut-offs when designing microdata release strategies.

KEY WORDS: Microdata, Disclosure Avoidance, Geographic Detail, Unique

1. INTRODUCTION

National statistical agencies and offices collect information about a nation's population and institutions and make the information available to the public. Statistical agencies have the responsibility of designing data release strategies which will not violate pledges of confidentiality either through intent or neglect. When a statistical agency releases microdata products, one of the important considerations is the geographic detail on the file. The finer the geographic breakout, the greater the risk that a respondent may be identified based on individual or household characteristics. In this paper, we regard the number of population uniques present on the microdata file as one of the components of a measure of disclosure risk and then relate this component of risk to identifiable geographic area size. One objective of this work is to contribute to the development of geographic area cut-offs when designing microdata release strategies.

Microdata files consist of records at the respondent level which contain characteristics of a sample of the individuals or households in a certain population. All obvious identifiers of respondents such as name or address have been removed. These records also contain geographic identifiers such as state or metropolitan area in which each respondent is located. The Census Bureau currently employs a general rule stating that no geographic region containing less than 100,000 people in the sampled area may be identified on a microdata file. However, for microdata from some surveys or censuses, the minimum number of people required per identified region may be larger than 100,000 if it is thought that the disclosure risk would be too great at that level. For example, for microdata from the Survey of Income and Program Participation (SIPP), no geographic region containing less than 250,000 people may be identified. One can reasonably assume that the smaller the identifiable geographic region on the file, the greater the disclosure risk.

We define the key variables on a set of microdata to be those variables which taken together may contribute to the linking of a record to its respondent (Bethlehem, Keller, and Pannekoek 1990; Greenberg 1990). In each identified geographic region, there may be records on the microdata file that represent individuals or households in that region which are unlike any other individuals or households in that region for the set of key variables. These

records will be called population uniques. The population uniques on a microdata file possess a high disclosure risk. A user of the microdata may know that an individual or a household in a given region has a unique combination of key variables, and if that combination of variables is represented in the microdata, the user would be able to link that respondent to its record. Also, if a user has access to a set of data records with individual identifiers and the same key variables as on the proposed public use microdata file, the user could match all records appearing in both sets of microdata that are unique with respect to the key variables. Under this scenario, unique individuals or households could be linked to their records and confidential information would be disclosed.

This paper attempts to describe the relationship between the percent of population uniques on a microdata file from a specific geographic region and the size of that geographic region. In most cases, when a geographic region is enlarged, the percent of individuals or households in that region which are unique decreases. This is because some of the individuals or households which are added to the region when the region is enlarged have the same combination of key variables as those individuals or households which were unique in the smaller region. Likewise, in most cases, when the size of a geographic region is reduced, the percent of individuals or households in that region which are unique increases. Some individuals which were not unique in the original region may be the only ones with their combinations of key variables remaining in the smaller region, thus they become unique.

In much of the work described below, we took simple random samples of a data set to model the change in the size of a geographic region and noted the difference between the percent of uniques in the original data set and in the subsets. We chose to use the procedure of taking random subsets from a data set rather than removing geographic areas from a specified geographic region in order to ensure that our work was controlled and replicatable and that our results would not be relevant solely to the region with which we were working. This method of modeling a change in the size of a geographic region is explained and described in detail in Section 2.

In Section 3, we discuss how the size of the geographic region affects the percent of unique individuals or households in that region for different sets of key variables and various categorical breakdowns of one of those key variables. We introduce the concept of equivalence classes in Section 4 and describe how the distribution of equivalence classes in a region affects the change in the percent of unique individuals or households brought about by a change in the size of the region. Our results from these two Sections help us to answer two main questions. When increasing the size of a geographic region in order to reduce the percent of unique individuals or households in that region, do we reach a point at which a further increase in size has no appreciable affect upon the percent of unique individuals or households in the region? And secondly, how does the similarity or dissimilarity of the individuals or households in a region affect the percent of population uniques as a function of the size of the region?

In Section 5, we examine how the enlarging of a given geographic region affects the records which are unique in the original region. The results from this section help us to answer a third question. How does the dispersion of the households in a region affect the percent of population uniques that remain unique when the region is enlarged? In Section 6, we examine how decreasing the size of a geographic region affects its population uniques and attempt to answer a fourth question. When a geographic region is reduced in size, how does the dispersion of the households in the original region affect the percent of uniques in the smaller region which became unique as a result of the reduction in the size of the region? Knowing the answers to these four questions may help in the designing of microdata release strategies. In the conclusion, we summarize our findings and offer recommendations concerning the

development of geographic area cut-offs for microdata files. An abbreviated version of this paper was presented at the 1990 Annual Meetings of the American Statistical Association and will appear in the proceedings of that meeting (Greenberg and Voshell 1990).

2. MODELING A REDUCTION IN THE SIZE OF A GEOGRAPHIC REGION

As stated earlier, in much of our work, we took simple random samples of a data set to model the change in the size of a geographic region and noted the difference between the percent of uniques in the original data set and in the subsets. We chose to use the procedure of taking random subsets from a data set rather than removing geographic areas from a specified geographic region in order to ensure that our work was controlled and replicatable and that our results would not be relevant solely to the region with which we were working. In order to compare the change in the percent of population uniques as geographic areas are removed from a region with the change in the percent of population uniques as randomly chosen elements are removed from a data set, we conducted the following experiment.

We began with a data set of 87959 household records from the 1980 Decennial Census. These 87959 households contained about 220000 individuals, and the 15 record variables were recoded to resemble possible key variables on SIPP microdata. See A.1 and A.3 in the Appendix for a description of the variables. These 87959 household records were from a total of 25 contiguous counties in Oregon. We then removed all records from two counties while maintaining the contiguousness of the remaining counties. We continued deleting counties by removing the records from counties until we obtained a total of 20 nested data sets of households from geographic regions of different sizes. See A.4 in the Appendix for a list of the counties included in each data set and the size of each data set. Using 6 variables and 15 variables, we calculated and plotted the percent of unique households in each data set. See Figures 1 and 2, plots A. We then created 19 random nested subsets of the original data set containing the exact same number of records as the data sets from the smaller geographic regions that we had created. Again using 6 variables and 15 variables, we calculated and plotted the percent of unique households in each data set. See Figures 1 and 2, plots B. From these plots, we can see that any difference in the effect on percent of uniques between the two procedures is small.

Thus, for example, suppose that an actual geographic region contains 50000 households, 10% of which are unique. If we consider a random subset containing 30000 of those households, we might find that say 25% of them are unique with respect to each other. It is reasonable to infer that if the size of that geographic region is reduced so that the region contains 30000 households, the percent of uniques will increase to approximately 25%. A further illustration of the quality of this model of taking random, nested subsets of a data set to simulate the change in the size of a geographic region is given in Section 5.1.b.

3. SIZE OF THE GEOGRAPHIC REGION VERSUS PERCENT OF UNIQUE HOUSEHOLDS

In this section, we discuss how the size of a geographic region affects the percent of households that are unique in that region for different sets of key variables and for various categorical breakdowns of one of those key variables. To model the effects of reducing the size of a geographic region, we conducted the following experiment. Starting with a "population" data set, we took simple random samples of the data set and noted the difference between the percent of uniques in the original data set and in the subsets.

3.1 Varying the Number of Key Variables

Starting with the same 87959 household records and using Poisson sampling, we randomly removed approximately 4398 records from the file to obtain a subset containing about 95% of the original records. We continued to randomly remove roughly this number of records until we had obtained 19 random, nested subsets containing approximately 95%, 90%, 85%, ..., 5% of the records in the original data set. Using 6, 10, and 15 record variables, we then counted the number of unique households in each of these data sets. See A.2 and A.3 in the Appendix for a description of the variables.

We plotted the percent of unique households versus the size of the data set for the data sets using the 6, 10, and 15 variables, as shown in Figure 3. These plots were decreasing and concave up. In each case, the percent of unique households leveled off considerably as the size of the data set increased. Note that we did reach a point where a further increase in the size of the data set had no appreciable affect upon the percent of unique households in that data set. Consider, for example, the data with 6 variables, as shown in Figure 3. When the size of the data set reached about 30000, a further increase in size offered almost no decrease in the percent of unique households in the data set. We were interested in observing the rate of change of the percent of uniques with respect to size of the data set. To analyze this rate of change, we approximated the derivative of the percent of unique records with respect to size of the data set for the 6, 10, and 15 variable cases. The measure that we used to approximate the derivative of the percent of uniques with respect to size s of the data set at size s is

$$\frac{\%\ \text{unique}\ (s+2199)\ -\ \%\ \text{unique}\ (s-2199)}{4398}$$

Our results are shown in Figure 4. We see in Figure 4, that the derivative is approximately 0 when the size of the data set reaches 30000 for the data with 6 variables. When comparing the results from the 6, 10, and 15 variable cases as shown in Figure 4, we note that the more variables used in the analysis, the larger the size of the subset needed to reach this point of diminishing returns.

In Figure 3, we see that the more variables used in the analysis, the larger the percent of unique households for data sets of corresponding size. This is to be expected because the larger the number of variables used, the more likely to find differences between those variables for different households. The more variables used, the more dissimilar the households can be. Also note that the greater the number of variables, the larger the decrease in the percent of unique households brought about by an increase in the size of the data set. Figure 4 shows that the more variables in a data set, the greater the rate of change in the percent of uniques at corresponding subset sizes. Thus the more dissimilar the households in a data set, the greater the rate of change of the percent of unique households in the data set with respect to size of the data set. A more detailed discussion on the method we used for quantifying the similarity or dissimilarity of households in a data set through the use of the entropy function will be presented in Section 4.

3.2 Varying the Categorical Breakdown of a Key Variable

Using the same 87959 household records with 15 variables and Poisson sampling, we randomly removed approximately 3159 records to obtain a subset containing about 96.4% of the records in the original data set. We continued to randomly remove approximately this number of records until we had obtained 10 random, nested subsets of this data set. Our smallest subset contained 56372 records. We calculated the percent of unique households in each data set six times

using different categorical breakdowns of the variable "payment" in order to see how geographic detail and the categorical breakdown of a key variable interact to affect the percent of unique households. In the SIPP context, the variable "payment" is the sum of utility costs and rent or mortgage payment, property taxes, and insurance. See A.5 of the Appendix for the six different categorical breakdowns of the variable "payment".

We plotted the percent of unique households in the data set versus the size of the data set for the various breakdowns of the variable "payment". These plots, shown in Figure 5, were decreasing and slightly concave up.

Entropy was used to measure the dispersion of the households over the categories of the variable "payment" for the original data sets of 87959 household records. If there were M categories of the variable "payment", and p_i was the probability that a household's "payment" was in category i, then

$$\text{ENTROPY} = -\sum_{i=1}^{M} p_i \times \ln(p_i)$$

Both the number of categories of the variable "payment" and the dispersion of the households over those categories affect the entropy value. For a fixed number of categories, the more evenly spread the household "payment" values over the categories, the higher the entropy. Entropy also increases as the number of categories increases given an even spread over the categories. We wanted to see whether this measure of dispersion was indicative of the percent of unique households in the data set. As seen in Figure 5, the larger this entropy value of the variable "payment", the larger the percent of households that were unique. Also note that the larger the entropy value, the larger (slightly) the rate of decrease in the percent of uniques as the size of the data set became larger. Thus the more disperse the households in the data set as measured by the entropy of one variable holding all others constant, the greater the decrease in the percent of population uniques brought about by an increase in the size of the data set. We extend the use of entropy to incorporate several variables jointly in Section 4 through the use of the equivalence class structure of the data set.

When examining Figures 3 and 5, it is interesting to note that, in this study, no matter how many variables are used in the analysis and no matter how the variable "payment" is broken into categories, the difference between the percent of unique households in a data set of 87959 household records and the percent of unique households in a data set of 56372 household records is never more than five percent.

4. EQUIVALENCE CLASSES, GEOGRAPHIC DETAIL, AND PERCENT OF POPULATION UNIQUES

Decreasing the size of a geographic region will cause some of the households which were not unique in the larger region to become unique in the smaller region. The number of households which become unique because of the reduction in the size of the region depends on the size of the reduction and on the similarity or dissimilarity of the households in the original region. This similarity is reflected in the distribution of the sizes of the equivalence classes (in a geographic region). An equivalence class consists of all households which have the same combination of key variables. All households within a region can be grouped with all other households exactly like them, and each group is an equivalence class. The number of households in each equivalence class is the size of that equivalence class. Unique households are equivalence classes of size 1. The distribution of the sizes of the equivalence classes for the data set of 87959 households using the 6, 10, and 15 variables listed in A.2 has been included in A.6 of the Appendix for the

purpose of allowing others to replicate much of the work described in this paper.

4.1 New Uniques

When a subset of a data set is considered, there will be some records in the subset which are unique with respect to all other records in the subset but which were not unique in the original data set. We will use the term new uniques for all such records. We will use the term original uniques for the records that were unique in the original data set. The expected number of new uniques in a random subset taken from a data set with a given equivalence class structure is calculated as follows.

Let N = number of records in the original data set
n = number of records in the subset
L = the size of the largest equivalence class in the original data set
t_k = the number of equivalence classes of size k in the original data set

Then the expected number of new uniques in a random subset of size n is

$$\sum_{k=2}^{L} t_k \times \frac{\binom{k}{1}\binom{N-k}{n-1}}{\binom{N}{n}}$$

The expected number of original uniques in a random subset of size n is

$$\frac{t_1 \times n}{N}$$

Thus, the expected percent of uniques in a random subset of size n is

$$\frac{t_1}{N} + \sum_{k=2}^{L} \frac{t_k}{n} \times \frac{\binom{k}{1}\binom{N-k}{n-1}}{\binom{N}{n}}$$

which is greater than or equal to t_1 / N, the percent of uniques in the original data set. As explained in Section 2, we have found that any difference between the change in the percent of population uniques brought about by the reduction in size of a geographic region and the change in the percent of population uniques brought about by removing a simple random sample of the households in that region is small. Therefore, when the size of a geographic region is reduced, it is expected that the percent of unique households in that region will increase. This formula also shows that the percent of household records that are unique with respect to other household records in a sample of a population is larger than the percent of households which are unique with respect to all other households in the entire population.

4.2 Equivalence Class Structure and Overall Entropy

We have shown that the expected increase in the percent of unique households

brought about by a reduction in the size of a geographic region depends upon the equivalence class structure of the households in the original region. We now attempt to quantify the dispersion or dissimilarity of the households in a region using a measure based upon the equivalence class structure of the households. We define this measure of dispersion as overall entropy which may be calculated as follows:

Let N = number of households in the original region
L = size of the largest equivalence class in the original region
t_k = number of equivalence classes of size k in the original region

We define

$$\text{OVERALL ENTROPY} = -\sum_{k=1}^{L} t_k \times [(k/N) \times \ln(k/N)]$$

The greater the dispersion of the households, the larger the value of overall entropy. Using the same 87959 household records and Poisson sampling, we created 9 random, nested subsets. We calculated the overall entropy of the original data set, and we calculated the percent of unique households in each subset ten times using sets of 6, 7, 8, ... , and 15 variables. For a description of these variables, see A.3 and A.7 of the Appendix. As one would assume, the larger the number of variables, the larger the overall entropy. The results are plotted in Figure 6. Note that the greater the dispersion as measured by overall entropy, the larger the percent of unique households for corresponding subset sizes and the larger the increase in the percent of unique households brought about by a decrease in subset sizes. So again, the more dissimilar the households in a data set, the greater the change in the percent of unique households brought about by a change in the size of the data set.

5. EFFECTS OF ENLARGING A GEOGRAPHIC REGION ON ORIGINAL UNIQUES

When a geographic region is enlarged, some households may be added to the region which have the same combinations of key variables as some of the households that were unique in the smaller region. Thus some households that were unique will remain unique, and others will not. We were interested in finding out how the dispersion of the households in the original data set affects the probability that a unique household will remain unique when the geographic region is enlarged.

5.1 Varying the Number of Variables

5.1.a Geographically Based Data Sets

Beginning with our 87959 household records from 25 contiguous counties, we removed all records from two counties while maintaining the contiguousness of the remaining counties. We continued deleting counties by removing the records from other counties until we obtained a total of 20 nested data sets of households from geographic regions of different sizes. Our smallest region contained 19034 households. See A.4 in the Appendix for a list of the counties included in each data set and the size of each data set.

Using 6 variables and 15 variables, we calculated and plotted the percent of unique households in the smallest data set that were also unique in each of the larger data sets. See Figure 7. Note that the more variables, the greater the percent of unique households that remained unique when the data set was enlarged. Recall that the addition of more variables into the analysis causes an increase in the dispersion of the households. Thus the

more disperse the households in a region, the greater the probability that a household that is unique in that region will remain unique if that region is enlarged.

When examining Figure 7, it is also interesting to note that in the 15 variable case, 70% of the records that were unique in the original region which contained 19034 households remained unique when the size of that region was quadrupled and contained about 80000 households.

5.1.b An Estimation Procedure and Random Data Sets

In Figure 7, we calculated and plotted the percent of unique households in a region that remained unique when that region was enlarged versus the size of the enlarged region. If the percent of unique households is known for some region and for that same region after it has been enlarged, then the percent of unique households in the original region that remained unique when the region was enlarged can be estimated as follows.

Let n_0 = number of households in original region
n_i = number of households in enlarged region
p_0 = percent of unique households in original region, known
p_i = percent of unique households in enlarged region, known
$u_0 = p_0 \times n_0$ = number of unique households in original region
$u_i = p_i \times n_i$ = number of unique households in enlarged region
$u_{0,i}$ = number of records that are unique in original region and remain unique in enlarged region

In Figure 7, we have plotted

$$\frac{100 \times u_{0,i}}{u_0} = \frac{100 \times u_{0,i}}{p_0 \times n_0} \quad \text{Versus} \quad n_i$$

for i = 0, ..., 19. If the original data set of 19034 records was not geographically based but was a <u>random subset</u> of the households in an "enlarged geographic region", then the expected value of $u_{0,i}$ would be

$$E (u_{0,i}) = p_i \times n_0$$

We will use this expected value as an estimate of the number of records that are unique in a <u>geographically based region</u> and remain unique when that region is enlarged. Thus we may estimate the percent of unique households in a region that remain unique when that region is enlarged as

$$\frac{100 \times p_i \times n_0}{p_0 \times n_0} = \frac{100 \times p_i}{p_0}$$

To illustrate this estimation procedure, we took a series of 16 random, nested subsets of our original 87959 household records using Poisson sampling. The smallest subset contained 19034 records, the same number of records as our smallest geographically defined data set described above. Using 6 and 15 variables, we then calculated the percent of uniques in each data set. In Figure 8, we have plotted

$$\frac{100 \times p_i}{p_0} \quad \text{Versus} \quad n_i$$

for i = 0, ..., 16 where p_0 is the percent of uniques in the smallest data set, p_i is the percent of uniques in the larger data set, and n_i is the size of the larger data set. Comparing Figures 7 and 8, we find almost no

differences between the two graphs. This shows not only that the estimation procedure works well, but also that, again, we find only small differences in results when comparing the effects of taking random subsets from a data set versus removing geographic sub-areas from a region.

It is also interesting to examine a version of Figure 8 where we let the size of the original data set as well as the size of the enlarged data set vary. Starting with the same 87959 household records and using Poisson sampling, we randomly removed approximately 4398 records from the file to obtain a subset containing about 95% of the original records. We continued to randomly remove roughly this number of records until we had obtained 19 random, nested subsets containing approximately 95%, 90%, 85%, ..., 5% of the records in the original data set. Using 6 and 15 record variables as listed in A.2 of the Appendix, we then calculated the percent of unique households in each of these data sets. Using methods previously explained, we estimated the percent of uniques in a data set that remain unique if that data set is increased in size by approximately 4398 households. In Figure 9, we have plotted $100 \times p_{i+1} / p_i$ versus n_i for $i = 1, \ldots, 19$ where

n_i = size of data set i
n_{i+1} = size of data set i+1
$n_{i+1} - n_i = 4398$
p_i = percent of uniques in data set i
p_{i+1} = percent of uniques in data set i+1

As one can observe, the pattern is more stable for the 15 variable case due to the low percentage of uniques when using 6 variables as seen in Figure 3. From Figure 9, we again conclude that the more variables used in the analysis, and thus the more disperse the households in a region, the greater the probability that a household that is unique in that region will remain unique if that region is enlarged. Figure 9 also supports the idea of a point of diminishing returns when it comes to enlarging geographic regions in order to decrease the percent of uniques in a region. The plots are concave down and level off as the size of the data sets increase. This implies that there is a point at which a further increase in the size of a region converts very few of the unique households in the region to non-uniques.

5.2 Varying the Categorical Breakdown of a Key Variable

Using the same 87959 household records with 15 variables and Poisson sampling, we randomly removed approximately 3159 records to obtain a subset containing about 96.4% of the records in the original data set. We continued to randomly remove approximately this number of records until we had obtained 10 random, nested subsets of this data set. Our smallest subset contained 56372 records.

Using six different categorical breakdowns of the variable "payment", we then calculated the percent of the uniques in the smallest data set that were also unique in the larger data sets. See A.5 of the Appendix for the six different categorical breakdowns of the variable "payment". We plotted the percent of uniques in the smallest data set that remained unique in the larger data sets versus the size of the larger data sets for each breakdown of the variable "payment". See Figure 10. The larger the entropy of the variable "payment", the larger the percent of original uniques which remained unique when the size of the data set was increased. Recall that the higher the entropy of the variable "payment", holding all other variables constant, the greater the dispersion of the records in the data set. Thus again, the more disperse the households in a region, the greater the probability that a household that is unique in that region will remain unique if that region is enlarged.

5.3 Using Overall Entropy to Measure Dispersion

Using the same 87959 household records and Poisson sampling, we created 9

random, nested subsets. We calculated the overall entropy of the original data set as defined in Section 4, and we calculated and plotted the percent of the unique households in the smallest subset which were also unique in the larger data sets ten times using sets of 6, 7, 8, ... , and 15 variables. See Figure 11. For a description of these variables, see A.3 and A.7 of the Appendix. Recall that the higher the value of overall entropy, the greater the dispersion of the households in the data set. Thus we see again that the more disperse the households in a region, the greater the probability that a household that is unique in that region will remain unique if that region is enlarged.

6. EFFECTS OF DECREASING THE SIZE OF A REGION ON POPULATION UNIQUES

When a geographic region is reduced in size, some households which were not unique in the larger region may be the only ones with their combinations of key variables remaining in the smaller region. Thus they become unique. In Section 4, we called such households "new uniques". We wanted to examine geographic regions that had been reduced in size to find out how the dispersion of the households in the original region affects the percent of uniques in the smaller region which became unique as a result of the reduction in the size of the region.

Using the same 87959 household records and Poisson sampling, we created 9 random, nested subsets. We calculated the overall entropy of the original data set as defined in Section 4, and we calculated and plotted the percent of the unique households in all of the subsets that were not unique in the original largest data set ten times using sets of 6, 7, 8, ... , and 15 variables. See Figure 12. For a description of these variables, see A.3 and A.7 of the Appendix. Recall that the higher the value of overall entropy, the greater the dispersion of the households in the data set. Note that the more disperse the households in the data set, the smaller the percent of uniques in each subset that were not unique in the original data set. Thus when the size of a geographic region is reduced, the more disperse the households in the original region, the smaller the percent of uniques in the smaller region that are new uniques.

7. CONCLUSIONS

As was stated earlier, we desired to answer two main questions from our results in Sections 3 and 4. When increasing the size of a geographic region in order to reduce the percent of unique individuals or households in that region, do we reach a point at which a further increase in size has no appreciable affect upon the percent of unique individuals or households in the region? And secondly, how does the dispersion of the individuals or households in a region affect the change in the percent of unique individuals or households brought about by a change in the size of the region? In our research, we have discovered that one does reach a point at which a further increase in the size of a region has almost no affect upon the percent of unique households in that region. The size at which this point occurs, however, varies for different data sets with different key variables. We have also noted that the more disperse the households from a region, the greater the increase in the percent of unique households brought about by a decrease in the size of the region.

The work described in Section 5 was completed in order to answer a third question. How does the dispersion of the households in a region affect the percent of population uniques that remain unique when the region is enlarged? In this Section, we saw that the more disperse the households in a region, the

larger the percent of population uniques that remain unique when that region is enlarged.

In Section 6, we answer our fourth question. When a geographic region is reduced in size, how does the dispersion of the households in the original region affect the percent of uniques in the smaller region which became unique as a result of the reduction in the size of the region? Here we found that when the size of a geographic region is reduced, the more disperse the households in the original region, the lower the percent of uniques in the smaller region that are new uniques.

Because different data sets contain different key variables, different numbers of key variables, and different categorical breakdowns of key variables, geographic detail has a different impact on each one. Each data set must be examined individually for possible disclosure risk. One may wish to use the percent of records in the data set which represent unique individuals or households as a way of quantifying disclosure risk. Although most sets of microdata records represent only a sample of a population, the percent of population uniques appearing on the file may be estimated using information from the sample (Willenborg, Mokken, and Pannekoek, 1990; Voshell 1990). Also, the percent of sample uniques may be used as an over-estimate of the percent of population uniques appearing on the file.

If a statistical agency chooses a certain maximum acceptable percent of either sample uniques or estimated population uniques required prior to the releasing of a set of microdata, it can change the number of key variables, the categorical breakdowns of those key variables, and the geographic detail on the microdata file until it has fulfilled that requirement. Dropping some key variables from the file, collapsing some of the key variable categories, and decreasing geographic detail are all ways of decreasing the percent of uniques on a file. The potential users of the microdata may express interest in some variables more than others or perhaps accept a decrease in the detail of all variables for an increase in geographic detail. In this way, the users may assist the statistical agency in arriving at a file providing as much data utility as possible with an acceptable disclosure risk. This type of interaction between agency and users and this type of trade-off between key variable detail and geographic detail will be incorporated in the design of the release strategies of the Public Use Microdata Samples (PUMS) from the 1990 Decennial Census and may be used in the future to develop a new geographic area cut-off for SIPP microdata.

APPENDIX: DESCRIPTION OF VARIABLES AND THEIR CATEGORICAL BREAKDOWNS

A.1

The microdata records used in this study were obtained from the 1980 Decennial Census and record variables were recoded to resemble possible key variables from SIPP. Data sets used in experimentation consisted of 6 and 15 variables as listed below. The categorical breakdowns of the variables are given in A.3.

6 Variables	15 Variables
Tenure	Tenure
Household Type	Household Type
Race	Race
Ethnicity	Ethnicity
Children	Children
Household Class	Household Class
	Payment
	Employment/Unemployment
	Veteran Status
	Disability
	Marital Status
	Household Income
	Social Security
	Public Assistance
	Other Income

The microdata records used in this study were obtained from the 1980 Decennial Census and record variables were recoded to resemble possible key variables from SIPP. Data sets used in experimentation consisted of 6, 10, and 15 variables as listed below. The categorical breakdowns of the variables are given in A.3.

6 Variables	10 Variables	15 Variables
Tenure	Tenure	Tenure
Household Type	Household Type	Household Type
Race	Race	Race
Ethnicity	Ethnicity	Ethnicity
Children	Children	Children
Marital Status	Marital Status	Marital Status
	Payment	Payment
	Employment/Unemployment	Employment/Unemployment
	Veteran Status	Veteran Status
	Disability	Disability
		Household Class
		Household Income
		Social Security
		Public Assistance
		Other Income

The categorical breakdowns of the variables used in analysis are found below.

1. Tenure

    a. NA
    b. Owner Occupied
    c. Renter with Cash Rent
    d. Renter with No Cash Rent

2. Household Type

    a. Everyone in Household Related
    b. At Least Two but Not All Persons in Household Related
    c. Single Person Household
    d. Otherwise

3. Race

    a. Class One, White Husband, White Wife
    b. Class One, White Husband, Black Wife
    c. Class One, White Husband, Indian Wife
    d. Class One, White Husband, Asian / Pacific Islander Wife
    e. Class One, Black Husband, White Wife
    f. Class One, Black Husband, Black Wife
    g. Class One, Black Husband, Indian Wife
    h. Class One, Black Husband, Asian / Pacific Islander Wife
    i. Class One, Indian Husband, White Wife
    j. Class One, Indian Husband, Black Wife
    k. Class One, Indian Husband, Indian Wife
    l. Class One, Indian Husband, Asian / Pacific Islander Wife
    m. Class One, Asian / Pacific Islander Husband, White Wife
    n. Class One, Asian / Pacific Islander Husband, Black Wife
    o. Class One, Asian / Pacific Islander Husband, Indian Wife
    p. Class One, Asian / Pacific Islander Husband, Asian / Pacific Islander Wife
    q. Class Two, Male Householder, White
    r. Class Two, Female Householder, White
    s. Class Two, Male Householder, Black
    t. Class Two, Female Householder, Black
    u. Class Two, Male Householder, Indian
    v. Class Two, Female Householder, Indian
    w. Class Two, Male Householder, Asian / Pacific Islander
    x. Class Two, Female Householder, Asian / Pacific Islander
    y. Class Three, White
    z. Class Three, Black
    aa. Class Three, Indian
    bb. Class Three, Asian / Pacific Islander
    cc. Otherwise

4. Ethnicity

   a. Class One, Both Spouses Spanish
   b. Class One, Male Spouse Spanish
   c. Class One, Female Spouse Spanish
   d. Class Two, Male Householder Spanish
   e. Class Two, Female Householder Spanish
   f. Class Three, Spanish
   g. Otherwise

5. Children

   a. NA
   b. Householder with Own Children Under 6
   c. Householder with Own Children Ages 6 - 17
   d. Householder with Own Children, Some Under 6 and Some 6 - 17
   e. Householder without children

6. Marital Status

   a. Now Married
   b. Widowed
   c. Divorced
   d. Separated
   e. Never Married

7. Payment (Rent or Mortgage Plus Utilities, Tax, Insurance, Etc.)

   a. PAYMENT = 0
   b. 1 ≤ PAYMENT < 50
   c. 50 ≤ PAYMENT < 75
   d. 75 ≤ PAYMENT < 100
   e. 100 ≤ PAYMENT < 125
   f. 125 ≤ PAYMENT < 150
   g. 150 ≤ PAYMENT < 175
   h. 175 ≤ PAYMENT < 200
   i. 200 ≤ PAYMENT < 250
   j. 250 ≤ PAYMENT < 300
   k. 300 ≤ PAYMENT < 400
   l. 400 ≤ PAYMENT < 500
   m. 500 ≤ PAYMENT < 600
   n. 600 ≤ PAYMENT < 700
   o. 700 ≤ PAYMENT < 800
   p. 800 ≤ PAYMENT < 900
   q. 900 ≤ PAYMENT < 1000
   r. 1000 ≤ PAYMENT

8. Employment / Unemployment

   a. Class One, Both Spouses Unemployed
   b. Class One, Husband Unemployed, Wife Employed
   c. Class One, Husband Unemployed, Wife Not in Labor Force
   d. Class One, Husband Employed, Wife Unemployed
   e. Class One, Husband Not in Labor Force, Wife Unemployed
   f. Class One, Both Spouses Not in Labor Force
   g. Class One, Husband Not in Labor Force, Wife Employed
   h. Class One, Husband Employed, Wife Not in Labor Force
   i. Class One, Both Spouses Employed
   j. Class Two, Male Householder Unemployed
   k. Class Two, Male Householder Not in Labor Force
   l. Class Two, Male Householder Employed
   m. Class Two, Female Householder Unemployed
   n. Class Two, Female Householder Not in Labor Force
   o. Class Two, Female Householder Employed
   p. Class Three, Unemployed
   q. Class Three, Not in Labor Force
   r. Class Three, Employed
   s. Other

9. Veteran Status

   a. Class One, Husband Veteran
   b. Class One, Wife Veteran
   c. Class One, Both Spouses Veterans
   d. Class Two, at Least One Male in Household is Veteran
   e. Class Two, at Least One Female in Household is Veteran
   f. Class Two, at Least One Male and at Least One Female are Veterans
   g. Class Three, Veteran
   h. Otherwise

10. Disability

   a. Class One, Husband Disabled
   b. Class One, Wife Disabled
   c. Class One, Both Spouses Disabled
   d. Class Two, Male Householder Disabled
   e. Class Two, Female Householder Disabled
   f. Class Three, Disabled
   g. Otherwise

11. Household Class

   a. Householder has Spouse Present
   b. Householder has No Spouse Present, Living with One or More Other Persons
   c. Single Person Household

12. Household Income

a.		HHINC	≤	0
b.	1	≤ HHINC	<	1000
c.	1000	≤ HHINC	<	3000
d.	3000	≤ HHINC	<	5000
e.	5000	≤ HHINC	<	7000
f.	7000	≤ HHINC	<	9000
g.	9000	≤ HHINC	<	11000
h.	11000	≤ HHINC	<	13000
i.	13000	≤ HHINC	<	15000
j.	15000	≤ HHINC		

13. Social Security

a.		SOCSEC	=	0
b.	1	≤ SOCSEC	<	500
c.	500	≤ SOCSEC	<	1000
d.	1000	≤ SOCSEC	<	1500
e.	1500	≤ SOCSEC	<	2000
f.	2000	≤ SOCSEC	<	2500
g.	2500	≤ SOCSEC		

14. Public Assistance

a.		PUBLIC	=	0
b.	1	≤ PUBLIC	<	500
c.	500	≤ PUBLIC	<	1000
d.	1000	≤ PUBLIC	<	1500
e.	1500	≤ PUBLIC	<	2000
f.	2000	≤ PUBLIC	<	2500
g.	2500	≤ PUBLIC		

15. Other Income

a.		OTHER	=	0
b.	1	≤ OTHER	<	500
c.	500	≤ OTHER	<	1000
d.	1000	≤ OTHER	<	1500
e.	1500	≤ OTHER	<	2000
f.	2000	≤ OTHER	<	2500
g.	2500	≤ OTHER	<	5000
h.	5000	≤ OTHER	<	10000
i.	10000	≤ OTHER	<	15000
j.	15000	≤ OTHER		

The households in these data sets were all in the state of Oregon. The counties included in each data set and the size of each data is listed below.

Data Set 1 87959 Records	Data Set 2 83198 Records	Data Set 3 78836 Records	Data Set 4 74446 Records
Linn	Linn	Linn	Linn
Marion	Marion	Marion	Marion
Lane	Lane	Lane	Lane
Deschutes	Deschutes	Deschutes	Deschutes
Crook	Crook	Crook	Crook
Wheeler	Wheeler	Wheeler	Wheeler
Grant	Grant	Grant	Grant
Wasco	Wasco	Wasco	Wasco
Hood River	Hood River	Hood River	Hood River
Sherman	Sherman	Sherman	Sherman
Gillian	Gillian	Gillian	Gillian
Malheur	Malheur	Malheur	Malheur
Harney	Harney	Harney	Harney
Lake	Lake	Lake	Lake
Benton	Benton	Benton	Benton
Douglas	Douglas	Douglas	Douglas
Klamath	Klamath	Klamath	Klamath
Tillamook	Tillamook	Tillamook	Tillamook
Polk	Polk	Polk	Polk
Lincoln	Lincoln	Lincoln	Lincoln
Union	Union	Union	Union
Coos	Coos	Coos	Coos
Jackson	Jackson	½ Jackson	
Curry			
Josephine			

A.4 continued

Data Set 5 70031 Records	Data Set 6 65363 Records	Data Set 7 62622 Records	Data Set 8 60751 Records
Linn	Linn	Linn	Linn
Marion	Marion	Marion	Marion
Lane	Lane	Lane	Lane
Deschutes	Deschutes	Deschutes	Deschutes
Crook	Crook	Crook	Crook
Wheeler	Wheeler	Wheeler	Wheeler
Grant	Grant	Grant	Grant
Wasco	Wasco	Wasco	Wasco
Hood River	Hood River	Hood River	Hood River
Sherman	Sherman	Sherman	Sherman
Gillian	Gillian	Gillian	Gillian
Malheur	Malheur	Malheur	Malheur
Harney	Harney	Harney	Harney
Lake	Lake	Lake	Lake
Benton	Benton	Benton	Benton
Douglas	Douglas	Douglas	Douglas
Klamath	Klamath	Klamath	Klamath
Tillamook	Tillamook	Tillamook	
Polk	Polk		
Lincoln			
Union			

Data Set 9 57462 Records	Data Set 10 54651 Records	Data Set 11 51700 Records	Data Set 12 47401 Records
Linn	Linn	Linn	Linn
Marion	Marion	Marion	Marion
Lane	Lane	Lane	Lane
Deschutes	Deschutes	Deschutes	Deschutes
Crook	Crook	Crook	Crook
Wheeler	Wheeler	Wheeler	Wheeler
Grant	Grant	Grant	Grant
Wasco	Wasco	Wasco	Wasco
Hood River	Hood River	Hood River	Hood River
Sherman	Sherman	Sherman	Sherman
Gillian	Gillian	Gillian	Gillian
Malheur	Malheur	Malheur	Malheur
Harney	Harney	Harney	Harney
Lake	Lake	Lake	Lake
Benton	Benton	Benton	
Douglas	½ Douglas		

A.4 continued

Data Set 13 44614 Records	Data Set 14 42190 Records	Data Set 15 40088 Records	Data Set 16 36418 Records
Linn	Linn	Linn	Linn
Marion	Marion	Marion	Marion
Lane	Lane	Lane	Lane
Deschutes	Deschutes	Deschutes	
Crook	Crook		
Wheeler	Wheeler		
Grant	Grant		
Wasco			
Hood River			
Sherman			
Gillian			

Data Set 17 32140 Records	Data Set 18 27814 Records	Data Set 19 23425 Records	Data Set 20 19034 Records
Linn	Linn	Linn	Linn
Marion	Marion	Marion	Marion
¼ Lane	½ Lane	¼ Lane	

Six categorical breakdowns of the variable "payment" used in the analysis and corresponding entropy values as defined in the body of the text.

Breakdown A
ENTROPY = 0.0417

a.		PAYMENT	=	0
b.	1	≤ PAYMENT		

Breakdown B
ENTROPY = 0.7276

a.		PAYMENT	=	0
b.	1	≤ PAYMENT	<	220
c.	220	≤ PAYMENT		

Breakdown C
ENTROPY = 1.3814

a.		PAYMENT	<	100
b.	101	≤ PAYMENT	<	220
c.	221	≤ PAYMENT	<	350
d.	350	≤ PAYMENT		

Breakdown D
ENTROPY = 2.0846

a.		PAYMENT	=	0
b.	1	≤ PAYMENT	<	25
c.	26	≤ PAYMENT	<	50
d.	51	≤ PAYMENT	<	75
e.	76	≤ PAYMENT	<	100
f.	101	≤ PAYMENT	<	125
g.	126	≤ PAYMENT	<	150
h.	151	≤ PAYMENT	<	175
i.	176	≤ PAYMENT	<	200
j.	201	≤ PAYMENT	<	500
k.	501	≤ PAYMENT	<	525
l.	526	≤ PAYMENT	<	550
m.	551	≤ PAYMENT	<	575
n.	576	≤ PAYMENT	<	625
o.	626	≤ PAYMENT	<	700
p.	701	≤ PAYMENT	<	750
q.	751	≤ PAYMENT	<	875
r.	875	≤ PAYMENT		

Breakdown E
ENTROPY = 2.5560

a.		PAYMENT	=	0
b.	1	≤ PAYMENT	<	50
c.	51	≤ PAYMENT	<	75
d.	76	≤ PAYMENT	<	100
e.	101	≤ PAYMENT	<	125
f.	126	≤ PAYMENT	<	150
g.	151	≤ PAYMENT	<	175
h.	176	≤ PAYMENT	<	200
i.	201	≤ PAYMENT	<	250
j.	251	≤ PAYMENT	<	300
k.	301	≤ PAYMENT	<	400
l.	401	≤ PAYMENT	<	500
m.	501	≤ PAYMENT	<	600
n.	601	≤ PAYMENT	<	700
o.	701	≤ PAYMENT	<	800
p.	801	≤ PAYMENT	<	900
q.	901	≤ PAYMENT	<	1000
r.	1000	≤ PAYMENT		

Breakdown F
ENTROPY = 2.9964

a.		PAYMENT	<	36
b.	37	≤ PAYMENT	<	50
c.	51	≤ PAYMENT	<	62
d.	63	≤ PAYMENT	<	78
e.	79	≤ PAYMENT	<	95
f.	96	≤ PAYMENT	<	113
g.	114	≤ PAYMENT	<	135
h.	136	≤ PAYMENT	<	163
i.	164	≤ PAYMENT	<	192
j.	193	≤ PAYMENT	<	217
k.	218	≤ PAYMENT	<	240
l.	241	≤ PAYMENT	<	264
m.	265	≤ PAYMENT	<	288
n.	289	≤ PAYMENT	<	315
o.	316	≤ PAYMENT	<	347
p.	348	≤ PAYMENT	<	383
q.	384	≤ PAYMENT	<	428
r.	429	≤ PAYMENT	<	484
s.	485	≤ PAYMENT	<	571
t.	571	≤ PAYMENT		

The distribution of the sizes of the equivalence classes for the data sets containing 87959 household records using the sets of 6, 10, and 15 variables listed in A.2.

6 VARIABLES

EQUIV. CLASS SIZE	NUMBER OF CLASSES	EQUIV. CLASS SIZE	NUMBER OF CLASSES	EQUIV. CLASS SIZE	NUMBER OF CLASSES
1	334	41	1	152	1
2	127	42	3	158	1
3	63	43	2	159	1
4	33	44	4	165	1
5	30	45	3	166	1
6	20	47	1	174	1
7	21	48	1	181	1
8	18	49	2	183	1
9	16	50	1	185	1
10	11	51	3	196	1
11	14	52	1	206	1
12	13	53	2	218	1
13	8	55	2	224	1
14	8	56	2	244	1
15	4	57	1	246	1
16	9	58	1	257	1
17	6	60	1	272	1
18	9	71	1	286	1
19	6	73	2	302	1
20	7	74	2	310	1
21	9	75	3	334	1
22	2	78	1	357	1
23	4	79	1	378	1
24	3	83	2	432	1
25	2	86	1	439	1
26	5	88	1	505	1
27	6	94	1	701	1
28	1	98	2	818	1
29	3	107	1	836	1
30	3	109	1	940	1
31	7	111	2	969	1
32	4	112	1	1397	1
33	1	113	1	1456	1
34	4	119	1	1840	1
35	2	131	1	2009	1
36	1	132	1	2019	1
37	1	133	1	2076	1
38	2	144	1	2165	1
39	2	146	2	3522	1
40	3	148	1	3541	1

A.6 continued

6 VARIABLES continued

EQUIV. CLASS SIZE	NUMBER OF CLASSES
3846	1
4136	1
4988	1
10994	1
23476	1

A.6 continued

10 VARIABLES

EQUIV. CLASS SIZE	NUMBER OF CLASSES	EQUIV. CLASS SIZE	NUMBER OF CLASSES	EQUIV. CLASS SIZE	NUMBER OF CLASSES
1	7860	45	4	91	5
2	1898	46	7	92	1
3	837	47	10	93	3
4	498	48	4	94	1
5	366	49	5	96	3
6	243	50	11	97	1
7	223	51	8	98	3
8	126	52	4	99	3
9	159	53	7	100	3
10	99	54	7	101	2
11	84	55	2	102	2
12	60	56	6	103	3
13	65	57	8	104	3
14	70	58	4	105	2
15	44	59	6	106	4
16	47	60	3	107	2
17	33	61	1	109	1
18	46	62	2	110	2
19	31	63	2	112	1
20	32	64	7	113	3
21	35	65	3	114	3
22	30	66	2	115	1
23	18	67	5	118	4
24	12	69	2	122	2
25	22	70	4	123	2
26	14	71	2	126	2
27	12	72	2	128	2
28	16	73	5	130	1
29	19	74	2	131	1
30	16	75	4	135	1
31	15	77	3	136	2
32	9	78	1	138	1
33	8	79	1	139	2
34	6	80	4	140	1
35	14	81	1	141	2
36	9	82	2	144	1
37	11	83	2	145	4
38	13	84	4	147	3
39	12	85	1	148	2
40	10	86	3	149	1
41	9	87	2	150	1
42	12	88	3	151	2
43	13	89	1	152	2
44	2	90	3	154	2

A.6 continued

10 VARIABLES continued

EQUIV. CLASS SIZE	NUMBER OF CLASSES
156	1
158	1
161	2
162	1
163	1
164	1
166	2
167	1
169	1
170	1
171	2
172	1
173	1
174	3
176	1
178	1
179	2
180	1
183	1
185	1
186	1
187	2
188	1
189	1
190	1
191	1
194	1
195	1
201	1
202	1
203	2
204	1
206	1
208	1
219	1
220	2
223	1
224	1
226	1
227	1
232	1
234	1
239	1
240	1
242	1
243	2
246	2
250	1
255	1
256	1
257	1
266	1
279	1
287	1
291	1
293	2
297	1
303	1
304	1
309	2
318	1
322	1
328	1
331	1
341	1
358	2
366	1
383	1
397	1
414	1
421	1
429	1
456	1
467	1
470	1
487	1
499	1
501	1
533	1
684	1

A.6 continued

15 VARIABLES

EQUIV. CLASS SIZE	NUMBER OF CLASSES	EQUIV. CLASS SIZE	NUMBER OF CLASSES	EQUIV. CLASS SIZE	NUMBER OF CLASSES
1	30908	45	10	100	2
2	4342	46	4	103	1
3	1666	47	3	105	1
4	899	48	6	106	2
5	515	50	5	107	2
6	359	51	4	108	2
7	265	52	3	109	1
8	169	53	7	110	1
9	158	54	3	113	1
10	123	55	3	114	1
11	92	56	4	115	1
12	80	57	2	118	1
13	76	58	3	122	1
14	55	59	3	125	1
15	63	60	7	126	1
16	47	61	3	129	1
17	34	62	4	136	1
18	30	63	6	137	1
19	37	64	2	138	1
20	31	65	1	154	1
21	27	66	3	162	1
22	28	68	2	200	1
23	26	71	2	393	1
24	20	72	2		
25	23	73	2		
26	20	74	2		
27	11	75	3		
28	10	76	3		
29	22	78	2		
30	11	79	2		
31	11	81	1		
32	15	82	2		
33	4	83	2		
34	13	84	2		
35	14	85	2		
36	9	86	1		
37	17	87	1		
38	7	89	1		
39	11	90	1		
40	5	91	2		
41	7	92	1		
42	6	94	2		
43	3	96	2		
44	8	97	2		

Data sets used in experimentation consisted of sets of variables as listed below. The categorical breakdowns of the variables are given in A.3.

A. 6 Variables
OVERALL ENTROPY = 3.00

Tenure
Household Type
Household Class
Race
Ethnicity
Children

B. 7 Variables
OVERALL ENTROPY= 3.41

Tenure
Household Type
Household Class
Race
Ethnicity
Children
Marital Status

C. 8 Variables
OVERALL ENTROPY = 4.06

Tenure
Household Type
Household Class
Race
Ethnicity
Children
Marital Status
Veteran Status

D. 9 Variables
OVERALL ENTROPY = 5.17

Tenure
Household Type
Household Class
Race
Ethnicity
Children
Marital Status
Veteran Status
Employment/Unemployment

E. 10 Variables
OVERALL ENTROPY = 5.73

Tenure
Household Type
Household Class
Race
Ethnicity
Children
Marital Status
Veteran Status
Employment/Unemployment
Disability

F. 11 Variables
OVERALL ENTROPY = 7.83

Tenure
Household Type
Household Class
Race
Ethnicity
Children
Marital Status
Veteran Status
Employment/Unemployment
Disability
Payment

A.7 continued

G. 12 Variables
OVERALL ENTROPY = 8.87

Tenure
Household Type
Household Class
Race
Ethnicity
Children
Marital Status
Veteran Status
Employment/Unemployment
Disability
Payment
Household Income

H. 13 Variables
OVERALL ENTROPY = 9.23

Tenure
Household Type
Household Class
Race
Ethnicity
Children
Marital Status
Veteran Status
Employment/Unemployment
Disability
Payment
Household Income
Social Security

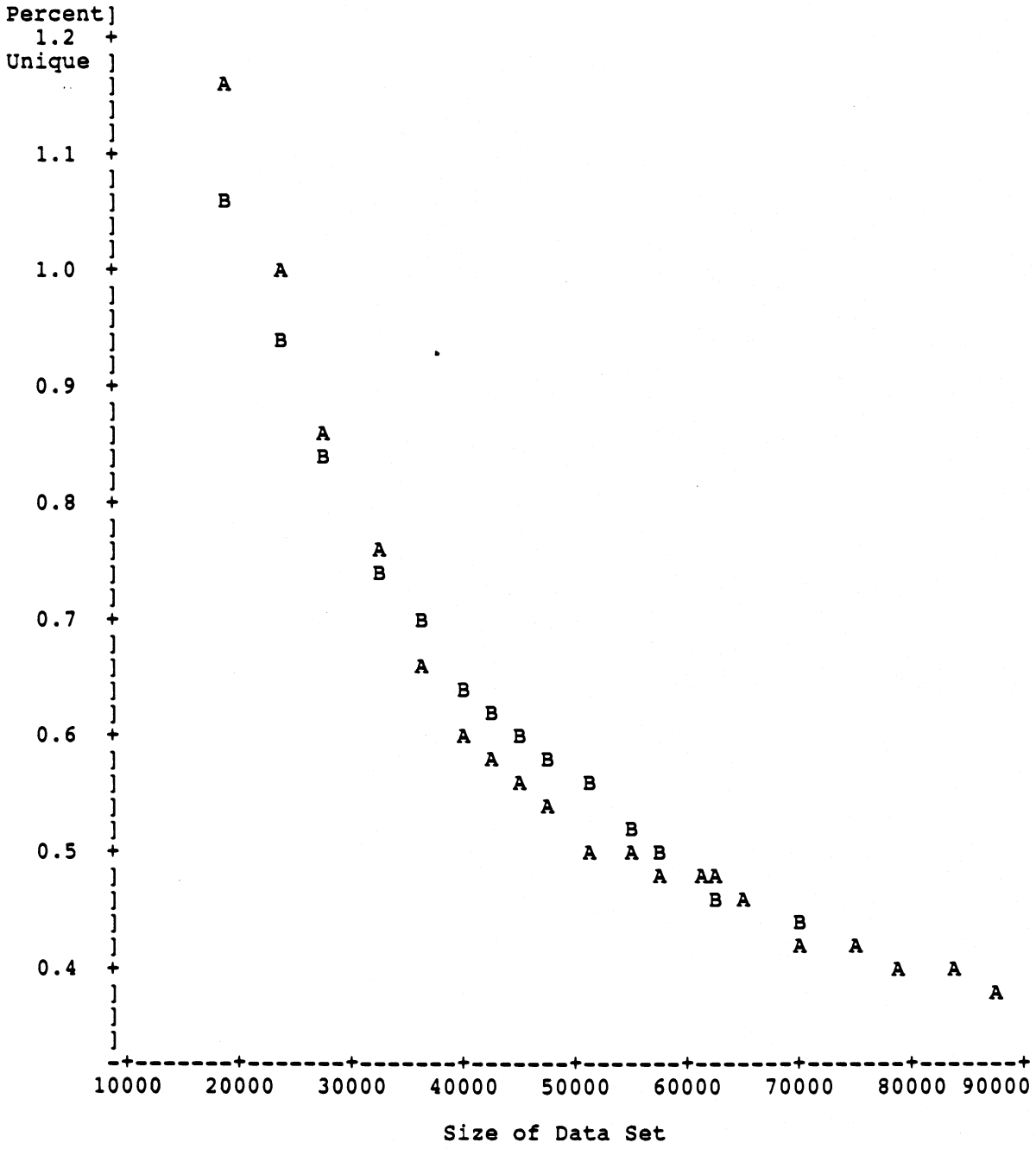
I. 14 Variables
OVERALL ENTROPY = 9.33

Tenure
Household Type
Household Class
Race
Ethnicity
Children
Marital Status
Veteran Status
Employment/Unemployment
Disability
Payment
Household Income
Social Security
Public Assistance

J. 15 Variables
OVERALL ENTROPY = 9.77

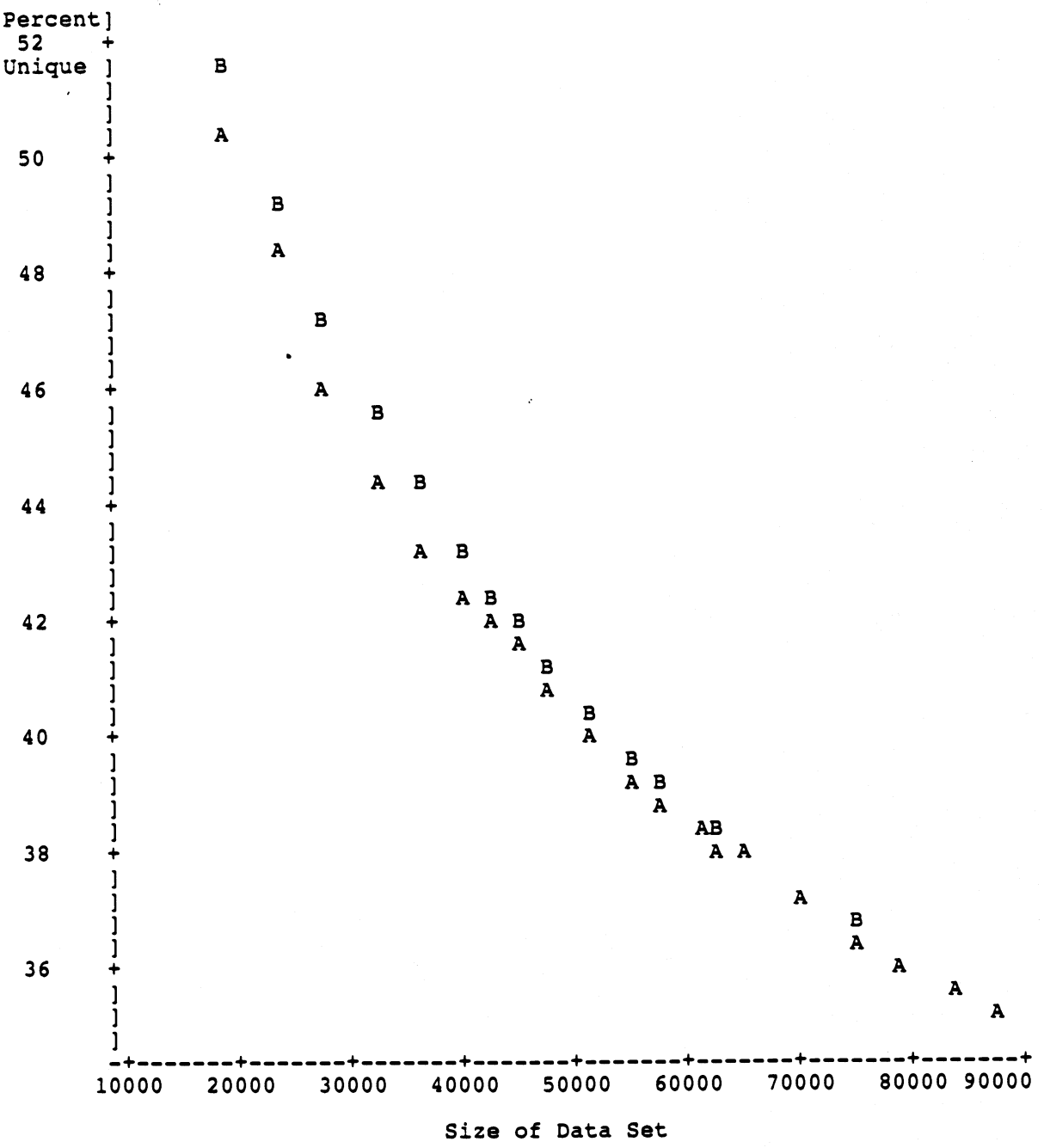
Tenure
Household Type
Household Class
Race
Ethnicity
Children
Marital Status
Veteran Status
Employment/Unemployment
Disability
Payment
Household Income
Social Security
Public Assistance
Other Income

Figure 1. Percent of Unique Households Versus Size of Data Set. 6 Variables.
A: Geographically Based Data Sets, B: Random Data Sets



NOTE: 6 OBS HIDDEN

Figure 2. Percent of Unique Households Versus Size of Data Set. 15 Variables. A: Geographically Based Data Sets, B: Random Data Sets



NOTE: 6 OBS HIDDEN

Figure 3. Percent of Unique Households Versus Size of Data Set. The symbols used in this figure represent the number of variables in the data set. A: 6 Variables, B: 10 Variables, and C: 15 Variables.

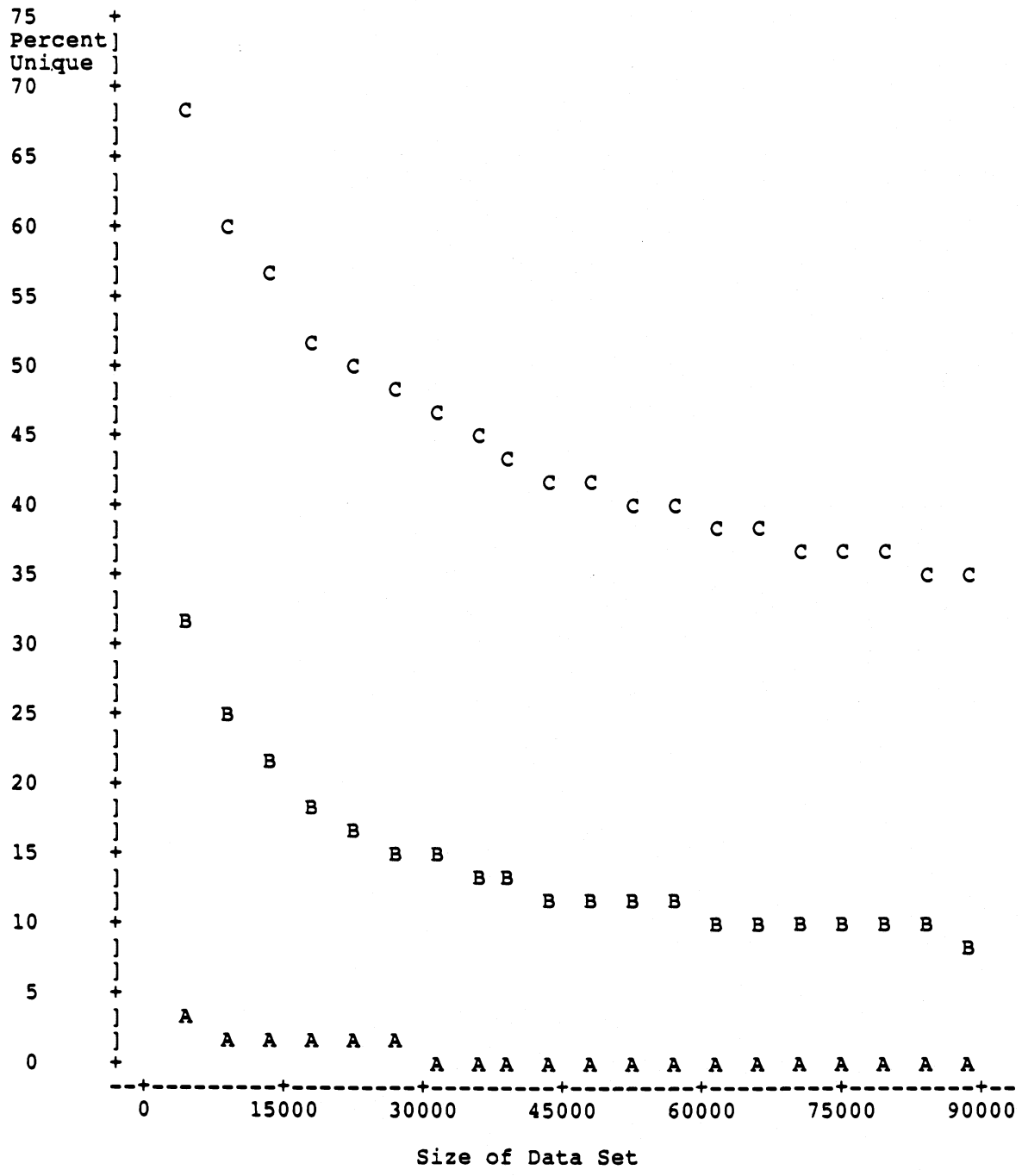


Figure 4. Derivative of Percent of Unique Households with Respect to Size of Data Set Versus Size of Data Set. The symbols used in this figure represent the number of variables in the data set. A: 6 Variables, B: 10 Variables, and C: 15 Variables.

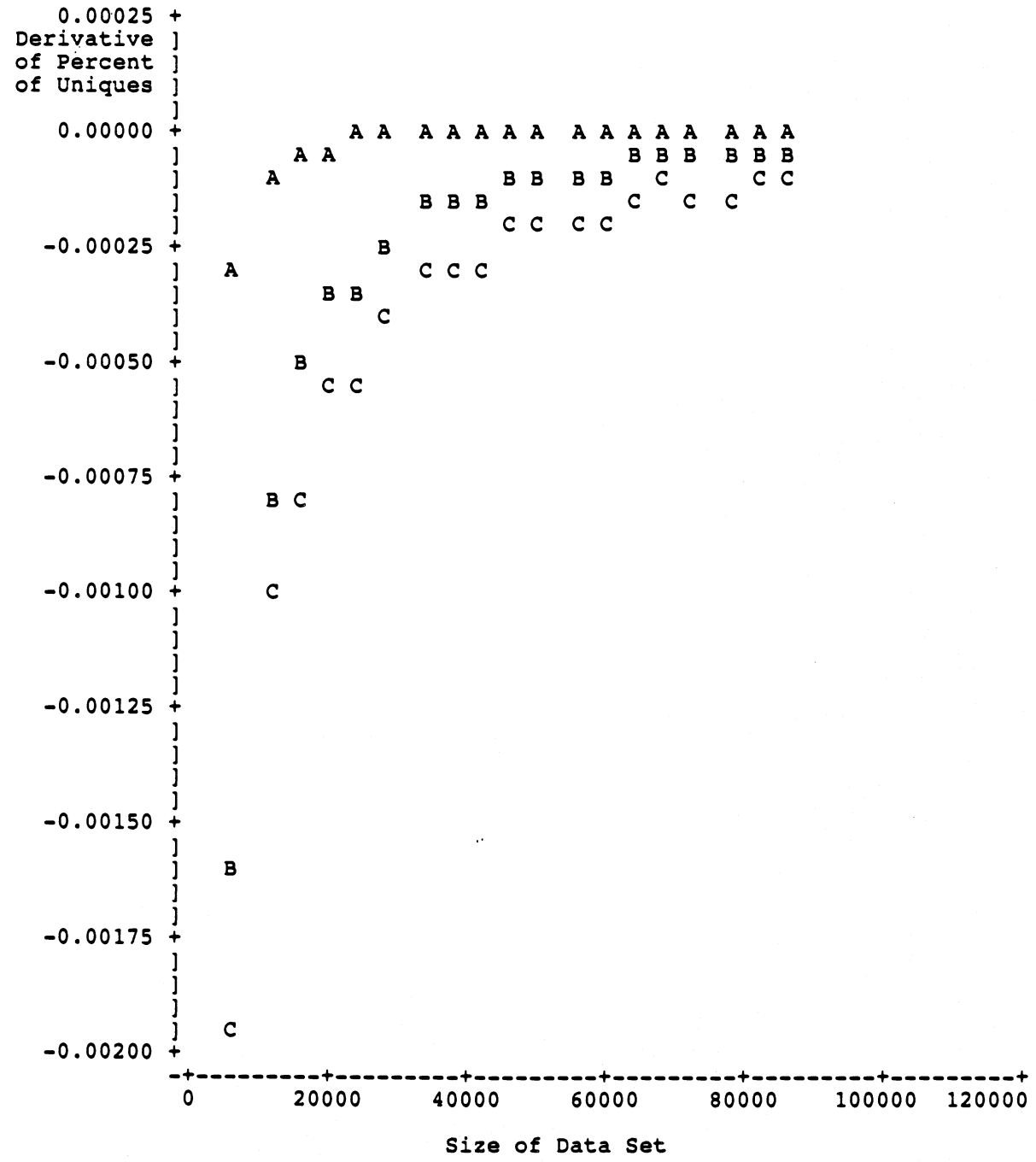


Figure 5. Percent of Unique Households Versus Size of Data Set. The symbols used in this figure represent the entropy of the variable "payment". A: Lowest Entropy of the Variable "Payment", ..., F: Highest Entropy of the Variable "Payment"

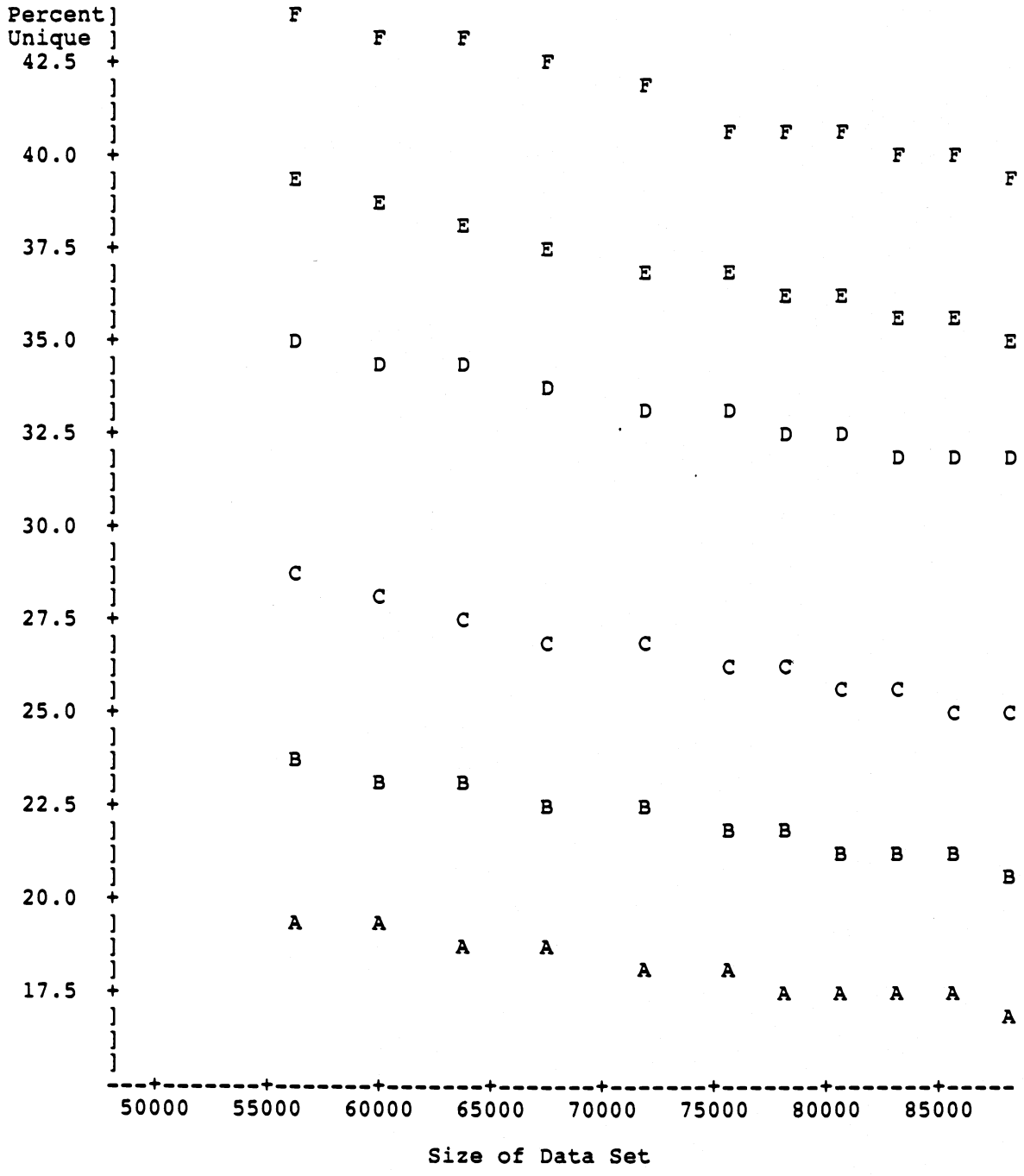
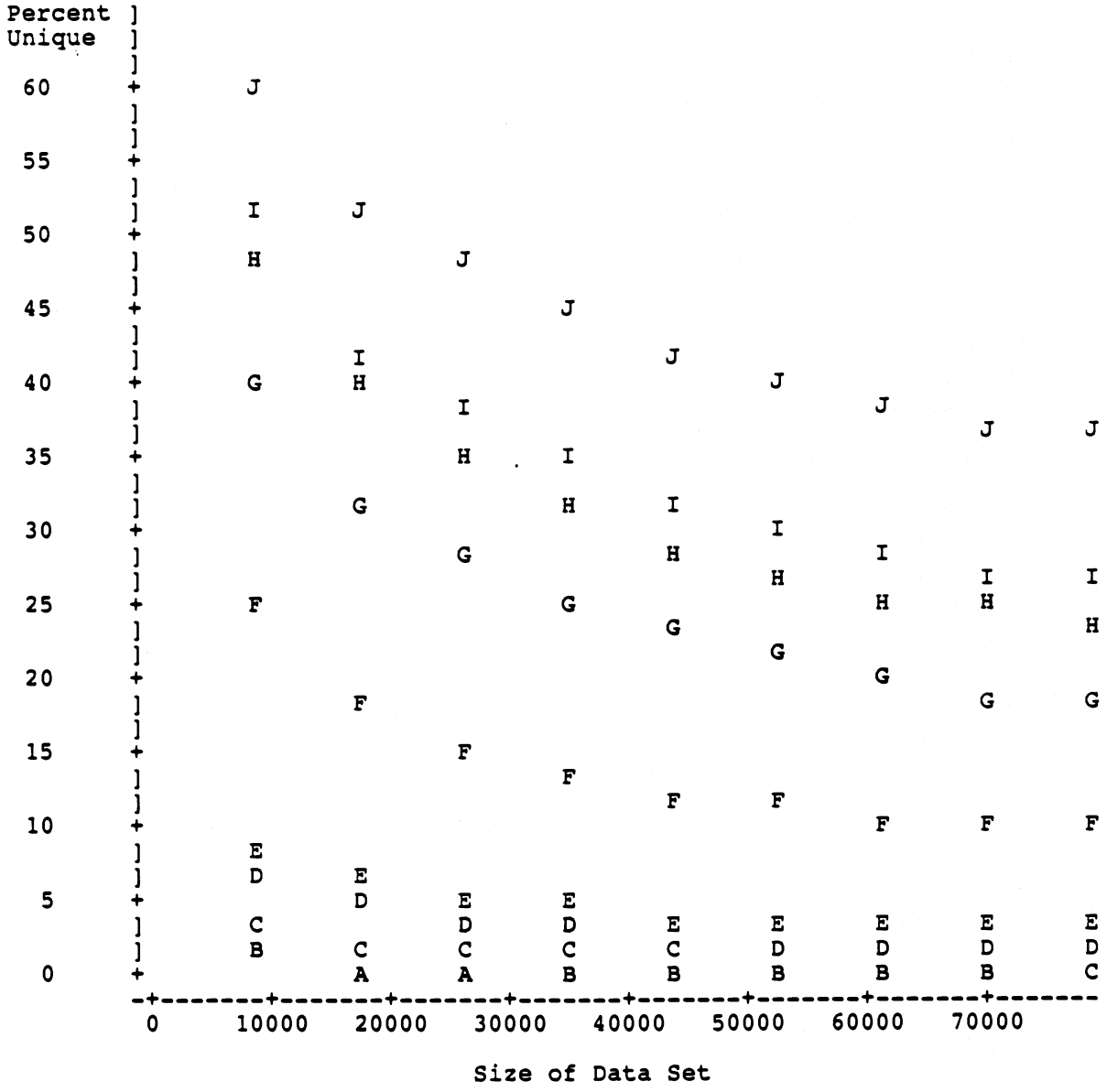
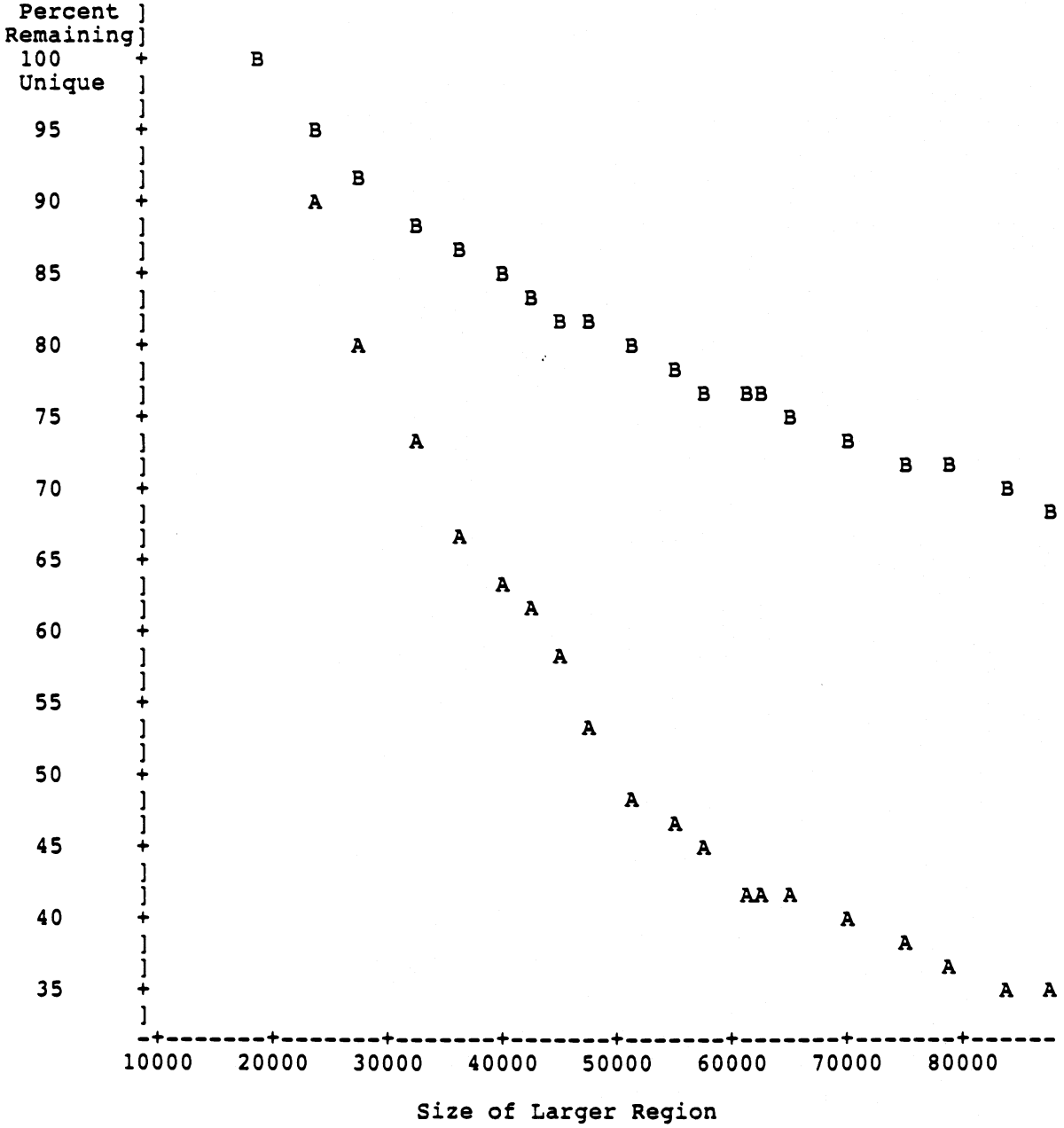


Figure 6. Percent of Unique Households Versus Size of Data Set. The symbols used in this figure represent the overall entropy of the original data set. A: Lowest Value of Overall Entropy, ..., J: Highest Value of Overall Entropy



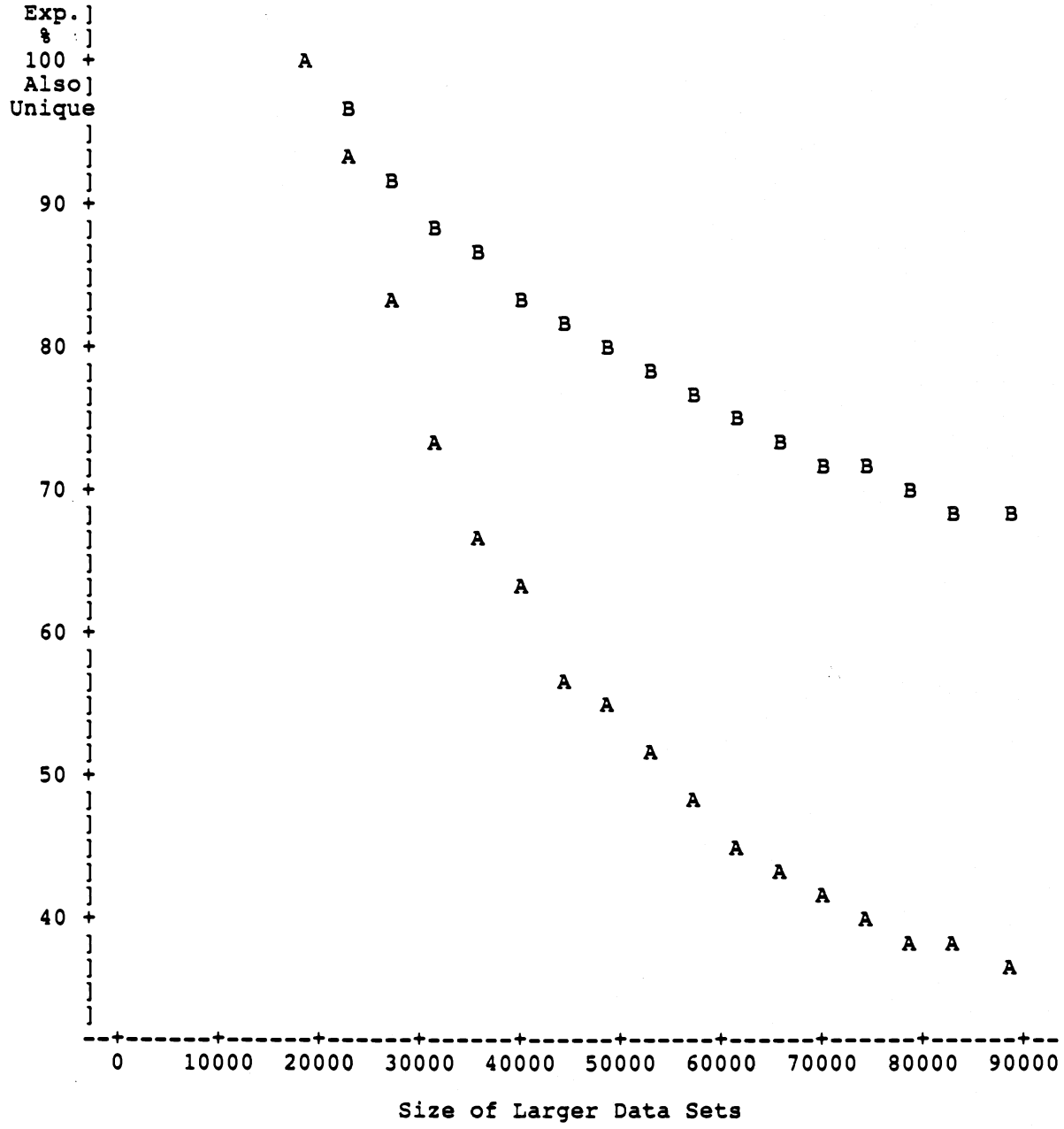
NOTE: 14 OBSERVATIONS HIDDEN

Figure 7. Percent of Uniques in Smallest Region that were also Unique in Larger Regions Versus Size of Larger Regions. The symbols in this figure represent the number of variables used. A: 6 Variables, B: 15 Variables



NOTE: 1 OBS HIDDEN

Figure 8. Expected Percent of Uniques in Smallest Data Set that were also Unique in Larger Data Sets Versus Size of Larger Data Sets. The symbols in this figure represent the number of variables used. A: 6 Variables, B: 15 Variables



NOTE: 1 OBS HIDDEN

Figure 9. Expected Percent of Unique Households in a Region that will Remain Unique if that Region is Increased in Size to Contain Approximately 4398 More Households than the Original Region Versus Number of Households in the Original Region. The symbols in this figure represent the number of variables used. A: 6 Variables, B: 15 Variables

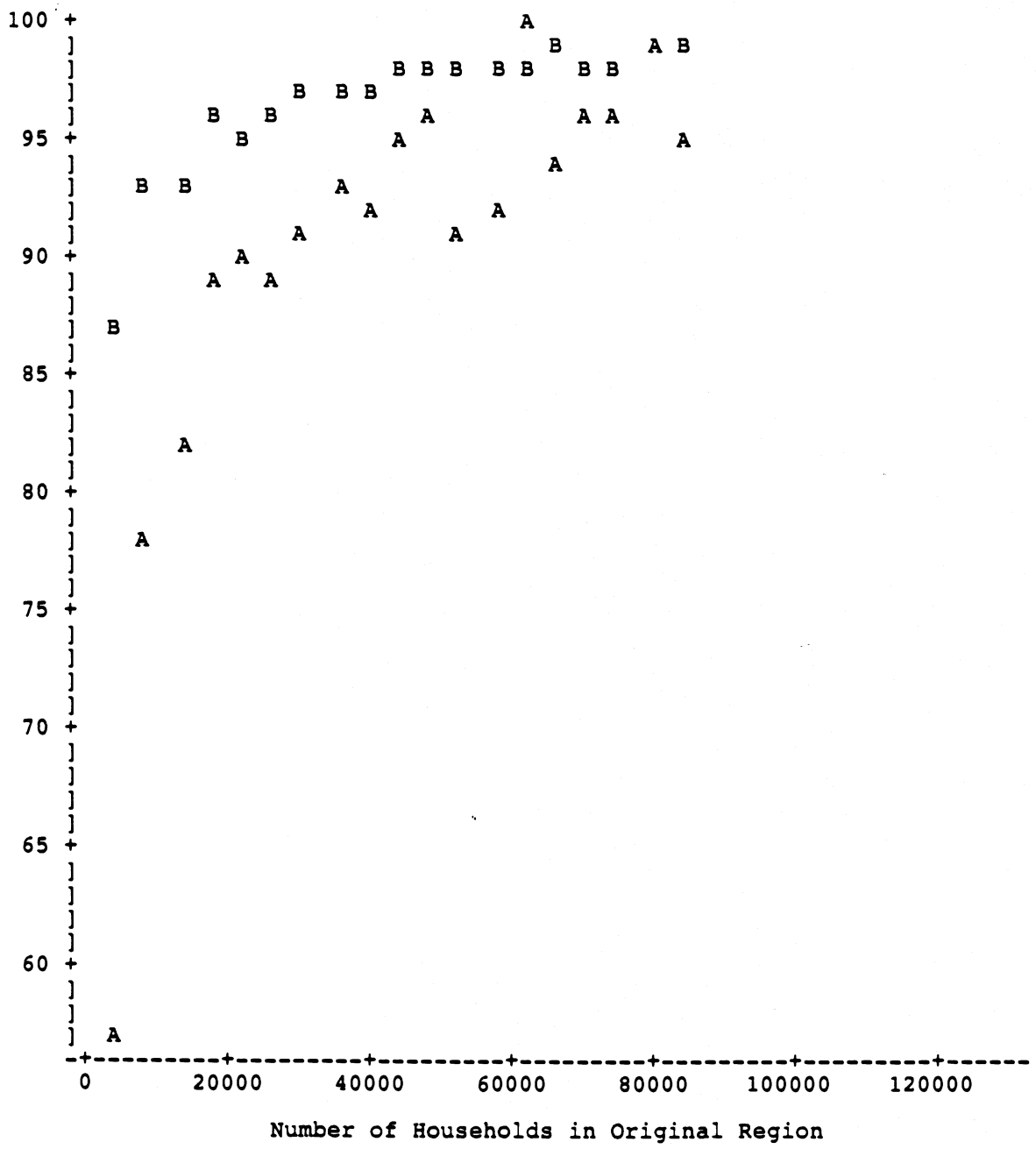


Figure 10. Percent of Uniques in Smallest Data Set that were also Unique in Larger Data Sets Versus Size of Larger Data Sets. The symbols in this figure represent the entropy of the variable "payment". A: Lowest Entropy of the Variable "Payment", ..., F: Highest Entropy of the Variable "Payment"

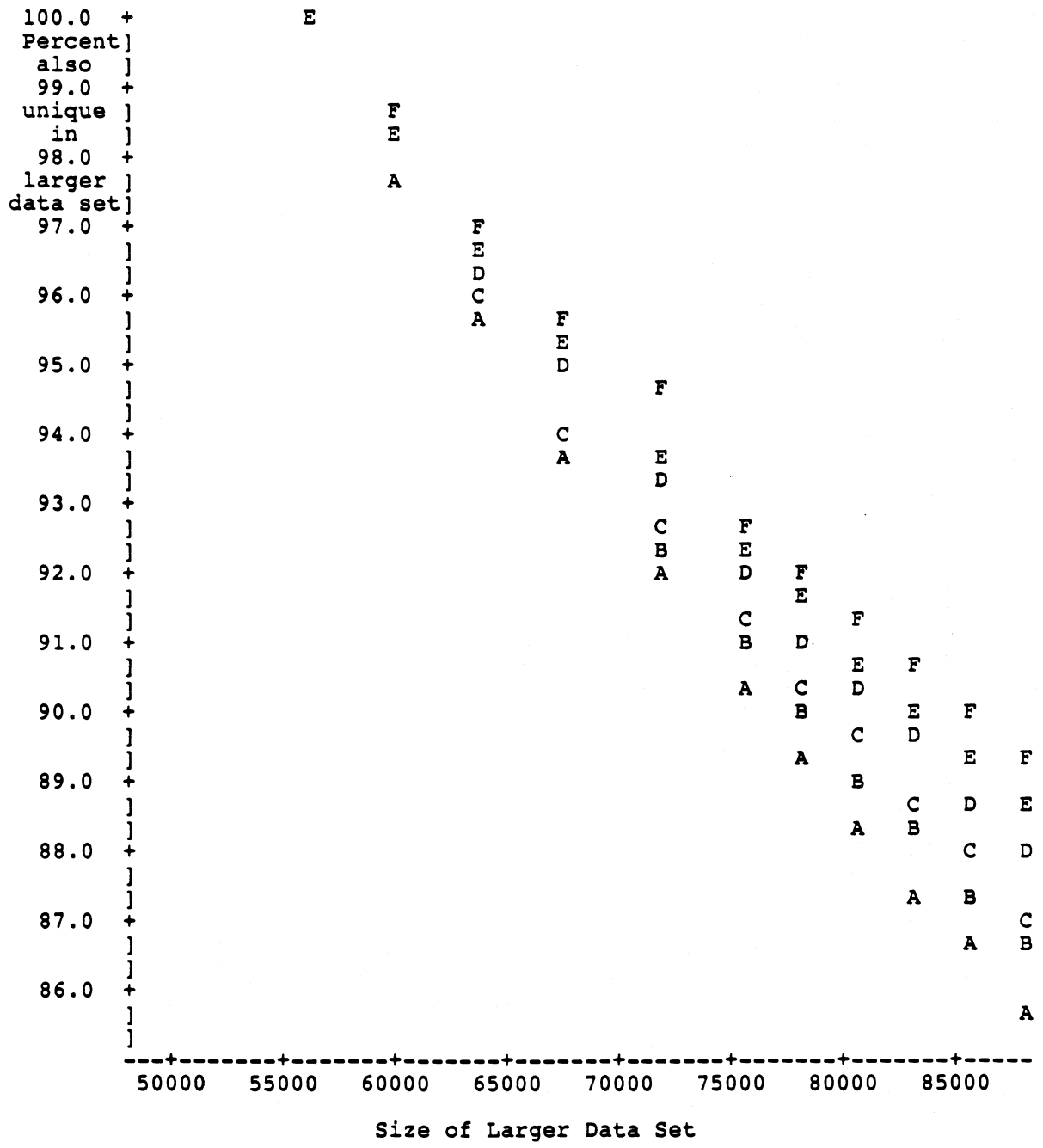
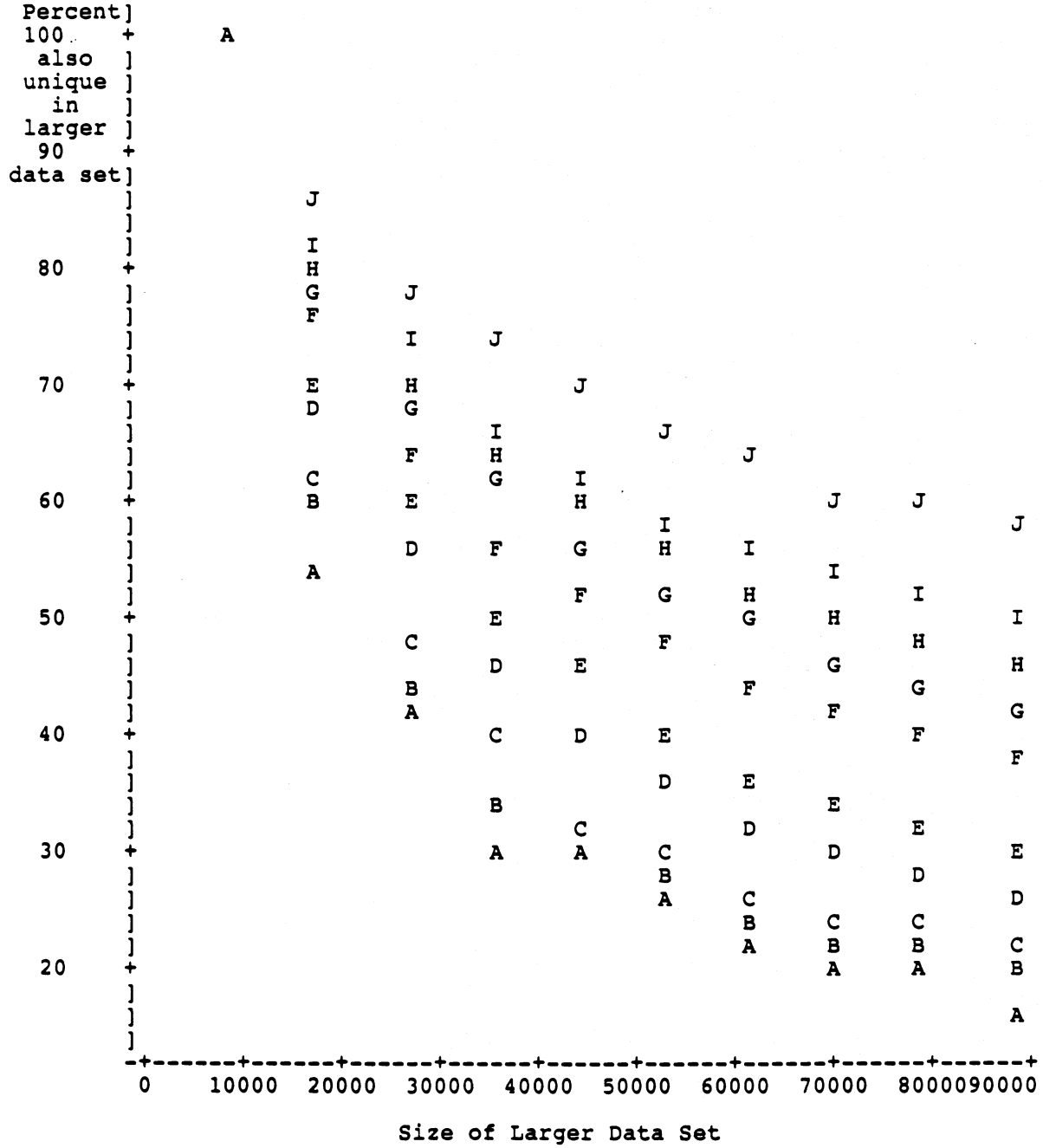
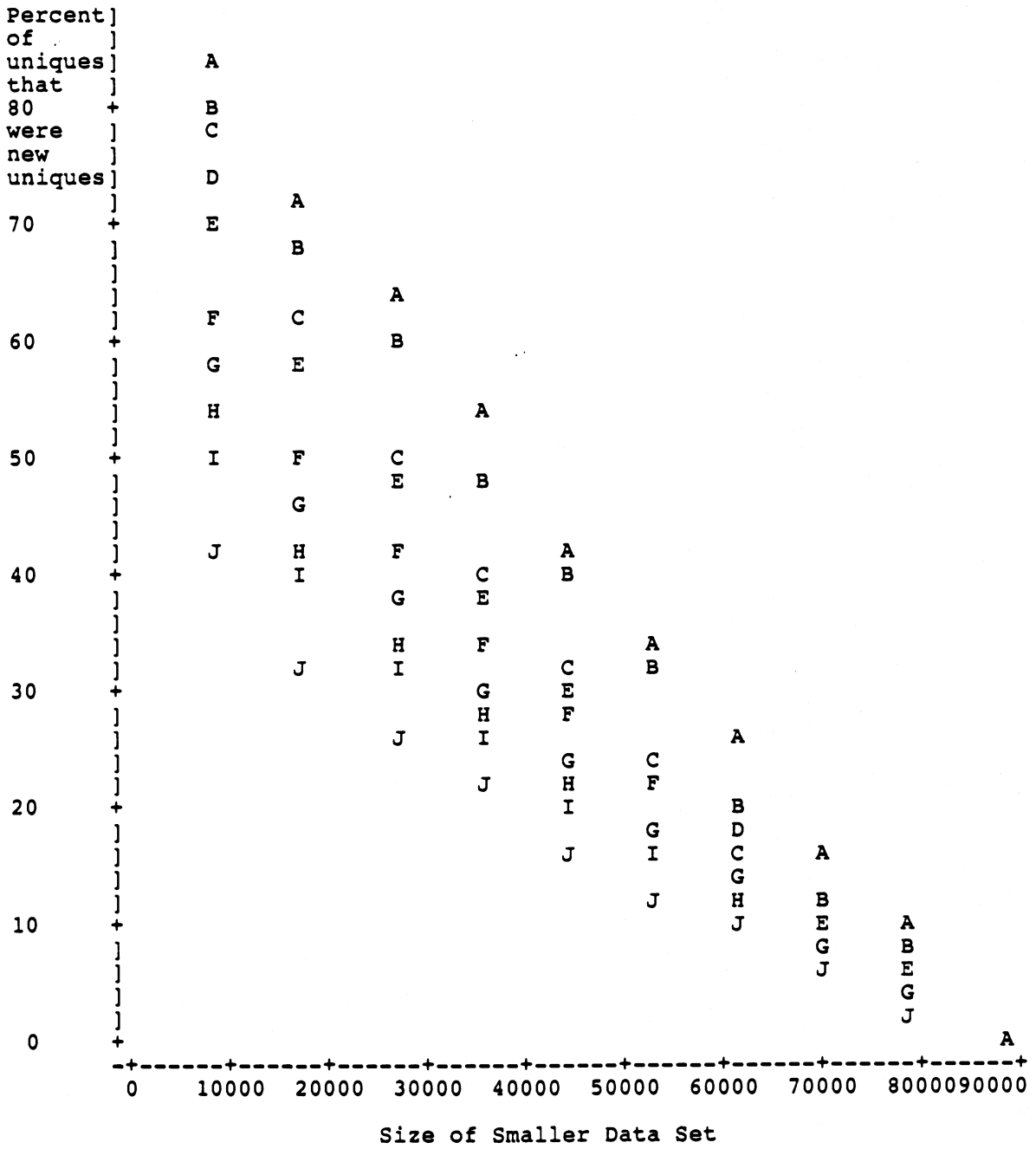


Figure 11. Percent of Uniques in Smallest Data Set that were also Unique in Larger Data Sets Versus Size of Larger Data Sets. The symbols in this figure represent the overall entropy of the original data set. A: Lowest Overall Entropy, ...; J: Highest Overall Entropy



NOTE: 10 OBS HIDDEN

Figure 12. Percent of Uniques in Smaller Data Sets that were New Uniques Versus Size of Smaller Data Sets. The symbols in this figure represent the overall entropy of the original data set. A: Lowest Overall Entropy, ..., J: Highest Overall Entropy



NOTE: 29 OBS HIDDEN

REFERENCES

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990), "Disclosure Control of Microdata," Journal of the American Statistical Association, Vol. 85, pp. 38-45. (An earlier version of this paper appeared in (1988), "Disclosure Control of Micro Data," Proceedings of the Bureau of the Census Fourth Annual Research Conference, pp. 181-192.)

Greenberg, B. (1990), "Disclosure Avoidance Research at the Census Bureau," Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, D.C., pp. 144-166.

Greenberg, B. and Voshell, L. (1990), "Relating Risk of Disclosure for Microdata and Geographic Area Size," Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, DC, to appear.

Voshell, L. (1990), "Estimating the Number of Population Uniques Using Information from a Sample," Statistical Research Division Report Series forthcoming.

Willenborg, L. C. R. J., Mokken, R. J., and Pannekoek, J. (1990), "Microdata and Disclosure Risks," Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, D.C., pp. 167-180.