United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research Division

RD Research Report
Number RD-97-01

March 1997

Comparison of 1993 Fruit Pesticide Use Survey Data and CALEPA Data

Mike Fleming

Comparison of 1993 Fruit Pesticide Use Survey Data and CALEPA Data. Mike Fleming, Survey Research Branch, Research Division, National Agricultural Statistics Service, United States Department of Agriculture, Washington, D. C. 20250-2000. March 1997. Research Report No. RD-97-01

ABSTRACT

In 1992, the National Agricultural Statistics Service (NASS) investigated using administrative data purchased from the California Environmental Protection Agency (CALEPA) to set estimates of the amount of pesticides applied in California for certain vegetables. The results proved to be somewhat disappointing, due to a high rate of rejection of the administrative data during the editing process. Follow-up analyses gave several reasons for the high rejection rate. A similar venture was tried again in 1993; this time for selected fruits grown in California. This report presents the results of research on the 1993 data. NASS surveys generally obtain more information from operators who make few applications of pesticides than is found in CALEPA reports. Both the CALEPA and the NASS data seem to be problematic in cases where multiple applications are made. It is impossible to link clearly the NASS sample with individual CALEPA records of operators who apply pesticides to more than one operation. The CALEPA data can be used in place of the survey data provided that proper quality control measures are followed to monitor its quality. It is recommended that computer assisted personal interviews with applicator/growers be used to access CALEPA files so sources of recording error can be eliminated at the time of the interview.

KEY WORDS

Administrative data; Pesticide Use; Chemical Use Survey; Computer Assisted Personal Interviews; Record Linkage; Statistical Process Control; Enterprise; Block; Site Location.

ACKNOWLEDGEMENTS

Jim Smith, Chris Messer, John Amrhein, and Dwaine Nelson provided very important technical assistance. Help in getting the enormous amount of data was given by Don Saboe, Doris Williams, Barbara Ison, and Perry Poe. Dale Atkinson supervised and edited the report.

CONTENTS

SUMMARY

The Survey Quality Research Section of the National Agricultural Statistics Service (NASS) examined the problems which arise when using fruit chemical use administrative data obtained from the California Environmental Protection Agency (CALEPA). This research was conducted at the request of the Environmental and Economics Survey Section of NASS. The CALEPA data are currently used in conjunction with NASS survey data to set estimates.

The problems first appear at the beginning of a four step process whereby records corresponding to the NASS sample are extracted from the CALEPA data. At the end of the process, a subset of the CALEPA data is produced. It is then submitted to the editing and summary systems from which various statistics related to the use of pesticides in the agricultural industry in California are analyzed and published by NASS.

A number of problems occurred when the 1992 vegetable CALEPA data were processed, and these created skepticism about the feasibility of using the CALEPA data for making reliable estimates. The initial investigation led by John Amrhein in 1993 [1] found three major problems with using the 1992 CALEPA data, namely:

1. Several NASS sample records did not correspond to any CALEPA grower record due to erroneous identification numbers.
2. Commodity classification and acreages did not always agree between the NASS and CALEPA data.
3. Many CALEPA identification numbers were associated with more than one operation, making these CALEPA records virtually impossible to use with NASS data without making significant personal inquiries to resolve the ambiguities.

This report is the product of another investigation into the same types of problems studied earlier. It was made after measures were taken to correct the problems cited in the Amrhein report. The current study addresses the use of the 1993 CALEPA data for targeted fruit crops. It was found that:

1. The same problems cited by John Amrhein were found to exist in the 1993 fruit data.
2. Erroneously recorded CALEPA ID's occurred about 10 percent of the time.
3. A unique mapping between NASS and CALEPA ID's did not exist in general. Therefore, the NASS sample at the enterprise level could not always be accurately identified with the CALEPA data.
4. An additional 20 percent of the NASS sample could be matched to the CALEPA data, if a record linkage program were used to compare the County Agriculture Commissioners' name and address files with the NASS sampling frame.
5. Not all operators in California submitted pesticide use reports to the State. Those who did not, tended to operate small acreages and use pesticides commonly sold for household use. For this portion of the population NASS may be able to obtain pesticide use information by means of its surveys.
6. The NASS and CALEPA data agreed reasonably well. When records could be accurately identified in the CALEPA data, the quality of the CALEPA data appeared to be excellent.
7. There was no substantial difference in the nature of the problems that arise when processing the CALEPA data for either the block or enterprise level records. The site location

identification numbers found in the block level records helped in identifying the correct records in the CALEPA data for this analysis.

Certain steps will need to be taken in order to use the CALEPA data more effectively. It is recommended:

1. That a computer assisted personal interview (CAPI) be used to reduce ID entry problems.
2. That a record linkage program be applied to the County Agriculture Commissioners' name and address files and the sampling frame used by NASS.
3. That the site location numbers be obtained at both the block and enterprise levels.
4. That a quality assurance program be considered to monitor the quality and processing of the purchased administrative data.
5. That the entire processing of the CALEPA data be performed in the California State Statistical Office.

INTRODUCTION

This report is a sequel to John Amrhein's study, "The Use of the California Environmental Protection Agency Pesticide Use Data by NASS" [1]. He assessed the quality of administrative data purchased from the California Environmental Protection Agency (CALEPA), because it was deemed to have potential in replacing the pesticide use data which otherwise would be collected by the National Agricultural Statistics Service (NASS). From his study of the 1992 CALEPA data, he identified three major problems which compromise the usefulness of the CALEPA data.

In California, every application of a pesticide in the agricultural industry is required to be reported monthly to the State. The County Agriculture Commissioners' (CAC) offices collect the reports and send tapes of the data to CALEPA. NASS purchases and uses these data instead of collecting the information with a survey. The objectives of this research into the quality of the CALEPA data are the same ones set by John Amrhein, except that he focused his study on the use of pesticides on vegetables. Here the focus is on the use of pesticides on fruits. This report presents an assessment of the quality of the CALEPA data, by comparing the CALEPA data to the data collected by NASS in its Fruit Chemical Use Survey.

Foremost among the concerns cited in the Amrhein report is the problem of not being able to obtain the correct CALEPA identification number (CALEPA ID). This study indicated that the problem exists for fruit and can be significantly minimized by using a record linkage program that matches the CAC name and address files to the sampling frame used by NASS. These name and address files are readily obtainable from the counties in California and offer an invaluable resource to the staff of the California State Statistical Office (SSO). The other two problems that John Amrhein identified, namely the problem of incompatible nomenclatures of the commodities and the problem of not being able to map NASS ID's to CALEPA ID's in a one-to-one fashion also occurred in the fruit data.

Every year, NASS conducts a survey of operators of agricultural enterprises in several states, including California. The scope of the survey follows an alternating pattern between vegetables and fruits from one year to the next. Thus, NASS collected data pertaining to vegetables in 1992 and to fruits in 1993. At the same time, applicators of pesticides in California must report to the State each use of a pesticide on an agricultural operation. That information is contained in the data which NASS obtains from CALEPA. Having survey data and administrative data concurrently for the same years provided an excellent opportunity to check the effectiveness of the NASS survey and to assess the quality of the CALEPA data.

In this study, information supplied by the growers and collected on NASS questionnaires was compared directly with the CALEPA data for successfully matched records. A successful match was not always achieved, since sometimes a NASS ID could not be mapped uniquely to a CALEPA ID. This problem reveals a particularly troublesome aspect of linking the NASS and CALEPA records. However, when a NASS survey data record was successfully matched and compared with the corresponding data found in the CALEPA data, the quality of the CALEPA data appeared to be excellent.

MAKING A SUCCESSFUL MATCH

The process of successfully matching ID's requires several steps. The enumerator must

first obtain the respondent's CALEPA ID during the NASS survey interview. This identification number is issued to an operator by a County Agriculture Commissioner's office for the county in which the operator registers. The number consists of eleven digits, with the last seven digits uniquely identifying the operator. It is this seven digit number which is tied to the NASS ID number. For a detailed description of the CALEPA ID, see Amrhein [1].

Appendix I shows the frequency distribution of CALEPA ID's found in the CALEPA data by county for fruit. The number of nut growers found in the CALEPA data is also given in Appendix I. The appendix also contains the counts of NASS ID's found in its 1993 sampling frame for fruit growers. The numbers in the column under nuts correspond to growers who reportedly applied pesticides only to nut bearing acres. The numbers in the column under fruit crops correspond to growers who reportedly applied pesticides to at least one target fruit crop. In Alameda County, for instance, 712 pesticide permits were issued by the Alameda County Agriculture Commissioner's office. They were issued to operators who had intentions of applying pesticides to their acreages. In this county, the 1993 CALEPA fruit and nut data contained records of 18 growers who used pesticides on any targeted fruit bearing plant. Some of the 18 growers may also have applied pesticides to nut crops. Besides these, there were three other operators who did not apply pesticides to any fruit crops but who had applied pesticides to nut crops.

More than one CALEPA ID may be associated with a specific NASS ID. Such a circumstance can happen if an operator selected for a NASS survey manages several operations, perhaps including his own, each of which had distinct CALEPA ID's. As Table 1 indicates, five cases were found, for instance, in which a NASS ID number was associated with two CALEPA ID's, and one case was found where the NASS ID was associated with three CALEPA ID's.

Conversely, two different operators selected by NASS can supply the same CALEPA ID number if they hire the same managing company. Though apparently not too common, the association of multiple NASS ID's with one CALEPA ID does occur, as we see from Table 2. Most of the time a NASS ID is linked uniquely with a CALEPA ID number. But, as Tables 1 and 2 clearly show, there are instances in which the mapping between the sets of NASS and CALEPA ID's is not one-to-one.

One connection between a NASS ID and a CALEPA ID occurs at the time of an interview, when the operator reports the CALEPA ID of the applicator who applied pesticides to his operation. The applicator may be the individual operator himself, a commercial applicator who was hired by the operator, or someone else like a neighbor, friend, or relative. Another connection between NASS and CALEPA ID's can be made with the use of a record linkage program which matches the CAC name and address files to the sampling

Table 1. Frequency of the Number of Valid CALEPA ID's for a Given NASS ID.

Number of Valid CALEPA ID's per NASS ID	1	2	3	4
Frequency	763	5	1	0

Table 2. Frequency of the Number of NASS ID's for a Given Valid CALEPA ID.

Number of NASS ID's per Valid CALEPA ID	1	2	3	4
Frequency	747	11	1	1

frame used by NASS. Both ways of connecting the NASS and CALEPA ID's should give consistent relationships.

If the relationships were one-to-one, it would be possible to go from a NASS ID to a CALEPA ID using the survey data and then return to the original NASS ID using the results of a record linkage program. The circuit would follow, in this ideal case, the schematic diagram shown in Figure 1. However, the statistics shown in Tables 1 and 2 indicate that the mapping between the NASS and CALEPA ID's is not one-to-one in general. In fact, as the illustration of a multivalued mapping depicted in Figure 2 shows, there are cases in which a single CALEPA ID can be linked not only to several NASS ID's but also to several operators who are not represented in the NASS sampling frame.

Because of the multivalued mappings, the records in the CALEPA data could not be accurately identified with elements of the NASS sample. Additional identifying information about them is needed for distinguishing records in the CALEPA as belonging to an operator selected for a survey. The worst cases involve operators who apply pesticides to more than one operation. Without NASS having adequate identifying information, it is impossible to associate CALEPA records with elements of the NASS sample.

Besides the complications which a multivalued mapping causes, a significant amount of information cannot be extracted from the CALEPA data due to other reasons such as typographical or format errors in recording the CALEPA ID's during the survey and the reporting of a wrong CALEPA ID by the respondent. It was found by inspecting the original NASS questionnaires and comparing the recorded CALEPA ID's to the CAC name and address files that illegible penmanship and recording errors commonly occurred.

There was one case, for example, in which a grower reported to NASS that he applied pesticides to over 19,000 acres of a crop. The CALEPA data indicated applications of pesticides to less than 100 acres. By using the CAC name and address files, it was discovered that a second CALEPA ID was not recorded on the NASS questionnaire. Once both CALEPA ID's were found, the acreages agreed.

After repeated attempts to deduce the correct CALEPA ID by inspection, a record linkage program was used to match records between a subset of the 1995 California fruit sampling frame and the 1995 CAC name and address files. The 1995 files were matched rather than the 1993 files in order to benefit the

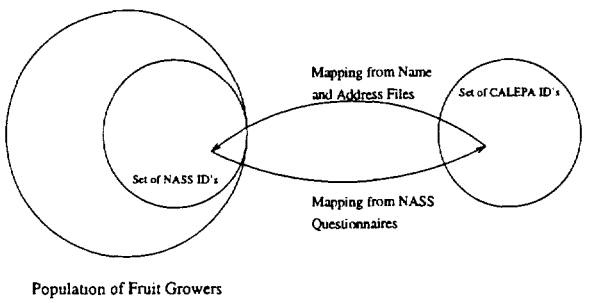

Figure 1. It is preferred to have a one-to-one mapping between NASS and CALEPA ID's, in order to extract the correct records from the CALEPA data.

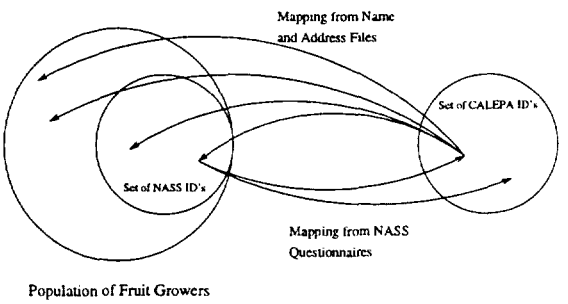

Figure 2: Typical mapping between NASS and CALEPA ID's is multivalued. Additional identifying information is needed to correlate records in the CALEPA data with the NASS sample.

1995 survey work [3]. The match was performed using the *Automatch* software package on the 1995 extracted sample. Some of the matches contained operators who were also selected for the 1993 survey. There were 195 such matches with the 1993 data of which 41 were not assigned a valid CALEPA ID on the NASS questionnaire. These 41 additional matches picked up by the record linkage would otherwise have been lost for the research on the 1993 fruit data. Other results produced by the record linkage program revealed some of the typical mistakes made in recording the CALEPA ID's during the 1993 survey.

Obvious recording and typographical errors in the CALEPA ID's can be seen in Table 3. For the computer to match the CALEPA ID's found in the CALEPA data with those in the NASS survey data, their digits must be absolutely identical. We see in the first example of Table 3 an instance in which the CALEPA ID recorded by NASS bears no resemblance to the one for that operator in the name and address files. The survey procedures stipulate that the CALEPA ID number used by the applicator of the pesticide should be recorded on the questionnaires. In this example, one of the CALEPA ID's may belong to an operator who applied pesticides to more than one operation and the other to the respondent. In any case, it is impossible to tell which ID number is the right one to use.

Table 3. Examples of Incorrect CALEPA ID's as Recorded by NASS.

CALEPA ID Recorded on NASS Questionnaire	Correct CALEPA ID
1040688	5401356
56P0039	56P0038
35C0006	56C0006
0350616	350616
0350709	3500709
510027A	5100027
511042A	5101042
0393044	3903044
3315002	331500Z
430609R	430669U
3940330	3340330

Since every record in the CALEPA data is identified by a CAC permit number it is possible, as discussed already, to match a NASS ID with a CALEPA ID by matching the names of the operators selected for a survey with the names found in the CAC name and address files. This method would guarantee the correct assignment of ID's, provided that the permit numbers in the CAC name and address files are always consistent with the ID's found in the CALEPA data. Although the permit numbers in the CAC name and address files and the CALEPA data are supposed to be the same, a few inconsistencies were discovered. For instance, the permit number 56PO102 contained the letter *O* in the name and address files, rather than the zero that it should have. Even though the frequency of inconsistencies among ID's between the two files may be very small, a periodic review of the consistency of the ID's would be part of a quality assurance program for the NASS Chemical Use Program.

While everyone who applies a pesticide on an agricultural enterprise in California is required to report each application to the California Environmental Protection Agency, some do not comply with the provisions of the law. This lack of compliance results in incompleteness in the CALEPA data. Further incompleteness results when registered users do not report every application of a pesticide on their operation. The extent to which the set of CALEPA data is not representative of the population of pesticide applications for fruit growers can be inferred by comparing the CALEPA data with the NASS survey data.

The NASS survey data is divided into two parts - block and enterprise level data. The enterprise level obtains information about the use of pesticides that may have been made on an entire operation. The block level obtains detailed information about the use of pesticides found on an operation for only one specific crop, possibly consisting of several non-contiguous sites. Block level samples were selected as replicates 2, 3, and 4, and enterprise level samples were selected as replicates 1, 5, and 6. During the processing of the CALEPA data, records in the NASS survey data were deleted when no match could be made with any record in the CALEPA data. Starting with records for 961 operators (including both block and enterprise level samples), the number of NASS data records decreased at each stage of the process. By the end of the process, records for 683 operators in the NASS sample remained. The breakdown of total records and successful matches is summarized in Table 4 and Appendix II. At the block level, 218 ID's (75 percent) could be successfully matched with the CALEPA data; at the enterprise level, 465 ID's (70 percent) could be matched.The characterization of the 278 cases that did not survive the processing is indicated in Figure 3.

Sometimes, a match could not be made because NASS received the wrong CALEPA ID from the respondent. Other times, the CALEPA ID was valid but the operator did not report the use of pesticides to the State for all of his crops for that year. It is often difficult to characterize all the reasons for the failure to make successful matches between the NASS sample and the CALEPA data.

Up to 10 percent of the operators sampled by NASS who had used pesticides did not

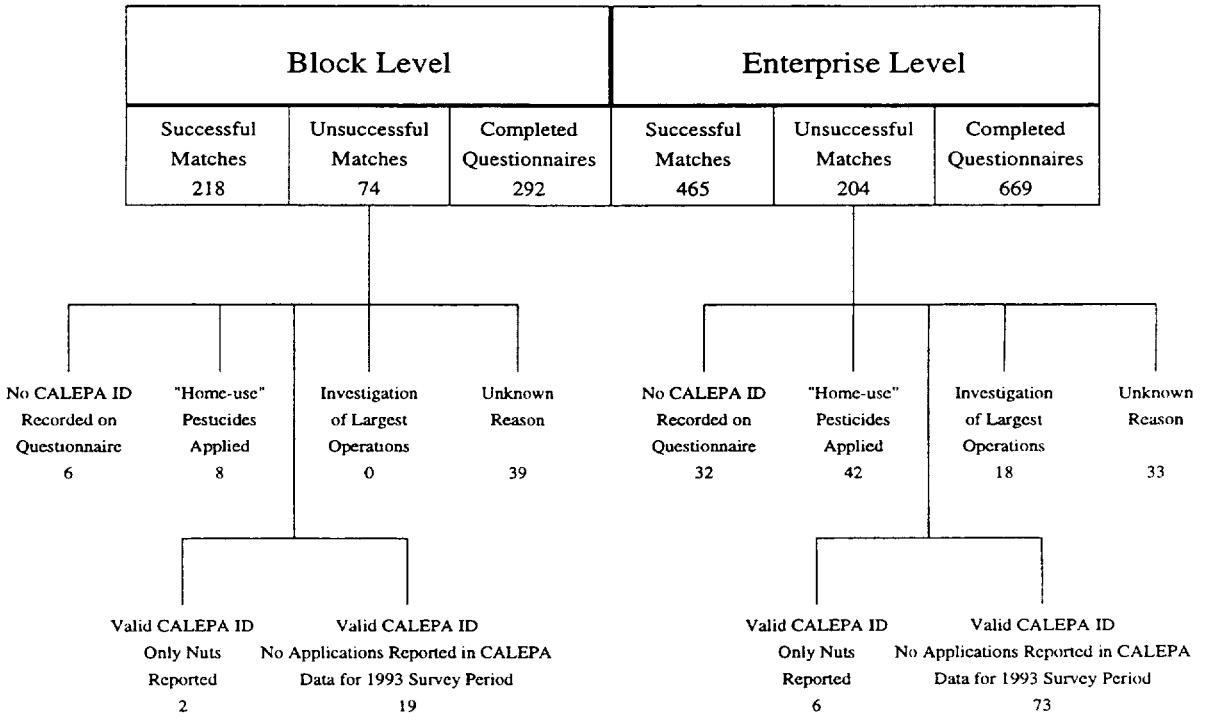

Figure 3. All operators in the NASS sample who indicated the use of a pesticide and had no records in the CALEPA data are characterized in various ways as the diagram shows. Categories of operators who had a valid CALEPA ID are indicated; in all other categories, the validity of the CALEPA ID's could not be determined.

Table 4. Number of Matches Between NASS and CALEPA ID's by Replicates

Replicate	Number of NASS ID's with Pesticide Data	Number of NASS ID's Matched with CALEPA Data
1	480	334
2 (Apples)	75	34
3 (Grapes)	131	114
4 (Oranges)	86	70
5	127	87
6	62	44
Total	961	683

obtain a permit to use pesticides. Others evidently had a valid CALEPA ID but may have applied pesticides that were purchased at a retailer for "home use" and failed to report the applications of these to the State. Some of the products reported in the NASS survey can be purchased in small quantities at a hardware store. Five of the most common of these are listed in Table 5. Typically the operators who applied pesticides purchased from a retailer tend to operate relatively small acreages. The sizes in terms of bearing acres of the top ten operators (for each level) found in the NASS sample who did not report any application to CALEPA are listed in Table 6.

The remaining operators in the NASS sample which could not be matched to any record

Table 5. Pesticides Probably Bought from a Retailer

CALEPA Product Code	Product Name
7102	Micro Flo Sulfur
4093	Roundup Herbicide
1278	Saf-T-Side for Grove Trees
1190	Safer Agro-Chem's Insecticidal Soap
7118	Dusting Sulfur

Table 6. Sizes of Ten Largest Operations in NASS Sample Which Did not Report a CALEPA ID by Level.

Block Level (bearing acres)	Enterprise Level (bearing acres)
55	300
30	110
21	100
19	67
10	55
9	48
8	47
7.5	40
5	35
4.5	25

in the CALEPA data were probably cases in which NASS was given an incorrect CALEPA ID. The ID's for the 18 largest operators in this group were examined, but the correct CALEPA ID's could not be found. Some of them were operations of considerable size. A list of the acreages of the top ten for each level is shown in Table 7. On every one of them, at least one registered pesticide had been used. Reports of the use of pesticides must have

Table 7. List of Sizes of Top Ten Operators in NASS Sample Which Could Not be Matched to CALEPA Data by Level.

Block Level (bearing acres)	Enterprise Level (bearing acres)
2380	1730
230	854
205	611
200	500
200	470
160	437
160	400
125	381
92	307
83	300

been submitted for them and, undoubtedly, are contained somewhere in the CALEPA data. Perhaps the operators had given someone else's CALEPA ID instead of their own.

For example, an instance was found in which an out-of-state insurance company owned an operation and had hired someone to manage it. The hired manager in turn also managed three other operations. None of the CALEPA ID's that he supplied to NASS could be matched to the CALEPA data. It is improbable that so many large operations never submitted any report of pesticide use to the State. Possibly, the manager did not report to NASS the CALEPA ID of the land on which the pesticides had been applied, but instead supplied the CALEPA ID's of other operations under his care. In the 18 cases that were studied, the CALEPA ID's supplied to NASS were probably incorrect.

There were other instances in which a valid CALEPA ID was given by a respondent, but the matching CALEPA records contained data for the use of pesticides on only bearing nut acres or had dates of application that were not within the scope of the 1993 survey. Each CALEPA record corresponds to one application of a specific pesticide. There are other reasons which also could explain the failure of making a successful match. The ones that seemed important enough to consider appear in Figure 3.

The elements in the 29 percent of the NASS sample at both levels that could not be matched to any record in the CALEPA data fall into three basic categories, the sizes of which appear in Table 8. The use of a record linkage program could be applied to the sampling frame and the CAC name and address files to improve the chances of making a successful match for the 9 percent of the NASS sample assigned an incorrect CALEPA ID. For the remaining 20 percent of the unmatched portion of the NASS sample, a special survey could be conducted to collect the necessary information from the population of fruit growers who do not submit pesticide use reports to the State.

Table 8. Summary of the Success in Matching NASS ID's with CALEPA Data (Both Levels)

Did Not Report to CALEPA	95	10%
Incorrect CALEPA ID	83	9%
Valid CALEPA ID But Data Not in Scope for 1993 Survey	100	10%
Successful Matches	683	71%
Completed Questionnaires	961	100%

MATCHING RECORDS

All chemicals registered by the United States Environmental Protection Agency are deemed to be pesticides, but some chemicals applied in the agricultural industry facilitate the application or performance of a pesticide. These additives which change the physical properties of a mixture are reported to the CALEPA, but are not included in the NASS Chemical Use Survey. Since NASS does not collect information about them, they need to be omitted from the CALEPA data when it is matched to the NASS sample. Accordingly, 123,313 records pertaining to these chemicals were omitted from the 1993 CALEPA data. The original number of records in the CALEPA data and the number of records after the matching process for each commodity appears in Appendix III.

A major difference arises at this point between the block and enterprise levels of the survey in the ability to extract CALEPA records according to the period of cultivation. Records at the block level contain the date of application of the pesticide. Since this infor-

mation is not available for enterprise level sample records, NASS uses a less direct method to extract the appropriate CALEPA data for these records. This method depends on the respondent's recollection of the dates of harvest of the previous and current crops. These dates are known as the beginning and ending dates. All records for which the dates of application lie between the beginning and ending dates for a particular crop are extracted from the CALEPA data. Therefore, the success of extracting records for the enterprise level operators from the CALEPA data depends not only on using the correct CALEPA ID, commodity name, and pesticide name, but also on using the correct beginning and ending dates.

Not all of the block level dates of application obtained by NASS agreed with the CALEPA data. Sometimes, a date of application in the NASS data for a given operator and commodity was later than the last date of application reported by the applicator in the CALEPA data. Other times, all of the NASS block level dates of application were found to be later than any entry in the CALEPA data. These partial and complete disagreements may have been caused by operators who did not always report their use of pesticides to the State. However, they could also have been caused by erroneous dates. There were many dates found in the NASS block level data that did agree with the CALEPA data. These provided an excellent benchmark for testing a computer program written to determine, independently of the survey data, the beginning and ending dates based solely on the CALEPA data.

This computer program was written to salvage for enterprise level records CALEPA data which would have been lost had the matching been performed on erroneous beginning and ending dates found in the NASS survey data. The benefit of using the computer program is indicated by Table 9. With its use, 49,343 records, the sum from the shaded area of the table, which otherwise would have been lost were saved and admitted for analysis.This approach should not be needed for future surveys, since the problem of beginning and ending dates has been addressed in the machine edit.

Table 9. Success of Using the Reported Beginning and Ending Dates at the Enterprise Level

Condition of Beginning and Ending Dates	Number of NASS ID's	Number of CALEPA Records
Correct	29	27,787
Inconsistent	409	37,605
Missing	37	11,738
Total	465	77,130

COMPARISON BY TREATMENT ACRES

CALEPA collects pesticide use reports for administrative purposes. The period over which permits are issued is the calendar year, which does not generally conform to the crop cycle. NASS on the other hand is interested in obtaining information about the use of pesticides beginning on the day after harvest of the previous crop and ending on the day of harvest of the current crop. For some crops, this period may span two calendar years. Correlating the beginning and ending of pesticide use in the CALEPA data with the reported beginning and ending dates in the NASS Chemical Use Surveys poses an additional complication. If it is not done correctly, then the cumulative number of acres to which pesticides are applied as reported in the CALEPA data will not agree with the NASS data. We will refer to the number of cumulative acres to which pesticides are applied as *treatment acres*. For example, if pesticides are applied to ten acres five times, then this would represent 50 treatment acres

Table 10. Difference in Treatment Acres Between NASS and CALEPA Data at the Block and Enterprise Levels.

Level	Name	Counts	Minimum Difference	25th Percentile	Median Difference	75th Percentile	Maximum Difference
Block	Apple	17	-575	-119	-18	13	1,966
	Grape	85	-2,307	-71	10	68	1,854
	Orange	53	-1,704	-26	31	68	435
Enterprise	Apple	35	-772	0	25	356	5,699
	Apricot	37	-2,304	-4	12	90	1,576
	Avocado	12	-3,629	-36	-5	28	420
	Cherry, Sweet	29	-806	-64	0	43	592
	Lemon	38	-6,489	-348	-10	57	2,964
	Date	10	-1,936	-14	28	350	1,567
	Fig	6	-8	0	386	2,980	3,898
	Grapefruit	17	-12,258	-122	-5	8	115
	Grape: Raisin	53	-1,616	-37	28	178	6,764
	Grape: Table	22	-9,265	-192	20	950	2,360
	Grape: Wine & Juice	43	-9,540	-90	20	158	19,147
	Kiwi Fruit	18	-29	-2	1	10	135
	Nectarine	40	-305	-10	24	94	1,234
	Olive	19	-1,328	-54	-4	74	3,896
	Orange	80	-19,861	-188	1	108	2,255
	Peach	83	-1,012	-2	16	112	1,903
	Pear	31	-1,641	6	160	1,218	5,427
	Plum	55	-1,067	-4	3	32	2,317
	Prune	54	-4,410	0	35	120	2,402
	Tangerine	12	-4,632	-55	4	42	134

Difference in treatment acres is the number of treatment acres reported in NASS data minus treatment acres reported in CALEPA data.

for each active ingredient in each pesticide product.

Table 10 shows the differences in the number of treatment acres between the NASS and CALEPA data for matched ID's by commodity at the block and enterprise levels. The values of the differences shown here were obtained by subtracting the number of treatment acres reported in the CALEPA data from the number of treatment acres computed from the NASS data. For example, if 9,000 treatment acres were reported in the NASS data and 100 treatment acres were reported in the CALEPA data, then the difference would be 8,900 treatment acres. A positive value implies that more treatment acres were reported to NASS, and a negative value implies that more treatment acres were reported to CALEPA.

Among the 34 records that could be matched between the NASS and CALEPA

data for apples at the enterprise level, the range of the differences in treatment acres extended from -772 to 5,699 with a median of 25. The medians listed in Table 10 suggest that slightly more treatment acres appear in the NASS data than in the CALEPA data. When using descriptive statistics like the median, it must be kept in mind that they describe the aggregate characteristics of a sample. They shed little light on the origin or extent of extreme values.

LARGE OPERATIONS

In the NASS data, the number of treatment acres can be easily computed. At the enterprise level for the 1993 Fruit Chemical Use Survey, it was sufficient for the respondent to report the number of times that a pesticide was applied rather than have the enumerator record the same information relating to each application over again. In the CALEPA data, on the other hand, the number of treatment acres corresponds only to the number of entries in the data. In Figure 4, the cumulative distribution of the number of treatment acres as recorded in the CALEPA data is plotted with circles. For each CALEPA point, the corresponding NASS point is plotted with a dot. The disparity that exists between the NASS and the CALEPA data, in terms of the number of treatment acres, becomes apparent when it is displayed graphically.

The divergence between the NASS and CALEPA data in Figure 4 shows that sometimes large differences between the NASS and CALEPA data do occur. The largest 48 differences were investigated by inspecting the questionnaires. Originally, 50 cases were chosen, but two questionnaires could not be found. The observed differences for these operations could be explained essentially by five reasons. The frequency for each reason is given in Table 11.

SITE LOCATION NUMBERS

At this point, we cannot tell whether the differences in treatment acres may be caused by the presence of operators whose records cannot be accurately identified with the NASS sample or by the burden of collecting too much information. To answer this question we will refer to the block level data again because these data contain important, additional detail. Recorded in the block level records is the site location identification number, which the enterprise level records do not have. An operation on which pesticides are applied receives not only a CALEPA ID, but also a site location identification number for every field receiving an application. This number, together with the CALEPA ID number and names of commodities, makes it possible to match the CALEPA and NASS block level data down to a particular field. However, a proper match between the CALEPA and NASS data at the block level depends on the accuracy of the site location ID given to NASS.

Out of 218 successful matches on ID number at the block level, 113 had the correct site location ID. By inspecting the incorrect site location ID's and comparing them with the those listed in the CALEPA data, it was possible to salvage 68 of them. Examples of typical mistakes that were found in the site location ID's are listed in Table 12. Salvaging the 68 resulted in 181 NASS block level records qualifying for analysis. Surprisingly, no records were found in the CALEPA data for 15 of these 181 good matches. It appears that the operators of those fields did not submit reports of their use of pesticides to the State. Ultimately, 166 elements of the NASS sample at the block level could be matched to the CALEPA data on a record by record basis.

There is also additional information in the block level data indicating whether a pesticide was applied by a commercial applicator, the

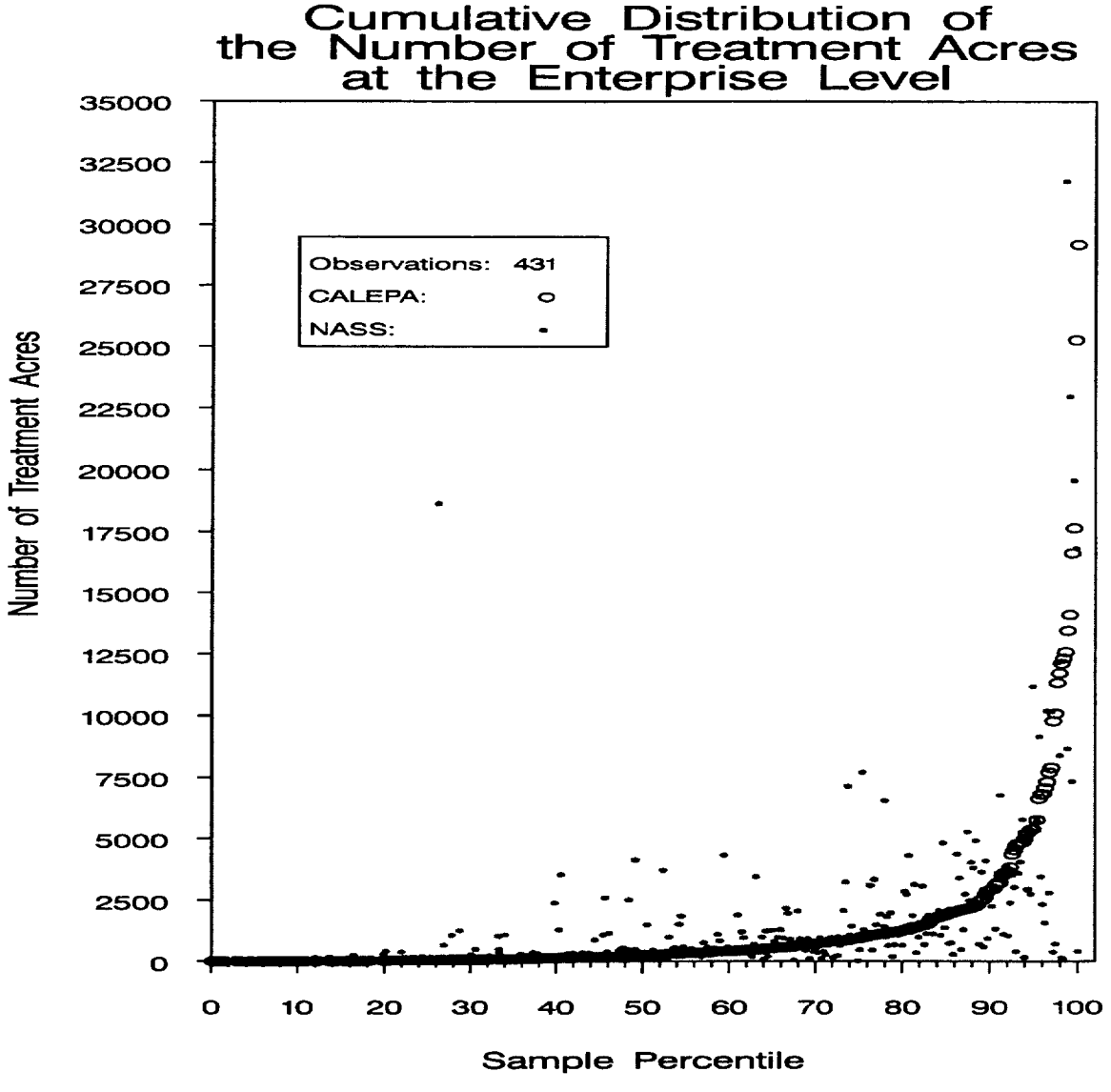

Figure 4. Each point in the figure represents the cumulative number of treatment acres of all pesticides that were applied on an operation which could be found in both the NASS and CALEPA data. The disparity in the number of treatment acres found in the NASS and CALEPA data generally reflects the presence of operators whose records can not be accurately identified with the elements of the NASS sample.

individual operator, or someone else. Being able to distinguish records on that basis makes it possible to see if the divergence that we see in Figure 4 is caused by any confusion due to the classification of the person who applied the pesticide.

The graph shown in Figure 4 was drawn again as Figures 5 and 6 using records from the block level data, so that whatever difference non-individual operators might make in the analysis can be seen. Since the use of pesticides is obtained at the block level for only one commodity, the scale for the number of applications will be much different in the plots for the enterprise and block level data. However, the outstanding features in Figure 4 are also seen in the plots of the cumulative distribu-

Table 11. Reasons for the Largest Differences in Treatment Acres between NASS and CALEPA DATA

Reason	Frequency
NASS Legitimately Got More Data	21
CALEPA Legitimately Got More Data	9
Confusion with Multiple Operations	7
Bad CALEPA ID's	9
Not All CALEPA ID's Used	1
Indeterminate	1

tions for the block level records. Since there is no substantial difference in the plots attributed to the classification of the operators as seen in Figures 5 and 6, there must be another reason for the disparity between the CALEPA and NASS data. This disparity may be caused by the presence of individual operators who apply pesticides to more than one operation or by the logistical and cognitive limitations of collecting large amounts of data related to a long period of time in a single-contact survey.

Figures 4, 5, and 6 may show that a disparity exits between the NASS and CALEPA data, but they do not indicate its extent. In order to consolidate the differences in treatment acres between the NASS and CALEPA data so that a distribution across commodities can be derived, the differences were standardized by means of a Möbius transformation. Using this transformation, differences for a given commodity are scaled so that the most negative difference is mapped to -1; the most positive difference is mapped to +1; and no difference is mapped to 0. These constraints dictate that our transformation cannot be a linear transformation.

Letting N_{ik} be the number of treatment acres reported to NASS by operator i for commodity k and C_{ik} be the number of treatment acres found in the CALEPA data for operator i and commodity k, we define the transformation z as follows:

Table 12. A Sample of Differences Between Recorded Site Location ID's and the Correct ID's.

Recorded Site ID	Correct Site ID
1-101	01-101
02HCP10	02-HCP10
1A	1A00
1105	01-105
01-401	401
1-3-C	1-3
01 101	01-101
26R1	2GR1
1	SITE 1
403	303
43	00000043
0K002	0K2
P0-E8C	PO-E8C
31001	310001
2400334	034

$$z_{ik} = \frac{\alpha_k \Delta_{ik}}{(\beta_k \Delta_{ik} - \gamma_k)}$$

where

$$\Delta_{ik} = N_{ik} - C_{ik}$$

$$\alpha_k = \max_i(\Delta_{ik}) - \min_i(\Delta_{ik})$$

$$\beta_k = \max_i(\Delta_{ik}) + \min_i(\Delta_{ik})$$

$$\gamma_k = 2 \max_i(\Delta_{ik}) \min_i(\Delta_{ik})$$

Note that when Δ_{ik} is the most positive difference in treatment acres between the NASS and CALEPA data for commodity k over all operators, then z_{ik}=+1; when Δ_{ik} is the most negative difference in treatment acres between the NASS and CALEPA data for commodity k over all operators, then z_{ik}= -1. In the event that the NASS and CALEPA data agree in treatment acres, then z_{ik}=0. If the NASS and CALEPA data agreed most of the

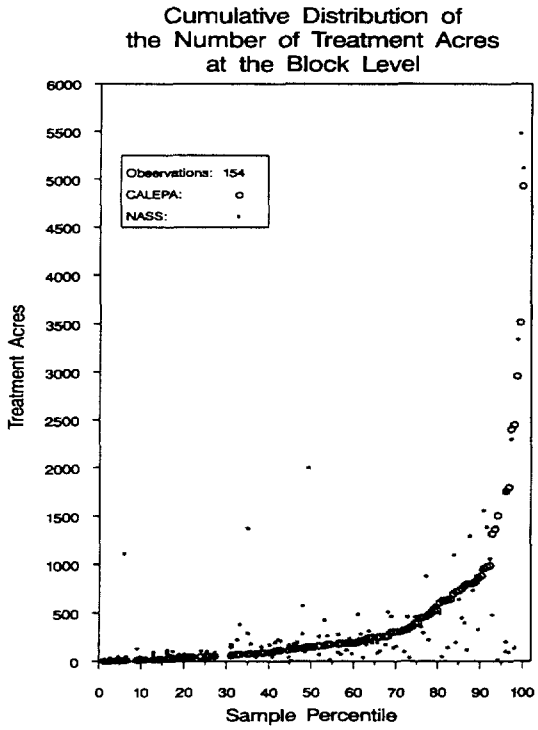

Figure 5. Block level records can be matched one for one with the CALEPA records down to the individual sites. Even so, the same sort of discrepancy still occurs between the CALEPA and NASS data as at the enterprise level.

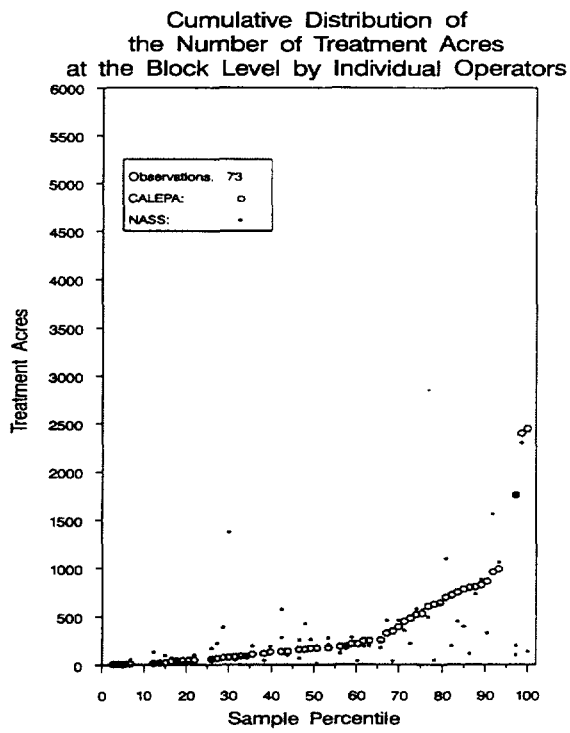

Figure 6. Records from only individual operators show that the persistent disparity between the NASS survey data and the CALEPA data narrows the explanation to the limitations of a survey in collecting data.

time, then z_{ik} should equal zero most of the time.

The shape and location of the distribution of the scaled differences shown in Figure 7 is determined by the many data points which comprise the middle range of the differences. Figure 7 indicates that the distribution of the z_{ik}'s is centered and clustered at zero. This would imply that most of the time the CALEPA and NASS data agree.

Even though we observed that, in the cases of few applications, NASS generally collects more information than is reported to CALEPA and that, in the cases of many applications, the CALEPA data is more complete than the NASS survey data, Figure 7 shows that neither of these cases play a important role in shaping the distribution of differences in treatment acres.

BEARING vs. NON-BEARING ACRES

Two types of treatment acres may be defined - bearing and non-bearing. Historically, NASS has collected data and set estimates on the use of pesticides only for bearing acres. There is no way to identify data pertaining only to bearing acres from the CALEPA data, since the reporting of the use of pesticides to CALEPA is made indiscriminately for bearing and non-bearing acres. No identifying information exists in the CALEPA data to dis-

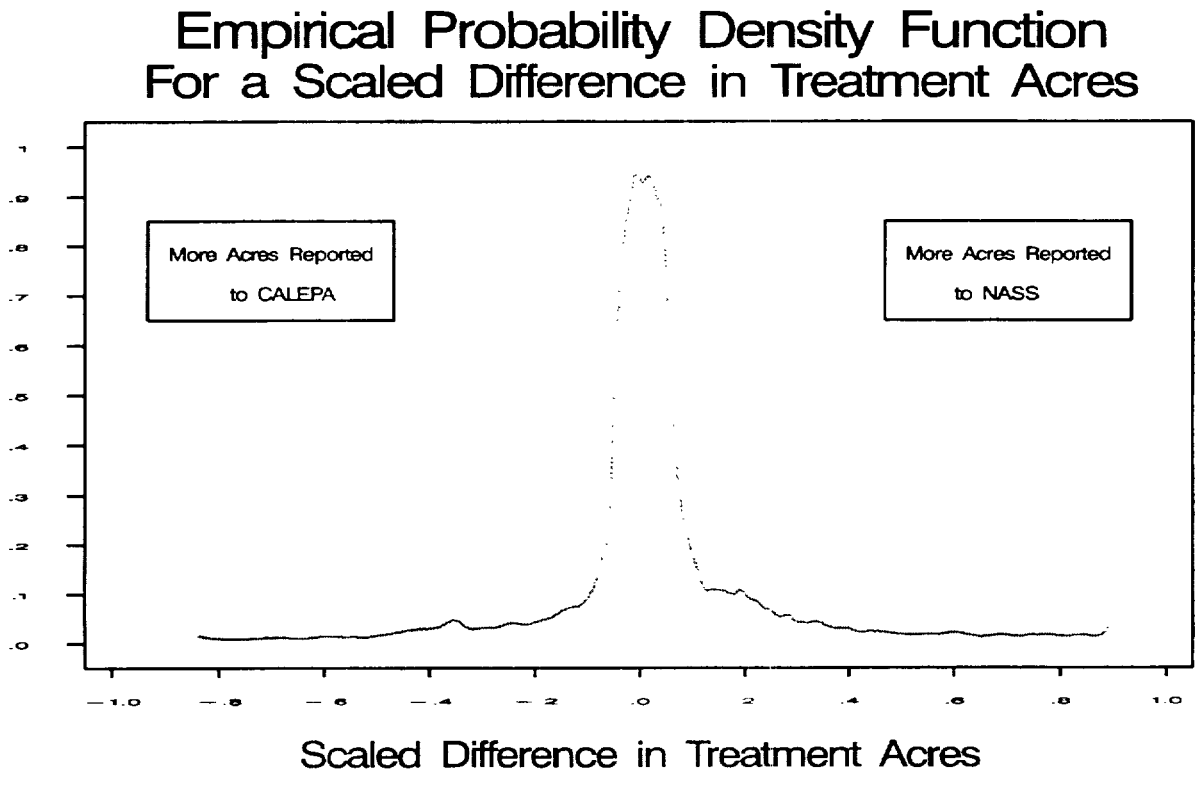

Figure 7. The location of the well defined peak at zero indicate that, in most cases, the CALEPA and NASS data agree in treatment acres.

tinguish bearing and non-bearing acres.

It may seem that, because treatment acres for non-bearing acres are confounded with those for bearing acres, more treatment acres will be found in the CALEPA data than are reported in the NASS data. As a consequence, the distribution of the differences in treatment acres shown in Figure 7 could be shifted accordingly. If the treatment acres found in the CALEPA data were somehow adjusted for bearing acres, then the distribution should shift to the right and rest more squarely on zero. To test that idea, estimates of bearing acres and total acres for the target fruit crops obtained from the California State Statistical Office were used to reduce uniformly the number of treatment acres found in the CALEPA data of every operator, using the ratios shown in Appendix IV. Some operators, of course, have a higher proportion of bearing acres than do others. However, this uniform adjustment should give us some indication of the potential effect of including non-bearing acres. The empirical probability density function of differences in treatment acres using the scaled CALEPA data was computed again and it appears in Figure 8 where it is superimposed on top of the previously computed density function. A noticeable difference does appear. For the 1995 survey, NASS set non-bearing acreage estimates and pesticide estimates were for bearing and non-bearing acreage combined.

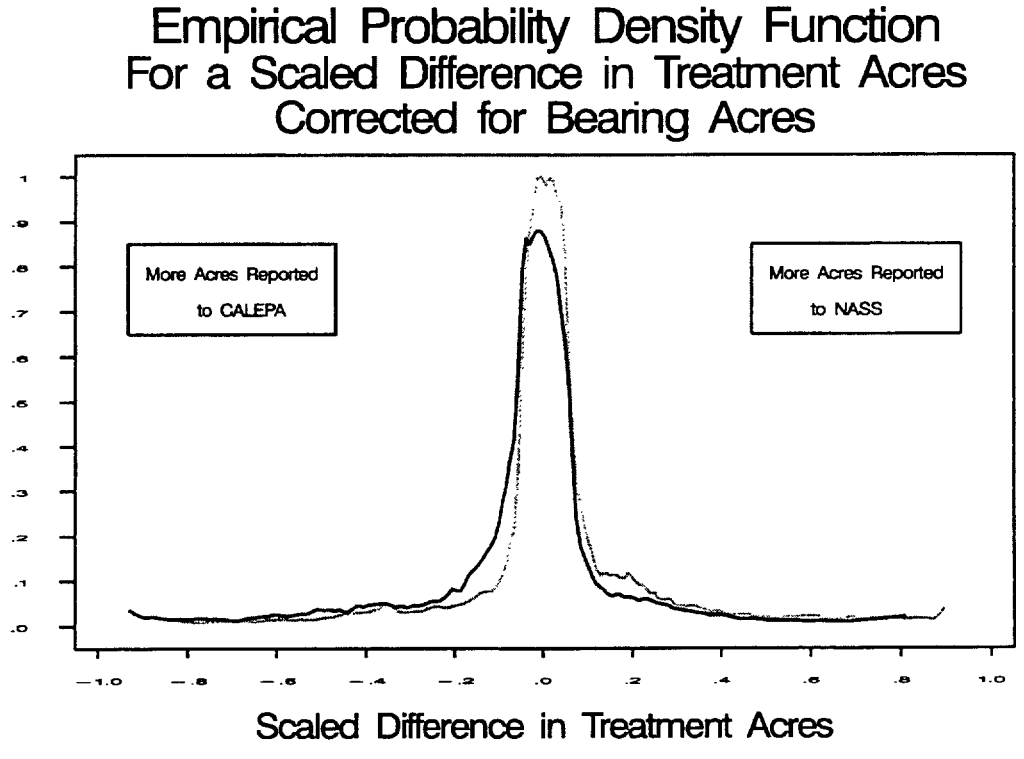

Figure 8. When the difference in treatment acres is corrected for bearing acres, some change occurs in the distribution of the scaled difference. The gray line is the same one that appears in Figure 7 and it represents the uncorrected treatment acres, while the black line represents the distribution corrected for bearing acres.

PROCESSING CALEPA DATA

Reported applications of pesticide use are put into different CALEPA data sets according to the calendar year during which the pesticides were applied. But because NASS requires information about the application of pesticides beginning with the cultivation of a crop and ending with its harvest, two sets of CALEPA data from consecutive years must be combined to meet the needs of NASS.

Before the sets are combined, subsets of each are made which contain the data for only those crops which pertain to a particular Chemical Use Survey. Those crops of interest to NASS are called *target crops.* A chart showing the processing of the CALEPA data is shown in Appendix V. Successful extraction of the pertinent data depends on the reliability of certain key steps in the process.

Important factors that limit the success of extracting CALEPA data are incorrect or missing information on the NASS questionnaire and key-entry errors. By substituting computer assisted personal interviewing (CAPI) instrument for the NASS paper questionnaires, the frequency of these errors could be substantially reduced. The utility of CAPI lies in its facility of using external files for interactive editing at the time of an interview. Furthermore, the results of using a record linkage program to match the names of operators in the NASS sample with entries in the CAC name and address files could be incorporated into a CAPI instrument. Verification of the CALEPA ID numbers given by the respondents can be immediately performed by comparing the reported CALEPA ID numbers with entries in

the external file.

As with any process, procedures should be implemented to monitor the processing of the CALEPA data. The monitoring would be part of a process control program.

It is important that records that are not kept during the process be examined to discover the cause of the attrition. For example, NASS may not be cognizant of CALEPA's decision to change the name of a commodity. With careful monitoring, that change would be discovered. Or some CALEPA ID numbers given to NASS may be valid but inconsistent with the CALEPA ID's recorded in the CAC name and address files. If undetected this might result in matching an operator in the NASS sample to the wrong records in the CALEPA data.

More than a matter of convenience, the processing of the CALEPA data in the California State Statistical Office on a computer dedicated for that purpose will facilitate the preparation of the data for the machine edit. Under that arrangement, the files from CAPI could be transferred directly to the computer on which the edit will be run. Moreover, the troublesome task of transmitting data to the mainframe and the uncertainty of the availability of the mainframe for use during the day would be eliminated. Off-line processing of the CALEPA data in the California SSO should also be more economical.

CONCLUSIONS

The CALEPA data contains an immense amount of pesticide use information for the agricultural industry in California. It is, however, incomplete. It is deficient in the information for operators of small acreages who do not report the use of pesticides to the State. The CALEPA data account for the use of pesticides in about 80 percent of California's agricultural industry. In general, the quality of the CALEPA data appears to be excellent, and the information for operators with fewer than 5,000 treatment acres agrees with NASS survey data reasonably well. On the other hand, the usefulness of the CALEPA data for operators with many applications, in particular operators who apply pesticides to more than one operation, is problematic because of association problems with the NASS sample. The NASS survey data are problematic for the largest operators because of incomplete and inaccurate reporting or recording. Hence NASS obtains imperfect information from the largest operators.

Regardless of the quality of the CALEPA data, its use is limited by processing problems. Extracting the appropriate data is compromised by incorrect identification numbers, incompatible nomenclatures used by NASS and CALEPA, and inaccurate beginning and ending dates. These problems exist to similar extents in the block and enterprise level data. Significant problems did exist with beginning and ending dates, which were correct only 40 percent of the time. It was possible to use a computer program to separate the CALEPA data according to the correct survey period without depending on the NASS survey respondent for the beginning and ending dates.

RECOMMENDATIONS

It is recommended:

1. That a computer assisted personal interview (CAPI) be use to reduce ID entry problems.

2. That record linkage program be applied to the County Agriculture Commissioners' name and address files and the sampling frame used by NASS.

3. That the site location number be

obtained for both the enterprise and block levels.

4. That a quality assurance program be considered to monitor the quality and processing of the purchased administrative data as described in Appendix VI.

5. That the entire processing of the CALEPA data be performed in the California State Statistical Office.

REFERENCES

[1] Amrhein, John. (1994), "The Use of California Environmental Protection Agency Pesticide Use Data by NASS", SRB-94-04, National Agricultural Statistics Service, U. S. Department of Agriculture, Washington, D. C.

[2] Amrhein, John. (1994), Unpublished Notes. National Agricultural Statistics Service, U. S. Department of Agriculture, Washington, D. C.

[3] Day, Charles. (1996) "Record Linkage II: Experience Using AUTOMATCH for Record Linkage", STB-96-01, National Agricultural Statistics Service, U. S. Department of Agriculture, Washington, D. C.

[4] Smith, Jim and Messer, Chris. (1995), "1991 - 1993 Agricultural Chemical Use Survey, Survey Administration Analysis. (1991 Fruit & Nut Chemical Use Survey, 1992 Vegetable Chemical Use Survey, 1993 Fruit Chemical Use Survey)". SAB-95-02, National Agricultural Statistics Service, U. S. Department of Agriculture, Washington, D. C.

APPENDIX I

Number of Different Identification Numbers Found in the CALEPA Data

Name of County	California County ID Number	Number of Permits Issued by County Ag. Commissioners	Reporting for Any Fruit in CALEPA Data	Reporting for Only Nuts in CALEPA Data	Reporting for Any Fruit or Nut Crops in CALEPA Data	Number of 1993 LSF ID's for Fruit Growers
Alameda	01	712	18	3	21	124
Alpine	02	0	0	1	1	0
Amador	03	155	75	8	83	47
Butte	04	1,072	340	478	818	336
Calaveras	05	177	23	13	36	14
Colusa	06	81	35	138	173	43
Contra Costa	07	637	95	37	132	88
Del Norte	08	0	0	0	0	0
El Dorado	09	338	138	0	138	97
Fresno	10	5,112	3,772	301	4,073	3,853
Glen	11	932	155	123	278	200
Humbolt	12	175	28	1	29	8
Imperial	13	3	21	0	21	31
Inyo	14	88	1	0	1	1
Kern	15	1,163	359	212	571	366
Kings	16	1,077	165	98	263	124
Lake	17	6	199	34	233	110
Lassen	18	1	1	0	1	4
Los Angeles	19	743	66	1	67	500
Madera	20	1,061	506	207	713	392
Marin	21	3	8	0	8	19
Marposa	22	230	7	1	8	5
Mendocino	23	566	421	2	423	223
Merced	24	1,883	279	818	1,097	236
Modoc	25	48	1	0	1	0
Mono	26	0	0	0	0	1
Monterey	27	1,016	70	3	73	51
Napa	28	636	546	3	549	600
Nevada	29	118	34	0	34	23
Orange	30	51	22	0	22	344
Placer	31	13	106	7	113	74
Plumas	32	52	0	0	0	3
Riverside	33	1,465	394	1	395	825
Sacramento	34	575	117	8	125	80
San Benito	35	371	95	22	117	114
San Bernadino	36	888	117	5	122	204
San Diego	37	3,212	1,103	4	1,107	1,994
San Francisco	38	3	0	0	0	26
San Joaquin	39	3,277	811	654	1,465	924
San Luis Obispo	40	1,035	211	47	258	226
San Mateo	41	147	7	0	7	23
Santa Barbara	42	762	483	24	507	276
Santa Clara	43	782	165	26	191	230
Santa Cruz	44	568	151	0	151	206
Shasta	45	11	26	36	62	29
Sierra	46	18	0	0	0	0
Siskiyou	47	517	10	0	10	6
Solano	48	588	171	73	244	131
Sonoma	49	1,313	676	9	685	919
Stanislaus	50	2,419	513	1,177	1,690	548
Sutter	51	1,032	714	351	1,065	380
Tehmia	52	846	239	113	352	304
Trinity	53	15	16	1	17	2
Tulare	54	4,285	2,949	359	3,308	2,118
Tuollumne	55	8	12	0	12	7
Ventura	56	1,631	693	2	695	740
Yolo	57	623	113	152	265	101
Yuba	58	265	228	73	301	92
Undetermined		19				
Total		44,824	17,505	5,626	23,131	18,422

APPENDIX II

Number of Records of Each Commodity in the CALEPA Data.

Name of Commodity	Commodity Code	Number of Records in CALEPA Data
Apple	211999	35,129
Apricot	217999	13,854
Avocado	221999	13,867
Cherry, Sweet	213199	9,146
Date	224999	2,883
Fig	225999	773
Grapefruit	202999	10,595
Grape: Raisin	216399	125,373
Grape: Table	216199	182,623
Grape: Wine & Juice	216299	213,957
Kiwi Fruit	218699	2,010
Lemon	204999	30,529
Nectarine	218199	77,881
Olive	226999	9,413
Orange	201999	147,310
Peach	212999	109,724
Pear	214999	27,902
Plum	215199	55,132
Prune	215299	18,855
Tangerine	203999	4,646

APPENDIX III

Characterization of the Matching on ID's Between NASS and CALEPA Data

Name of County	ID's in 1993 LSF	Block Level Sample				Enterprise Level Sample			
		Successful Matches	Unsuccessful Matches	Out of Scope	Total In Sample	Successful Matches	Unsuccessful Matches	Out of Scope	Total In Sample
Alameda	124	0	0	0	0	1	1	0	2
Alpine	0	0	0	0	0	0	0	0	0
Amador	47	1	0	0	1	1	0	0	1
Butte	336	0	0	0	0	16	3	4	23
Calaveras	14	2	0	0	2	0	0	0	0
Colusa	43	0	0	0	0	0	0	1	1
Contra Costa	88	0	0	0	0	5	0	3	8
Del Norte	0	0	0	0	0	0	0	0	0
El Dorado	97	3	0	1	4	7	0	0	7
Fresno	3,853	72	2	1	75	97	1	6	104
Glen	200	0	0	0	0	7	1	2	10
Humbolt	8	0	0	0	0	0	0	0	0
Imperial	31	0	0	0	0	0	0	0	0
Inyo	1	0	0	0	0	0	0	0	0
Kern	366	8	0	2	10	5	3	0	8
Kings	124	2	1	0	3	0	2	0	2
Lake	110	0	1	0	1	0	6	2	8
Lassen	4	0	0	0	0	0	0	0	0
Los Angeles	500	5	1	0	6	4	3	3	10
Madera	392	3	0	0	3	8	3	0	11
Marin	19	0	0	0	0	0	0	0	0
Marposa	5	0	0	0	0	0	0	0	0
Mendocino	223	3	3	0	6	6	3	1	10
Merced	236	1	0	0	1	10	0	0	10
Modoc	0	0	0	0	0	0	0	0	0
Mono	1	0	0	0	0	0	0	0	0
Monterey	51	2	0	0	2	1	1	0	2
Napa	600	11	0	0	11	7	0	0	7
Nevada	23	0	0	0	0	1	0	0	1
Orange	344	1	0	0	1	2	6	0	8
Placer	74	1	0	0	1	4	0	1	5
Plumas	3	0	0	0	0	0	0	0	0
Riverside	825	4	1	0	5	29	9	7	45
Sacramento	80	0	1	0	1	8	1	2	11
San Benito	114	0	1	0	1	2	2	1	5
San Bernadino	204	2	0	0	2	7	1	0	8
San Diego	1,994	3	2	0	5	14	14	4	32
San Francisco	26	0	0	0	0	0	0	0	0
San Joaquin	924	10	1	1	12	38	8	4	50
San Luis Obispo	226	4	2	0	6	5	1	0	6
San Mateo	23	0	0	0	0	1	0	0	1
Santa Barbara	276	0	0	0	0	4	5	2	11
Santa Clara	230	0	0	1	1	3	4	0	7
Santa Cruz	206	11	4	1	16	6	3	2	11
Shasta	29	0	0	0	0	0	0	0	0
Sierra	0	0	0	0	0	0	0	0	0
Siskiyou	6	0	0	0	0	0	0	0	0
Solano	131	0	0	0	0	11	0	3	14
Sonoma	919	0	23	0	23	1	16	1	18
Stanislaus	548	7	3	1	11	22	1	2	25
Sutter	380	1	0	1	2	24	5	3	32
Tehmia	304	0	0	0	0	10	2	2	14
Trinity	2	0	0	0	0	0	0	0	0
Tulare	2,118	54	4	3	61	76	8	10	94
Tuollumne	7	0	0	1	1	0	1	0	1
Ventura	740	13	3	2	18	26	9	4	39
Yolo	101	0	0	0	0	3	0	0	3
Yuba	92	0	0	0	0	3	1	0	4
Total	18,422	224	52	15	292	475	125	69	669

APPENDIX IV

1993 California SSO Estimates of Bearing and Non-bearing Acres for Fruits

Name of Commodity	Bearing Acres	Non-bearing Acres	Total Acres	Ratio of Bearing to Total Acres
Apple	34,700	1,600	36,300	.96
Apricot	19,300	1,500	20,800	.93
Avocado	72,900	1,500	73,200	1.00
Cherry, Sweet	11,700	2,700	14,400	.81
Date	5,500	200	5,700	.96
Fig	12,200	500	12,700	.96
Grapefruit	17,800	2,600	20,400	.87
Grape: Raisin	266,000	10,200	276,200	.96
Grape: Table	77,800	5,000	82,800	.94
Grape: Wine & Juice	312,000	25,200	337,200	.93
Kiwi Fruit	7,200	10	7,210	1.00
Lemon	46,400	2,100	48,500	.96
Nectarine	27,100	8,300	35,400	.77
Olive	30,100	4,400	34,500	.87
Orange	184,000	16,600	200,600	.92
Peach	59,700	14,400	74,100	.81
Pear	24,100	1,000	25,100	.96
Plum	41,600	3,500	45,100	.92
Prune	78,200	8,400	86,600	.90
Tangerine	7,700	1,300	9,000	.86