FARM ENVIRONMENTAL STEWARDSHIP SAMPLING PROTOCOL

MARK L. KINSEL, DVM, PHD

THE FARM ENVIRONMENTAL STEWARDSHIP MODULE

The FARM Environmental Stewardship module provides a comprehensive estimate of the greenhouse gas (GHG) emissions and energy use associated with dairy farming. The tool is based on a life cycle assessment (LCA) of fluid milk conducted by the Applied Sustainability Center at the University of Arkansas, incorporating data from more than 500 dairy farms across the United States. The FARM ES module asks a limited set of questions to assess a farm's carbon and energy footprint – reducing the burden on farmers while still providing reliable, statistically robust estimates.

STRATIFIED RANDOM SAMPLING

This selection protocol is based on the concept of "stratified random sampling". Stratified random sampling is a method of sampling based on dividing the population of interest (e.g. dairies of interest) into groups ("strata") based on common factors likely to influence the outcome to be measured (in this case, greenhouse gas emissions) (Kelsey et al., 1986; Martin et al., 1987; Dohoo et al., 2010). The goal of stratified random sampling is to generate a sample that has the same proportions of the grouping factors as the population of interest (a "representative" sample). It makes sure that members of all the strata are equally likely to be selected for the sample. It is most easily explained by an example:

Let's say we want to investigate the effect of hat color on baldness. Our population of 100 hats is broken down into the following colors: 10 blue, 70 red, and 20 yellow. If we wanted to select 10% of the 100 hats (10 hats) for our sample and we use simple random sampling (paying no attention to color), it would be highly likely that we would pick a sample of 10 hats with no blue hats. However, if we stratify each hat by color first and then randomly select 10% from each strata, we will get 1 blue, 7 reds, and 2 yellows … exactly matching the proportions of the overall population thus giving us a "representative sample".

The three advantages of stratified random sampling is that: 1) it ensures all strata are represented in the sample, 2) the precision of the overall estimates are more likely to be greater than a simple random sample as the differences between strata are explicitly removed from the estimates¸ and 3) it produces stratum-specific estimates of the outcome of interest (Dohoo et al, 2010).

The general steps used in generating a stratified random sample are as follows:

1. Identify the characteristics that will be used to divide the population into subgroups ("strata"),
2. Organize the population of interest into non-overlapping strata (each member can be in only one subgroup / strata),

3. Multiply the proportion to be selected from the population by the number of members in the strata to determine the number to be selected in each strata,
4. Randomly select members from each strata until the number to be selected is reached.

Once again, the goal is to have the proportions of stratifying characteristics in the sample match as closely as possible to the proportions of stratifying characteristics in the population (a "representative sample").
Stratified random sampling works best when the following conditions exist:

1. The population of interest consists of a finite number of members
2. The population consists of subgroups and each of these subgroups must be investigated to increase the accuracy of the results
3. Each member of the population can be assigned to only one stratum
4. Proportional samples can be obtained from within each stratum

Another important concept when we stratify by more than one factor is that the number of members in each group does not become too small. In this situation, there is a substantial reduction in precision of the effect estimates (Kelsey et al, 1986). Using our hat sampling example, if we stratified by color, size, manufacturer, shape, and weight, many of the strata would have only one member (called "thin data"). Thus, it is important to strike a balance between number of stratification factors and the size of each stratum. A good rule of thumb is to use no more than 3 factors unless there is a large population size and a low percentage to sample. Ideally, we would like more than 5-10 members in each strata.

STRATIFICATION FACTORS FOR ENVIRONMENTAL STEWARDSHIP

Beginning in 2008, a group of studies were conducted to begin to understand the potential effects of dairy production on greenhouse gas (GHG) emissions (Asselin-Balencon et al., 2013; Popp et al., 2013; Thoma et al., 2013). Results from these studies indicated that GHG emissions on dairies were influenced by geographic region, dairy size, level of milk production, type of feeds and feeding practices, and manure management, along with a host of other factors. In one study, more than 162 animal-rations were investigated (Thoma et al., 2013), whereas a second study was able to achieve similar results when considering 12 feed-rations (Asselin-Balencon et al., 2013). Regardless of these studies, the number of stratifying factors greatly exceeds our requirement for 3 or less. In all three of these survey studies, important factors were geographic region, herd size, and level of milk production. A second "driver" in our choice of stratification factors was availability of data. It was important to select items that were not only important predictors of GHG emissions, but also data that was relatively easy for dairy cooperatives and milk marketing organizations to obtain from their member dairies. Number of milk cows was not data that was routinely known to milk processors and was correlated to level of milk production so that the ending stratification factors were level of milk production and region.

<u>Level of Milk Production</u>

Level of milk production should be calculated using the daily Fat & Protein Corrected Milk (FPCM) produced. Users should enter the number calculated by using the IDF formula (2010):

FPCM = Milk in pounds x [(0.1226 x Fat percent) + (0.0776 x Protein percent) + 0.2534]

Nearly all milk handlers have daily data on milk production for each dairy as well as the component data required to calculate FPCM. In situations where the daily FPCM production cannot be calculated because milk fat and milk protein percentages are not known, daily milk production in pounds may be substituted. The downside of this substitution is that some breeds maybe moved to a different stratum than they would be without the substitution (e.g. Jersey dairies) due to differences in component levels. This total amount of milk shipped per farm per day will be used later to calculate production quartiles.

Geographic Region

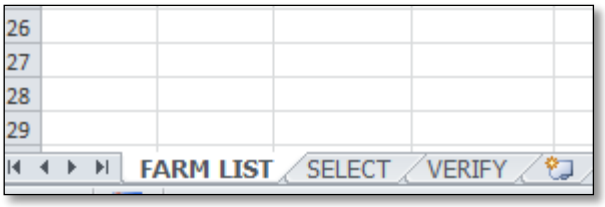
Geographic region is by far the most difficult to define, but important stratifying factor to include. While the parsimonious model described by Asselin-Balencon et al. (2013) includes many feed and feed management factors important in predicting GHG emissions, it would be impossible to use all these characteristics to generate a stratified random sample. While certainly not perfect, geographic region allows us to capture the effects of many predictors in a single factor, including feed management, if we make a few assumptions. First, we assume that the climate is the same for all dairies in the geographic region. This is problematic if we look at too big a region such as a state. For example, the climate in the western / coastal region of Washington State is tremendously different than the eastern / arid region of the state. Same issue in the state of Idaho where the north is mountainous and the south is large valleys. But in a small enough region, this is a safe assumption. The second assumption is that the feeds and the feeding practices used are similar within the region. Again, this is not true on a state level, but is probably a safe assumption if the area is small enough. Third, it is assumed that the herd management practices, including manure management, are similar if the area is small enough. However, if too small an area is selected, the problem of "thin data" arises.

So if state boundaries are too big to use, what definition of geographic region can be used that defines a region small enough that the above assumptions are true? Possibilities that were considered included grouping dairies by which processing plant they shipped to, grouping dairies in the same BTU, and grouping dairies in the same Federal Milk Order. Grouping dairies according to which processing plant received their milk wouldn't work since for large processors, a given plant may receive milk from many states leading to the same problem as using state boundaries. Grouping dairies by BTU wouldn't work because there are many BTU's with only one dairy in them leading to thin data. This is the same for grouping by county. Grouping by Federal Milk Order led to too big an area as many FMO's include multiple states.

The geographic region definition that seems to hit the "sweet spot" is the first three numbers of the zip code known as the "Zip-3" code. The US National Weather Service has used to the Zip-3 code to generate regional numbers for rain fall totals so they roughly represent a common climate. They

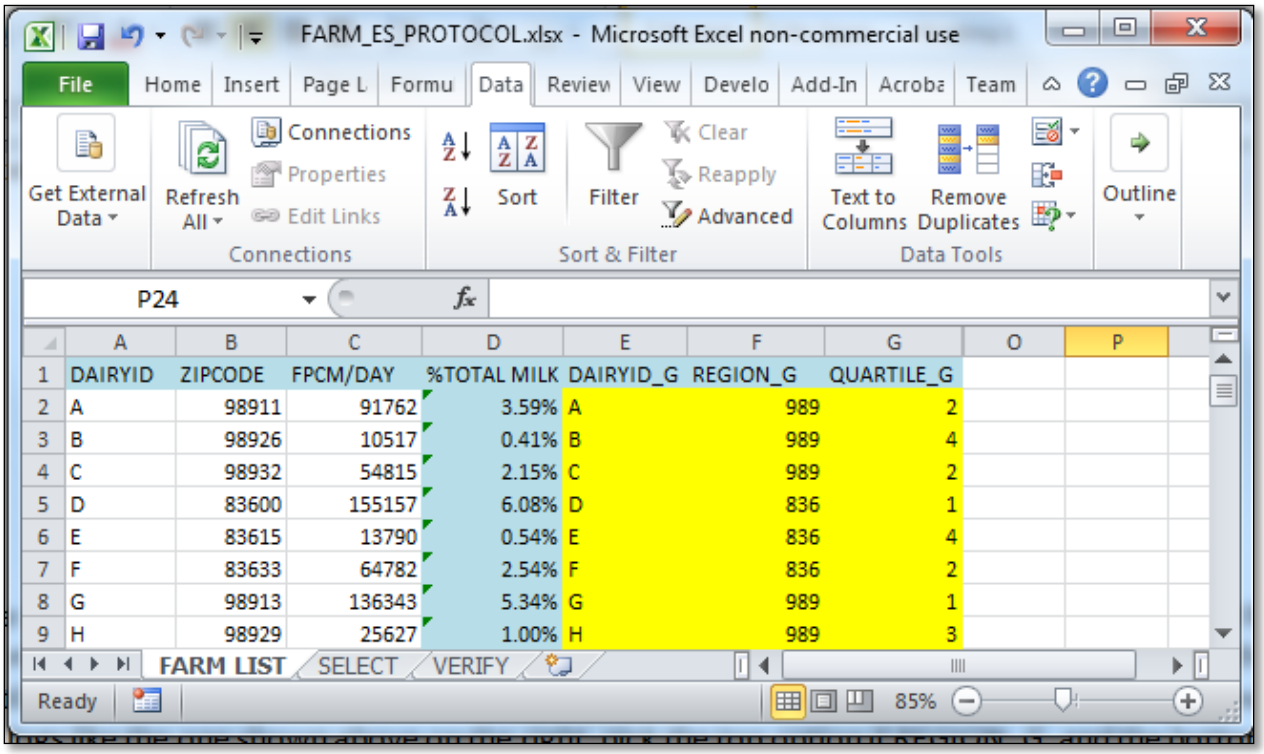
are big enough to avoid “thin data” issues, yet small enough to relatively safely assume similar feeds and feeding practices, similar herd management practices, and similar manure practices. The figure below shows the current “Zip-3” code map for the United States.

MICROSOFT EXCEL SPREADSHEET FOR SELECTING DAIRIES



In order to streamline the selection process for dairy cooperatives and milk marketing organizations, a Microsoft Excel spreadsheet has been created that performs the necessary calculations to select farms according to the protocol. You can obtain this spreadsheet by contacting NMPF. The spreadsheet is broken into three worksheets listed as tabs on the bottom of the workbook: FARM LIST, SELECT, and VERIFY (shown at upper left). The FARM LIST worksheet is used to sort the farms of interest into strata based on region and milk production. Once the farms are sorted into their respective groups, the user copies each group of farms to work with into the SELECT worksheet to randomly select the appropriate number of farms from the group (strata). Data from the SELECT worksheet is copied to the VERIFY worksheet to verify that the farms chosen were selected properly according to protocol.

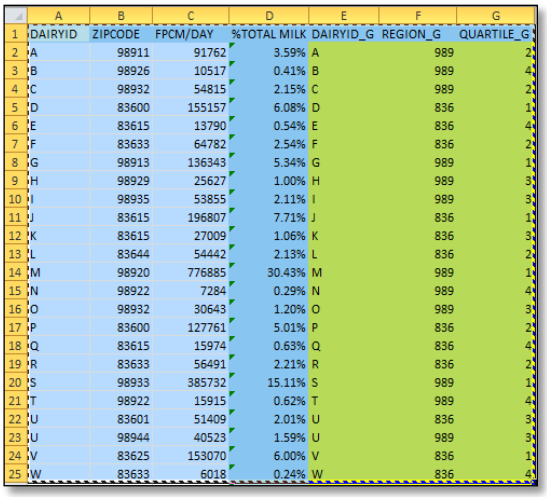
The FARM LIST worksheet is shown in the figure below. Columns A through C are the columns in white where user data is entered. This data consists of: 1) the DAIRYID in column A, a unique identifier for the
dairy that can be either the producer number or dairy name, 2) the ZIPCODE of the dairy in column B, which will be used to generate the ZIP-3 region code, and 3) the daily amount of FPCM milk produced in
Column C used to generate the milk production quartile. Column D uses the data in Column C to calculate the percent of milk produced by each dairy out of the total produced by all dairies entered.



	A	B	C	D	E	F	G	O	P
1	DAIRYID	ZIPCODE	FPCM/DAY	%TOTAL MILK	DAIRYID_G	REGION_G	QUARTILE_G		
2	A	98911	91762	3.59%	A	989	2		
3	B	98926	10517	0.41%	B	989	4		
4	C	98932	54815	2.15%	C	989	2		
5	D	83600	155157	6.08%	D	836	1		
6	E	83615	13790	0.54%	E	836	4		
7	F	83633	64782	2.54%	F	836	2		
8	G	98913	136343	5.34%	G	989	1		
9	H	98929	25627	1.00%	H	989	3		

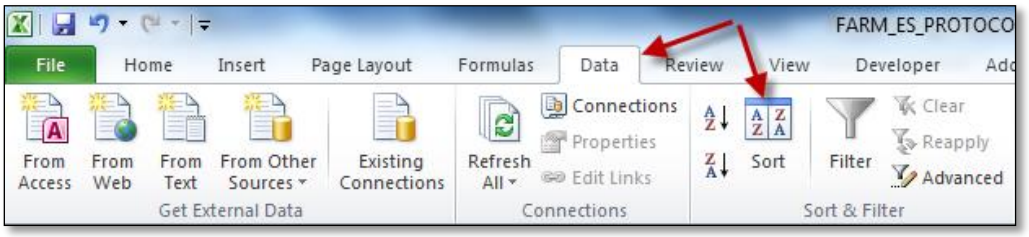
You will notice that certain cells are highlighted in light blue (e.g. the first row). You should not change / edit these cells. You will also notice that columns H through N in this worksheet are hidden. These columns contain the calculations necessary to break the dairies into the strata shown in the yellow columns E through G. You should not unhide columns H through N, change the data in columns E through G, or change the column names. If the data in columns E through G accidently gets changed, try undoing the change first with the Undo button. You may also copy columns E through G of a row that was not changed and paste it to the row that was. If these two methods do not return the data, close the spreadsheet without saving it, and start the process over.

The spreadsheet is designed to sample from up to 3000 dairies of interest (the "population"). The data in Columns E through G will automatically calculate and update as you enter data. Column F consists of the first 3 numbers of the zip code. Column G ranks the herds on daily milk production and breaks them into quartiles, with Quartile 1 being the top 25% of dairies by total milk produced and Quartile 4 being the lowest 25% of dairies by total milk produced. Once you have the farm data entered, you need to sort the data by the grouping columns (columns F and G). You start this sorting process by highlighting all the data rows for columns A through G, including the column heading as shown here on the right:

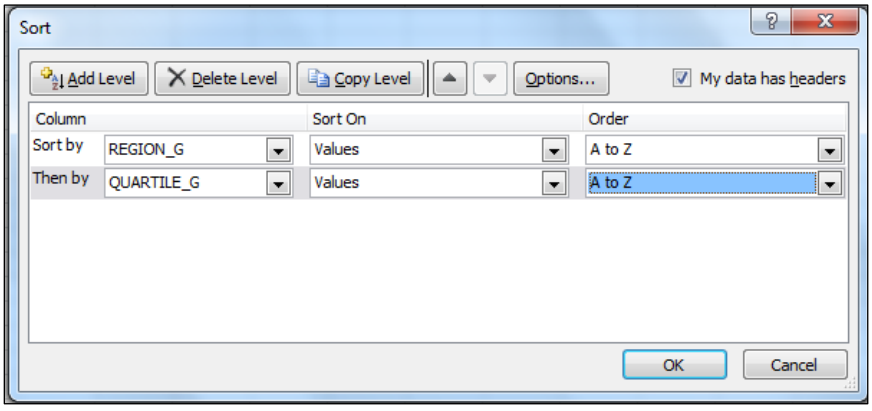


	A	B	C	D	E	F	G
1	DAIRYID	ZIPCODE	FPCM/DAY	%TOTAL MILK	DAIRYID_G	REGION_G	QUARTILE_G
2	A	98911	91762	3.59%	A	989	2
3	B	98926	10517	0.41%	B	989	4
4	C	98932	54815	2.15%	C	989	2
5	D	83600	155157	6.08%	D	836	1
6	E	83615	13790	0.54%	E	836	4
7	F	83633	64782	2.54%	F	836	2
8	G	98913	136343	5.34%	G	989	1
9	H	98929	25627	1.00%	H	989	3
10	I	98935	53855	2.11%	I	989	3
11	J	83615	196807	7.71%	J	836	1
12	K	83615	27009	1.06%	K	836	3
13	L	83644	54442	2.13%	L	836	2
14	M	98920	776885	30.43%	M	989	1
15	N	98922	7284	0.29%	N	989	4
16	O	98932	30643	1.20%	O	989	3
17	P	83600	127761	5.01%	P	836	2
18	Q	83615	15974	0.63%	Q	836	4
19	R	83633	56491	2.21%	R	836	2
20	S	98933	385732	15.11%	S	989	1
21	T	98922	15915	0.62%	T	989	4
22	U	83601	51409	2.01%	U	836	3
23	U	98944	40523	1.59%	U	989	3
24	V	83625	153070	6.00%	V	836	1
25	W	83633	6018	0.24%	W	836	4

Next, you will need to click the Excel Data tab and choose the Sort icon:



Next, set up your sort order as shown below:



The sort dialog shown above will initially start out with only the "Sort by" row filed. Change DAIRY_ID to REGION_G in the "Sort by" dropdown list and then click the "Add Level" button to add the second level. Select the QUARTILE_G item from the "Then by" dropdown list so your sort dialog looks like the one above and then click the OK button. If you make a mistake, you can click the

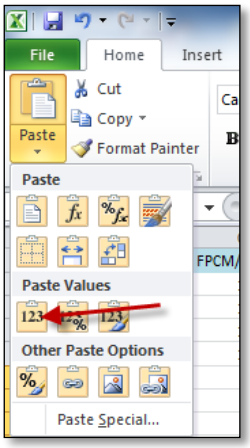
Undo button at the top left of Excel to return your spreadsheet to how it was. Your spreadsheet should now look like this:

	A	B	C	D	E	F	G
1	DAIRYID	ZIPCODE	FPCM/DAY	%TOTAL MILK	DAIRYID_G	REGION_G	QUARTILE_G
2	D	83600	155157	6.08%	D	836	1
3	J	83615	196807	7.71%	J	836	1
4	V	83625	153070	6.00%	V	836	1
5	P	83600	127761	5.01%	P	836	1
6	F	83633	64782	2.54%	F	836	2
7	L	83644	54442	2.13%	L	836	2
8	R	83633	56491	2.21%	R	836	2
9	K	83615	27009	1.06%	K	836	3
10	U	83601	51409	2.01%	U	836	3
11	E	83615	13790	0.54%	E	836	4
12	Q	83615	15974	0.63%	Q	836	4
13	W	83633	6018	0.24%	W	836	4
14	G	98913	136343	5.34%	G	989	1
15	M	98920	776885	30.43%	M	989	1
16	S	98933	385732	15.11%	S	989	1
17	A	98911	91762	3.59%	A	989	2
18	C	98932	54815	2.15%	C	989	2
19	H	98929	25627	1.00%	H	989	3
20	I	98935	53855	2.11%	I	989	3
21	O	98932	30643	1.20%	O	989	3
22	U	98944	40523	1.59%	U	989	3
23	B	98926	10517	0.41%	B	989	4
24	N	98922	7284	0.29%	N	989	4
25	T	98922	15915	0.62%	T	989	4

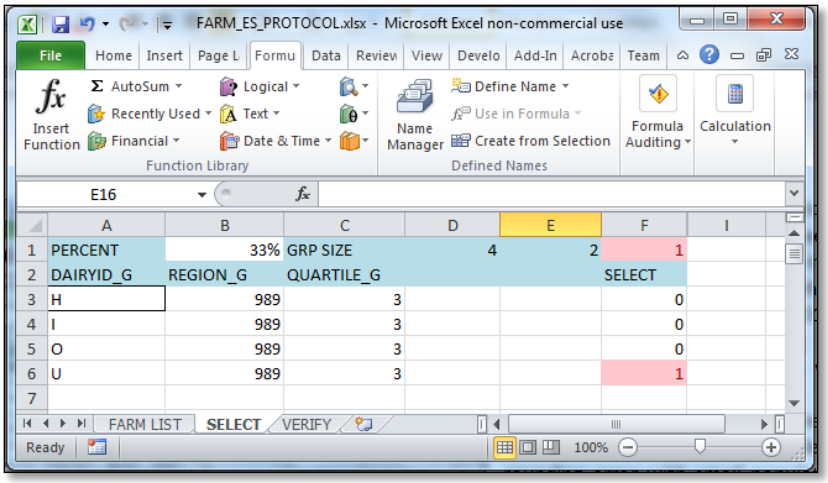
Sometimes with dairies with the same zip code, Excel doesn't sort things correctly. If the quartiles do not appear to be grouped correctly, click the Sort button and rerun the sort. This usually fixes the problem. The subgroups / strata are the dairies that have matching data in columns F and G. The figure below shows the 8 different strata.

	A	B	C	D	E	F	G
1	DAIRYID	ZIPCODE	FPCM/DAY	%TOTAL MILK	DAIRYID_G	REGION_G	QUARTILE_G
2	D	83600	155157	6.08%	D	836	1
3	J	83615	196807	7.71%	J	836	1
4	V	83625	153070	6.00%	V	836	1
5	P	83600	127761	5.01%	P	836	1
6	F	83633	64782	2.54%	F	836	2
7	L	83644	54442	2.13%	L	836	2
8	R	83633	56491	2.21%	R	836	2
9	K	83615	27009	1.06%	K	836	3
10	U	83601	51409	2.01%	U	836	3
11	E	83615	13790	0.54%	E	836	4
12	Q	83615	15974	0.63%	Q	836	4
13	W	83633	6018	0.24%	W	836	4
14	G	98913	136343	5.34%	G	989	1
15	M	98920	776885	30.43%	M	989	1
16	S	98933	385732	15.11%	S	989	1
17	A	98911	91762	3.59%	A	989	2
18	C	98932	54815	2.15%	C	989	2
19	H	98929	25627	1.00%	H	989	3
20	I	98935	53855	2.11%	I	989	3
21	O	98932	30643	1.20%	O	989	3
22	U	98944	40523	1.59%	U	989	3
23	B	98926	10517	0.41%	B	989	4
24	N	98922	7284	0.29%	N	989	4
25	T	98922	15915	0.62%	T	989	4

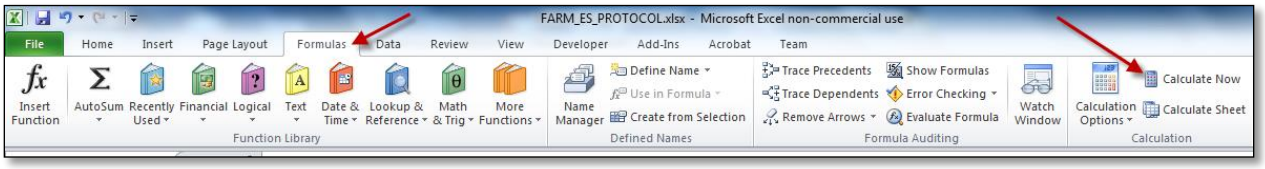
For this small example population with only 24 farms, the strata are quite thin (small number of members). Once the dairies have been stratified into their subgroups, the next step is to randomly select farms from each stratum to create our sampling list. Let's look at how to select farms from strata #7 above.



The first step is to highlight and copy the strata #7 data from the FARM LIST worksheet to the SELECT worksheet as shown below. When copy and pasting in this spreadsheet, it is essential to paste data as values. We do this by clicking the triangle under the Paste icon in the Excel control pane when we're ready to paste and selecting the first icon in the Paste Values row. This will copy the values in the cells, not cell references or formulas. In Cell B1 of this worksheet, we enter the percentage of farms we want to randomly select from the population. In this example, it's 33%. Notice that this worksheet automatically counts the number of rows (group size) we've copied to this worksheet in Cell D1.



The worksheet will automatically calculate the maximum number of farms to select from the group (Cell E1) by multiplying the desired percentage to sample (Cell B1) times the number of farms in the group (Cell D1) times a weighting factor and rounding it up to the nearest whole number. This weighting factor allows us to choose a higher percentage of farms from the highest quartile of milk production since they are likely to have more greenhouse gas emissions. Because of rounding error with small population sizes, we may or may not get exactly the percentage of farms we're trying to get when we're done, but we'll be close and err on getting more than needed. Column F of the SELECT worksheet is where farms are randomly selected based on hidden calculations in Columns G and H. Selected rows (farms) are marked with a red "1" with a pink highlighted cell. The value in Cell F1 is the total number of farms currently selected. Notice in our example above that 1 farm is currently selected (shown in Cell F1) and the maximum to select is 2 (shown in Cell E1). If Cell F1 is greater than Cell E1, randomize again by going to the Excel Formulas tab and clicking the Calculate Now icon.



A new set of farms will be randomly selected. Continue clicking the Calculate Now icon until the number of farms in Cell F1 is less than or equal to the number of farms in Cell E1 (in this case, "2"). These farms will be the randomly selected farms to include in our sample from the group. In the example below, Cells E1 and F1 match, and the two randomly selected farms are H and U:

	A	B	C	D	E	F
1	PERCENT	33%	GRP SIZE	4	2	2
2	DAIRYID_G	REGION_G	QUARTILE_G			SELECT
3	H	989	3			1
4	I	989	3			0
5	O	989	3			0
6	U	989	3			1

The last task to do after selecting farms for a given stratum is to transfer data from the SELECT worksheet to the VERIFY worksheet to make sure we did things properly. First we need to create a

row with the REGION and QUARTILE for each stratum in the VERIFY worksheet. You can do this before completing the SELECT worksheet or as you go selecting farms in the SELECT worksheet. For each stratum (row), you will need to enter the number of farms selected in Column C (it is also the number in Cell F1 in SELECT worksheet) and the number of farms in the group / strata in Column E (also the number in Cell D1 of SELECT worksheet). The figure below shows the worksheet for the eight stratum / groups in our example.

	A	B	C	D	E	F	G
1	OK	CHECK	11		24	45.83%	
2	REGION_G	QUARTILE	SELECTED	SELECT %	STRAT SIZE	STRAT %	FLAG
3	836	1	3	27.27%	4	16.67%	
4	836	2	1	9.09%	3	12.50%	
5	836	3	1	9.09%	2	8.33%	
6	836	4	1	9.09%	3	12.50%	
7	989	1	2	18.18%	3	12.50%	
8	989	2	1	9.09%	2	8.33%	
9	989	3	1	9.09%	4	16.67%	?
10	989	4	1	9.09%	3	12.50%	

This worksheet has several error checking features. Cell A1 indicates whether we ended up with the right number of samples based on the number of farms selected (Cell C1), the number of total farms (Cell E1), the percentage we wanted to select, and weighting factor. The spreadsheet is designed to err on the side of sampling too many farms instead of too few based on rounding errors. A green "OK" means we have an appropriate number of selected farms, while a red "CHECK" means something doesn't appear correct. Cell B1 does a similar error check, but is looking row by row to make sure each row is correct. If it finds a row that appears to have an error, it will put a "?" with a pink background in the FLAG column (Column G) of that row and set Cell B1 to "CHECK". In this example, Row 9 has a problem because only 1 farm was selected (1/4 = 25%) when we should have more than 33% and have selected two. Column D and F express the same data in Column C and E as percentages. If you have an error, check the steps in the FARM LIST and SELECT worksheets. When you have a green "OK" in both Cell A1 and Cell B1, you have selected your samples correctly.

CONCLUSION

The document presents details describing the methodology used in the FARM Environmental Stewardship sampling protocol and a Microsoft Excel spreadsheet to assist dairy cooperatives and milk marketing organizations in selecting a representative group of farms for evaluations. Due to differences in organization size and rounding errors, this protocol will not always select exactly the expected percentage of farms to evaluate. It has been designed to err on selecting more dairies than needed rather than to select too few. For questions regarding this protocol, please contact Ryan Bennett at NMPF.

REFERENCES

Asselin-Balencon AC, Popp J, Henderson A, Heller M, Thoma G, Jolliet O (2013). Dairy farm greenhouse gas impacts: A parsimonius model for a farmer's decision support tool. International Dairy Journal 31:S65-S77.

Dohoo IR, Martin SW, and Stryhn H (2010). Veterinary Epidemiologic Research, 2nd Edition. VER Inc, Charlottetown, Prince Edward Island, Canada. pp 38-39.

IDF (2010). A common carbon footprint approach for dairy – the IDF guide to standard life cycle assessment methodology for the dairy sector. *In Bulletin 445 of the International Dairy Federation.* Brussels, Belgium: International Dairy Federation.

Kelsey JL, Thompson WD, and Evans AS (1986). Methods In Observational Epidemiology, 1st Edition. Oxford University Press, New York, NY. pp 260-262.

Martin SW, Meek AH, and Willeberg P (1987). Veterinary Epidemiology: Principles and Methods, 1st Edition. Iowa State University Press, Ames, Iowa. pp 27-28.

Popp JS, Thoma GJ, Mulhern J, Jaeger A, LeFranc I, and Kemper N (2013). Collecting complex comprehensive farm level data through a collaborative approach: A framework developed for a life cycle assessment of fluid milk production in the US. International Dairy Journal 31:S15-S20.

Thoma G, Popp J, Shonnard D, Nutter D, Matlock M, Ulrich R, Kellogg W, Kim DS, Neiderman Z, Kemper N, Adam F, and East C (2013). Regional analysis of greenhouse gas emissions from USA dairy farms: A cradle to farm-gate assessment of the American dairy industry circa 2008. International Dairy Journal 31:S29-S40.