J. Dairy Sci. 97:5645–5656
http://dx.doi.org/10.3168/jds.2013-7724

A statistical filtering procedure to improve the accuracy of estimating population parameters in feed composition databases

P. S. Yoder,*[1] N. R. St-Pierre,† and W. P. Weiss*[2]
*Department of Animal Sciences, Ohio Agricultural Research and Development Center, The Ohio State University, Wooster 44691
†Department of Animal Sciences, The Ohio State University, Columbus 43210

ABSTRACT

Accurate estimates of mean nutrient composition of feeds, nutrient variance (i.e., standard deviation), and covariance (i.e., correlation) are needed to develop a more quantitative approach of formulating diets to reduce risk and optimize safety factors. Commercial feed-testing laboratories have large databases of composition values for many feeds, but because of potentially misidentified feeds or poorly defined feed names, these databases are possibly contaminated by incorrect results and could generate inaccurate statistics. The objectives of this research were to (1) design a procedure (also known as a mathematical filter) that generates accurate estimates of the first 2 moments [i.e., the mean and (co)variance] of the nutrient distributions for the largest subpopulation within a feed in the presence of outliers and multiple subpopulations, and (2) use the procedure to generate feed composition tables with accurate means, variances, and correlations. Feed composition data (>1,300,000 samples) were collected from 2 major US commercial laboratories. A combination of a univariate step and 2 multivariate steps (principal components analysis and cluster analysis) were used to filter the data. On average, 13.5% of the total samples of a particular feed population were removed, of which the multivariate steps removed the majority (66% of removed samples). For some feeds, inaccurate identification (e.g., corn gluten feed samples included in the corn gluten meal population) was a primary reason for outliers, whereas for other feeds, subpopulations of a broader population were identified (e.g., immature alfalfa silage within a broad population of alfalfa silage). Application of the procedure did not usually affect the mean concentration of nutrients but greatly reduced the standard deviation and often changed the correlation estimates among nutrients. More accurate estimates of the variation of feeds and how they tend to vary will improve the economic evaluation of feeds and risk assessment of diets, and provide the ability to implement stochastic programming.
Key words: feed composition, variation, data auditing

INTRODUCTION

Diet formulation models assume that the nutrient composition of individual ingredients is known with certainty and that the dietary concentration of a nutrient equals the weighted average of the nutrients provided by each ingredient. Both assumptions are invalid to varying degrees. Based on a survey, variation in composition of feeds within and among dairy farms is substantial (Weiss et al., 2012). Variability in composition can be incorporated into diet optimization by using stochastic programming, which considers the uncertainty of ingredient composition (St-Pierre and Harvey, 1986), but accurate data on nutrient variation (variances and covariances) are required. Some commercial feed analysis labs and the NRC (2001) provide standard deviation estimates for common feeds. However, small sample sizes, old data, misidentified feed samples, and other problems could cause substantial errors in the standard deviation estimates (Weiss and St-Pierre, 2009).

Mathematically, nutrients are either simple (e.g., CP) or complex linear (e.g., soluble protein) or complex nonlinear (e.g., NE_L) functions of simple nutrients. To estimate the variance of complex nutrients, the covariance between the simple nutrients must be considered or the variance estimate will be erroneous if simple nutrients within a feed do not vary independently. For example, NDF and starch concentrations in corn silage have a strong negative correlation (Dardenne et al., 1993); therefore, these 2 nutrients should not be considered independent when estimating the distribution of NE_L in corn silage.

Univariate methods have not been able to identify feed populations with high precision (Yu, 2005). Because nutrient concentrations within a feed sample are inherently correlated, multivariate methods have been

Received November 15, 2013.
Accepted May 23, 2014.
[1]Current address: Perdue AgSolutions LLC, Salisbury, MD.
[2]Corresponding author: weiss.6@osu.edu

suggested as a means to discriminate and classify feed samples (Yu, 2005; Gallo et al., 2013). Cluster analysis and principal components analysis (**PCA**) have been used to distinguish feed samples within a feed population by using the correlated nutrient concentrations of feed samples (Yu, 2005; Gallo et al., 2013) and these multivariate analyses allow for a large number of variables to be explained in fewer dimensions without the collinearity associated with using linear regression. To our knowledge, correlation matrices have not been summarized for feed populations; this information is needed when estimating the variances of complex linear and nonlinear nutrients. Our first objective was to develop a procedure that generates accurate and robust estimates of the first 2 moments [mean and (co)variance] of the nutrient distributions for the largest subpopulation within a feed in the presence of outliers and multiple smaller subpopulations. The second objective was to use that method to develop composition tables that contained accurate estimates of variation and correlation matrices among nutrients within feedstuffs. This information is necessary for the application of stochastic diet formulation.

MATERIALS AND METHODS

Feed Database

Feed composition data were provided by 2 major commercial laboratories [Cumberland Valley Analytical Services Inc. (Hagerstown, MD) and Dairyland Laboratories Inc. (Arcadia WI)]. Our objective was not to compare laboratories but to obtain a diverse set of data. Therefore, all statistical procedures were carried out within each laboratory-feed classification. The original database was large (>1,300,000 feed samples) and contained information on most common feeds used in the United States. Data were limited to samples submitted from 2003 through 2011, except for corn silage. Because of the large number of samples and because of changing genetics, corn silage samples submitted from 2007 through 2011 were used. For most samples, the feed classification code provided by the laboratory (which usually was provided by the original supplier of the sample) was used; however, apparently synonymous feeds were combined into single feeds (e.g., feeds designated as milo silage were added to the sorghum silage data set).

Statistical Methods

A 3-step statistical procedure was developed to identify univariate and multivariate outliers within a feed class and to identify outlying subpopulations (e.g., vegetative wheat silage within the broader population of wheat silage). When identified, outliers and subpopulations were removed and then means, standard deviations, and covariance matrixes were estimated from the remaining records. For all feeds except animal protein meals, statistics were calculated for concentrations of DM, CP, NDF, ash, lignin, neutral detergent-insoluble CP (**NDICP**), acid detergent-insoluble CP (**ADICP**), soluble CP, fat, starch, and sugar. For animal proteins, statistics were calculated for DM, CP, ash, NDICP, ADICP, soluble CP, and fat. The analytical methods used are described at each laboratory's websites (Cumberland Valley Analytical Services, 2014; Dairyland Laboratories, 2014). All statistical procedures were conducted using SAS/STAT software (version 9.3; SAS Institute Inc., Cary, NC).

Univariate Analysis (Step 1). The UNIVARIATE procedure was used to calculate the mean and standard deviation for each nutrient, and any sample that had a nutrient concentration more than 3.5 standard deviation units from the mean for 1 or more nutrients was removed. During this step, some anomalies within feed classes were noted, especially for feeds that can exist in different forms with respect to DM (e.g., hay or silage). Histograms for DM concentrations were generated and when bimodal distributions were observed, the wet and dry feed samples were separated visually based on the sample distributions. Forage samples with >70 to 80% DM (the specific breakpoint for a given forage was chosen based on natural breaks in the sample distribution) were moved to the appropriate hay classification and samples classified as hay that had less than 70 to 80% DM were moved to the appropriate silage category. No clear bimodal distribution was found for dry or high-moisture corn grain; therefore, corn grain samples with ≥84% DM were classified as dry corn and samples with <84% DM were classified as high-moisture corn. The univariate step was rerun after these changes in feed classification were made.

PCA (Step 2). For multivariate analyses (steps 2 and 3), only 3 to 6 nutrients were used because too many samples did not have information for all 11 nutrients and with multivariate analysis, a data record cannot be used unless all required variables are present. For the majority of feeds, DM, CP, NDF, and ash were the minimum set of nutrients used for multivariate analyses [Table 1; Supplemental Tables S1–S3 and supplemental spreadsheet file (http://dx.doi.org/10.3168/jds.2013-7724)]. One or 2 additional nutrients were added for certain feeds when they might be important nutritionally. Lignin was included for most forages, starch was included in feeds that contained grain, fat was included in feeds that could contain substantial concentrations of fat, and NDICP was included in distillers dried grains.

Table 1. Nutrients used in the multivariate steps and sample removal rates for selected feedstuffs (full list can be found in Supplemental Tables S1–S3; http://dx.doi.org/10.3168/jds.2013-7724)

Feed	Additional nutrients used[1]	Starting no.[2]	Removal rate,[3] %				Final clusters[4]
			UNI	PCA	CLS	Total	
Forage							
Alfalfa silage	NDF, LIG, ash	13,276	4.3	0.9	0	5.1	2[5]
Barley silage	NDF, LIG, ash, STA	4,123	5.5	1.9	0	7.3	2
Corn silage	NDF, LIG, ash, STA	109,190	3.9	1.7	0	5.5	1
Grass hay	NDF, LIG, ash	16,709	4.5	1.5	3.6	9.2	2
Rye silage	NDF, LIG, ash	7,849	8.2	1.0	5.4	14.0	3[5]
Corn grain							
Dry ground	NDF, ash, STA	4,876	5.7	3.8	14.5	23.1	2
High moisture	NDF, ash, STA	8,376	2.1	1.0	0	3.2	2
Oilseed meals							
Canola meal	NDF, ash	642	5.5	3.2	12.1	19.5	1
Soybean meal	None	247	5.3	2.5	0	7.6	2
By-product							
Brewers grain, wet	NDF, ash	1,010	3.0	1.1	4.6	8.4	3
Cottonseed hulls	NDF, ash	197	4.6	3.2	57.1[6]	60.9	1
Corn gluten feed (dry)	NDF, ash	438	4.8	2.4	14.5	20.5	2
Corn gluten meal	NDF, ash	127	3.1	2.4	46.7[7]	50.4	2
Feather meal	Fat	260	1.5	0	4.7	6.1	3
Hominy	NDF, ash, STA, fat	144	4.9	4.3	9.0	17.2	2[8]
Mean[9]	—	5,719	4.6	1.7	7.7	13.5	2.1

[1]Additional nutrients used in multivariate procedures; LIG = lignin; STA = starch. All feeds used DM and CP.

[2]Starting no. = number of records within a specific feed classification provided by a laboratory.

[3]Removal rates for univariate (UNI), principal components analysis (PCA), and clustering (CLS) steps. Removal rate = (number of records to which the step was applied − number of records remaining after the step)/records to which the step was applied × 100. Total removal = (starting number of records − final number)/starting records × 100.

[4]Number of clusters that were retained and combined into the final data set.

[5]Clusters were tentatively identified as representing different maturity classes.

[6]A removed cluster appeared to be whole cottonseed.

[7]A removed cluster appeared to be corn gluten meal.

[8]The 2 clusters were tentatively identified as representing low- and high-fat hominy.

[9]Data in this row is for entire data set (see Supplemental Tables S1–S3; http://dx.doi.org/10.3168/jds.2013-7724).

Prior to PCA, the variables were standardized to a mean of zero and a standard deviation of 1 using PROC STANDARD. The standardized variables were then used in PCA using PROC PRINCOMP with the COV option. The COV option forces the covariance matrix rather than the correlation matrix to be used for estimating principal components. The COV option causes the variables with large variances to be associated with the principal components that explain the most variance. The OUT statement was used to generate a data set containing the PCA scores for each principal component of the feed samples.

Principal components analysis transforms a set of correlated variables into a set of uncorrelated variables, each one being a linear function of the original variables, with the complete set of transformed variables explaining the total variance of the data. Each principal component is a linear transformation of the original variables, with derived coefficients that explain a portion of the variance of the data, and all new variables being independent of each other. The derived coefficients are based on the covariance matrix and these principal components can be useful for multivariate outlier detection (Gnanadesikan, 1977). However, PCA does not describe all types of association structures within a data set, specifically nonlinear structures. Using the derived coefficients of each principal component, PCA scores were generated for each feed sample. An improbable structure of nutrient concentrations within a feed sample results in an improbable (i.e., outlier) PCA score for a principal component (Gnanadesikan, 1977; Walczak and Massart, 1995; Xie et al., 2008). Outliers are samples that do not fit the principal component linear model (e.g., variation not explained by PCA) and can be evaluated by plotting the PCA scores or using the Hotelling's T^2 statistic (Eriksson et al., 2006). For our purposes, we used the standard deviation method for distinguishing which PCA scores were outliers. Standard deviations of PCA scores were estimated using PROC UNIVARIATE and samples with PCA scores $\geq$3.5 standard deviations from the mean PCA score for each principal component were removed. All principal components were used for identifying outliers because even the smallest principal components (i.e., explain-

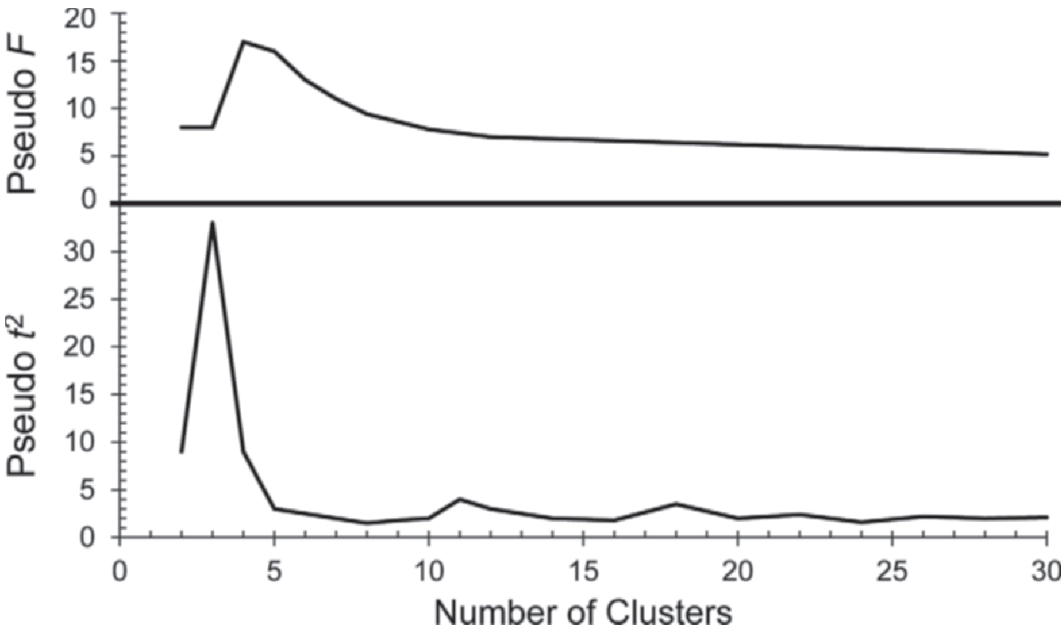

Figure 1. Pseudo F statistic (**PSF**) and pseudo t^2 statistic (**PS2**) by number of clusters following clustering analysis. When the maxima for PSF and PS2 occurred at similar levels of clustering (i.e., PS2 preceded PSF by 1 or 2 clustering levels), then the level at the maxima PSF was used for creating the number of clusters. In this example of dry citrus pulp (n = 154), the level of clustering was determined to be 4. The maxima for PS2 occurred at a clustering level of 3 (PS2 score = 32.6) and the maxima for PSF at a clustering level of 4 (PSF score = 18.8).

ing the least amount of variance) are useful for outlier detection (Shen, 2007).

Clustering (Step 3). Two-stage density linkage cluster analysis was performed on feed samples remaining after PCA analysis to identify whether outlier groups of samples were present. Two-stage clustering was determined to be the optimal method for dealing with clusters that have very different covariance matrices or are not normally distributed (SAS Institute, 2014), and it was the most accurate method of 9 cluster algorithms for assigning simulated data to proper clusters (Chen et al., 1995). Clustering was accomplished using PROC CLUSTER with the TWOSTAGE method and number of neighbors (k) was set at $k = n^{0.3}$, where n = the number of samples in the data set Wong (1982). The pseudo F statistic (**PSF**) and pseudo t^2 statistic (**PS2**) of PROC CLUSTER were evaluated and plotted by the number of clusters to determine the optimal number of clusters within the data (Figure 1). The separation among clusters at a current clustering level is measured by PSF (i.e., a high PSF score at a given level of clustering indicates the clusters are well separated). The other statistic, PS2, measures whether the current level of clustering has resulted in poorly separated clusters (i.e., a high score indicates the clusters are poorly separated). As an example, a high PSF at a clustering level of 3 followed by a high PS2 at a clustering level of 2 indicates that the population contains 3 well-defined clusters and that combining these 3 clusters into 2 clusters is not optimal. When the local maxima for PSF and PS2 occurs at similar levels of clustering (i.e., PS2 preceded PSF by 1 or 2 clustering levels), then the level at the local maxima PSF indicates an appropriate level of clustering for the data (SAS Institute, 2014). Once the number of clusters was determined, the TREE procedure with the n option equaling the specified number of clusters was performed. If the local maxima for PSF and PS2 did not occur at similar levels of clustering, only 1 cluster was formed.

If more than 1 cluster was identified, the following procedure was followed. Arbitrarily, clusters with frequencies that were $\leq$10% of the data set used in the clustering procedure were considered outliers and removed. Clusters were removed in a stepwise fashion and the cluster algorithm was rerun following each cluster removal until no additional clusters had $\leq$10% of the observations in the data set. For most feeds, clusters with >10% of the data set were recombined and designated as the final feed population. However, for several feeds, clusters could be provisionally identified based on mean nutrient composition and those clusters were moved to the corresponding population or were identified as a unique feed. For example, clustering identified a subpopulation within the corn gluten meal population that, based on composition data, was probably dry corn gluten feed.

The data in the samples remaining after the univariate, PCA, and cluster steps were then analyzed using PROC UNIVARIATE and CORR to calculate statistics and the correlation matrix for each defined population. If samples contained concentrations of nutrients not used in the first 3 steps (e.g., ADICP) that were more than 3.5 standard deviation units from the mean, then these samples were removed. Example SAS input statements for the 3-step procedure are in the supplemental SAS input file (http://dx.doi.org/10.3168/jds.2013-7724).

Procedure Validation

A common way of evaluating a new statistical procedure is through Monte Carlo simulations (Monahan, 2001). Data sets with known properties are generated using a computer, the proposed procedure is applied to the generated data, and its outcomes are compared with the known properties of the data set. Various simulated data sets were used in the evaluation of our procedure. The question that was addressed through these simulations was whether estimates of the first 2 moments of the primary distribution were accurately estimated using the procedure when data originated from more than 1 population (i.e., from a primary and multiple secondary distributions).

Uncontaminated Data Set. A population of 100,000 simulated corn silage samples with hypothetical CP, NDF, and starch concentrations that were normally distributed was generated using Monte Carlo techniques (Fan et al., 2002). The specified population parameters were as follows: mean = 8.00, 40.00, and 32.00 and standard deviation = 1.00, 5.00, and 6.00 for CP, NDF, and starch, respectively, which came from a large set of samples analyzed by a commercial laboratory. The 3-step procedure was applied to the simulated data. Intentionally, this data set contained no outliers. The purpose of this step in the validation was to assess the procedure in a situation where it is not needed. A desirable procedure would produce estimates of the means, variances, and covariances that are similar to those used to generate the data.

Data Set with Normal Outlier Distributions. Three simulated outlier populations of 2,500, 2,500, and 5,000 samples were then added to the simulated 100,000 uncontaminated data set for a total population size of 110,000 samples (Table 2). Two of the simulated outlier populations had mean nutrient concentrations that were greater than 3.5 standard deviation units from the mean of the uncontaminated data set but similar variances and covariances as the uncontaminated data set. Multiple outlier subpopulations were simulated to mimic a feed population that is highly contaminated with multiple different feed populations (e.g., by-product feeds originating from multiple manufacturing sites). A multivariate normal distribution was used to generate these records (Fan et al., 2002). The third simulated outlier population contained 5,000 records from a multivariate normal distribution with means less than 3.5 standard deviation units from the mean of the uncontaminated data set but with variances and covariances that differed greatly from those used for the uncontaminated data set. The variances were 50% less and the covariances were almost the opposite (i.e., −0.89 vs. 0.70) compared with the uncontaminated data set. The resulting data set (110,000 records) contained 100,000 records from perfectly multivariate normal data and no outliers plus 10,000 records from multivariate normal data that differed in their distribution (either in their means, variances, or covariances) from the distribution of the uncontaminated data set. This does not mean that each of the additional 10,000 records was itself an outlier, but that the resulting combined data set of 110,000 records contained data from 4 entirely different distributions, all multivariate normal.

Data Set from Log-Normal Distributions with Outliers. Although normality was never implicitly assumed during the development of the procedure, some of its parameters (i.e., 3.5 SD in the univariate step) were selected based on normal theory. More importantly, the distribution of some nutrients within feedstuffs is possibly nonnormal. Therefore, the performance of the proposed procedure must be evaluated under nonnormal conditions. For the normal and log-normal distributions (unlike the binomial, Poisson, negative binomial, exponential, and gamma distributions), the first 2 moments are independent of each other (i.e., the mean and the variances are 2 entirely separate parameters). But unlike the normal distribution, the log-normal is a highly skewed distribution, showing very long right tails (positive skewness), something that is characteristic of the distribution of some nutrients (e.g., ash). Similar to the data set with normal outlier distributions, 110,000 records (Table 2) were generated from 1 multivariate normal distribution (n = 100,000) and 3 (total n = 10,000) multivariate log-normal distributions (Fan et al., 2002).

RESULTS AND DISCUSSION

Procedure Validation

Uncontaminated Data Set. When applied to the data set containing 100,000 records generated from a multivariate normal distribution without any outliers, the procedure removed 242 samples or 0.24% of the population, and the mean, standard deviation, and cor-

Table 2. Simulated data sets to assess the accuracy and robustness of the proposed procedure

Data set	Mean	SD	Correlation		
			CP	NDF	Starch
Uncontaminated data set[1]					
CP	8.00	1.00	1		
NDF	40.02	5.02	0.262	1	
Starch	31.98	6.03	−0.471	−0.891	1
Uncontaminated data set after procedure[2]					
CP	8.00	1.00	1		
NDF	40.02	4.99	0.261	1	
Starch	31.98	6.00	−0.470	−0.891	1
Data set with normal outlier distributions[3]					
CP	8.16	1.51	1		
NDF	39.22	7.57	−0.468	1	
Starch	32.93	9.08	0.378	−0.946	1
Data set with normal outlier distribution after procedure[4]					
CP	8.00	1.00	1		
NDF	40.02	5.01	0.261	1	
Starch	31.98	6.02	−0.469	−0.891	1
Data set with log-normal outlier distributions[5]					
CP	8.24	1.81	1		
NDF	39.44	7.29	−0.455	1	
Starch	32.90	9.19	0.436	−0.949	1
Data set with log-normal outlier distributions after procedure[6]					
CP	8.00	1.00	1		
NDF	40.02	5.01	0.261	1	
Starch	31.98	6.02	−0.469	−0.891	1

[1]Data set of 100,000 records generated from a multivariate normal distribution with parameters set to those of the population parameters.

[2]Data set of 99,758 remaining records after an uncontaminated data set (n = 100,000) was processed by the 3-step procedure.

[3]Data set of 110,000 records; 100,000 were generated from a multivariate normal distribution with parameters set to those of the population parameters and 10,000 were generated from 3 multivariate normal distributions with very different means, variances, and correlations than those of the uncontaminated data set.

[4]Data set of 99,950 remaining records after an uncontaminated data set (n = 100,000) combined with 3 multivariate normal distributed populations (n = 10,000) were processed by the 3-step procedure.

[5]Data set of 110,000 records; 100,000 were generated from a multivariate normal distribution with parameters set to those of the population parameters and 10,000 were generated from 3 multivariate log-normal distributions with very different means, variances, and correlations than those of the uncontaminated data set.

[6]Data set of 99,925 remaining records after an uncontaminated data set of (n = 100,000) combined with 3 multivariate log-normal distributed populations (n = 10,000) were processed by the 3-step procedure.

relation estimates were all within 0.6% of their true values (Table 2). The removal of a few extreme records from the simulated data had little influence on the estimates of the parameters of the first 2 moments of the distribution. Therefore, the consequences of applying the procedure in cases where it is not needed would be negligible with regard to the accuracy of the estimates of the means and (co)variances.

Data Sets with Normal Outlier Distributions. The 3-step procedure removed 10,050 or 9.1% of the total 110,000 records. The mean, standard deviation, and correlation estimates of the resulting population was almost the same (differences <1.0%) as those of the primary population of 100,000 feed samples (Table 2). Various simulations conducted with differing types and frequency of contamination from outlier distributions resulted in similar outcomes (differences <2.0%). Thus, the proposed procedure produced accurate estimates of the first 2 moments of the primary population when the sample set was contaminated by a large number of records from different populations, all being normally distributed.

Data Sets from Log-Normal Distributions of Outliers. The 3-step procedure removed 10,075 or 9.2% of the total 110,000 records. The mean, standard deviation, and correlation estimates of the resulting population was almost the same (differences <1.0%) as those of the primary population of 100,000 feed samples. Various simulations conducted with differing types and frequency of contamination from multivariate log-normal outlier distributions resulted in similar outcomes (differences <2.0%). Thus, the proposed procedure produced accurate estimates of the first 2 moments of the primary population when the sample set was contaminated by log-normal-distributed outlier populations.

Application of the Procedure

The 3-step procedure was applied to 115 feed populations (i.e., feed classifications provided by the laborato-

ries) from a total of approximately 1,300,000 analytical records provided by 2 commercial feed laboratories. On average (Table 1), the univariate steps removed 4.6% of the samples from each starting population (range from 0.3 to 11.2%). If a nutrient distribution was perfectly normal, approximately 0.02% of the records should fall outside 3.5 standard deviation units from the mean or about 0.22% of samples when applied to 11 different nutrients (assuming that the nutrients are uncorrelated to each other). On average, our removal rate was approximately 20 times greater (and in the worst case, 50 times greater) than what is expected from a multivariate normal distribution with null correlations, indicating substantial misclassification for many of these feed samples (90% of all feed populations had ≥2% of samples removed by the univariate step) or that some of the populations were not multivariate normal.

The PCA step removed an average of 1.7% of the samples remaining after the univariate step (range was 0 to 5.3%), and the clustering step removed an average of 7.7% of the samples remaining after the PCA step (range was 0 to 61.1%). The term "removed" as used here and in the previous paragraph should be interpreted in the context of a statistical filter. In each step, the procedure produces virtual bins; each record can only belong to 1 bin. The bin with the largest number of records in the end is considered the primary bin from which the means, variances, and covariances of the primary population are estimated. A by-product of the multivariate analysis is an improved ability to identify unique subpopulations within individual feeds. For example, outlier PCA scores for distillers dried grains represented samples that individually were distinctly different in their feed covariance structure (Figure 2). After the distillers dried grains samples were identified as outliers to the primary population, we examined the user-identified names of the samples and 24% of the outliers were identified as high-protein distillers grains, 20% as wheat middlings, 18% as wheat distillers grains, 3% as soybean meal or canola meal mixtures, and 35% either as distillers grains or an unrecognizable name. Almost 61% of the outliers (n = 11) identified by the first principal component and 81% of the outliers (n = 13) in the second principal component were identified by the user as high-protein distillers grains. Approximately 80% of the outliers in the third principal component were identified as wheat-based distillers grains or wheat middlings (n = 36). The univariate method (discarding samples that were >3.5 SD units from the mean) did not detect samples that likely should not be labeled as distillers dried grains. This illustrates the robustness of multivariate techniques to identify misclassified samples. Gallo et al. (2013) was able to identify forage populations and differentiate between high- and poor-quality forages using PCA.

On average, 13.5% of the observations were removed as outliers [i.e., not belonging to the primary population; Tables 1 and 3; Supplemental Tables S1–S3 (http://dx.doi.org/10.3168/jds.2013-7724)]; however, the percentage removed varied from <3% (e.g., sorghum silage, sudangrass hay, timothy hay, and mixed hay) to >45% for corn stalks, corn gluten meal, and cottonseed hulls. More than half of all feed populations had at least 10% of their records removed (16.5% of the feed classes had ≥20% of the records removed) which indicates (1) substantial misclassification, (2) outliers in the original data set, or (3) more than one subpopulation in a given feed. The largest removal rate occurred at the clustering step, which suggests that many of these records contained a substantial number of misidentified samples (e.g., dry corn gluten feed labeled as corn gluten meal) or that the populations were broadly defined and included subpopulations (e.g., immature alfalfa and mature alfalfa within a single population).

Identification of Subpopulations of Feeds

Across all feeds, 18% consisted of a single cluster, 56% had 2 identified clusters, 20% had 3 clusters, and 6% had >3 clusters. More than 60% of the samples were removed from the cottonseed hull population but of the 120 samples removed, 101 of them belonged to a single cluster that averaged 23.6% CP, 46% NDF, and 21% fat. This cluster almost certainly comprises whole cottonseed samples that were misidentified as cottonseed hulls. Almost 50% of the samples in the corn gluten meal class were identified as outliers and removed; however, 57 of the 64 samples removed belonged to 1 cluster that averaged 23.8% CP and 37.3% NDF, which is near the expected composition of corn gluten feed. Expellers soybean meal had a high removal rate (34.4%) because the initial set contained several samples that were probably conventional soybean meal (cluster that averaged 1.9% fat). Another identified cluster averaged 17% fat, which suggests that many of those samples were likely full-fat soybean products.

For some feeds, large subpopulations (>10% of the initial set) were removed from the initial set and provisionally considered a unique, definable feed with its own composition data. For example, distillers dried grains formed 2 well-separated clusters (Table 4). The smaller cluster had substantially greater average concentrations of NDF and NDICP than the larger cluster. We provisionally identified the larger and smaller populations as samples that had been assayed for NDF with and without the use of sodium sulfite, respectively

Table 3. Simple statistics for the feed populations (115 different populations representing different feed classifications and laboratories) and sample removal rates

Feed population	Number of samples	Samples removed,[1] %
Starting population		
Mean	5,719	13.5
Median	470	10.6
Range	63 to 166,545	2.2 to 60.9
Final population[2]		
Mean	5,256	
Median	412	
Range	54 to 158,458	

[1]Percentage of samples in the starting populations provided by the commercial feed laboratories that were identified as outliers or subpopulations and removed to obtain the final populations.

[2]Statistics obtained after the data sets were filtered and outliers were removed using the 3-step statistical procedure outlined in the paper.

(some nutritional models require NDICP to be assayed without sodium sulfite). The NDF in distillers grains has substantial protein contamination when assayed without sodium sulfite (Dong and Rasco, 1987), resulting in higher concentrations of NDF and NDICP. The value of using a multivariate approach rather than a simple, univariate cut-off value for NDICP is evident in the ranges remaining in the clusters (Table 4). The average NDICP is much greater in cluster A (i.e., samples likely assayed without sodium sulfite) than cluster B but substantial overlap is observed between the 2 clusters.

Almost all haycrop forage populations (including hays and silage, legumes, cool-season grasses, warm-season grasses, and whole-plant small grains such as wheat and oats) separated into 2 or 3 clusters. Orchardgrass silage, pasture hay, sudangrass silage, and wheat silage separated into 2 clusters, which, based on average composition data, were identified as less mature or more mature. Sorghum silage separated into 2 clusters that were tentatively identified as silage made from forage sorghum and silage made from grain sorghum. A large set of data (n = 158,357 and 61,637 silage and hay samples, respectively) with nondescriptive classification codes (i.e., haycrop silage and haycrop hay) was processed using this procedure and separated into 4 (hay) or 5 (silage) major clusters (Supplemental Table S1; http://dx.doi.org/10.3168/jds.2013-7724). Based on average nutrient composition of the clusters and correlations among nutrients (e.g., CP and NDF and lignin and NDF) we could putatively identify the subpopulations based on maturity (i.e., immature, medium maturity, and more mature mixed grass and legume) and on forage type (i.e., predominantly grass or predominantly legume). Because we do not have specific sample information, the accuracy of these identifications cannot be determined. Many feed libraries in currently available ration software programs characterize forage maturity classes using a univariate approach based on NDF (e.g., mature grass silage has >60% NDF; NRC, 2001). The variation in nutrient concentrations within those populations is likely underestimated because using univariate cut-off points remove the tails of a distribution of a nutrient. Corn silage had 1 cluster but samples identified as brown midrib corn silage separated into 2 clusters

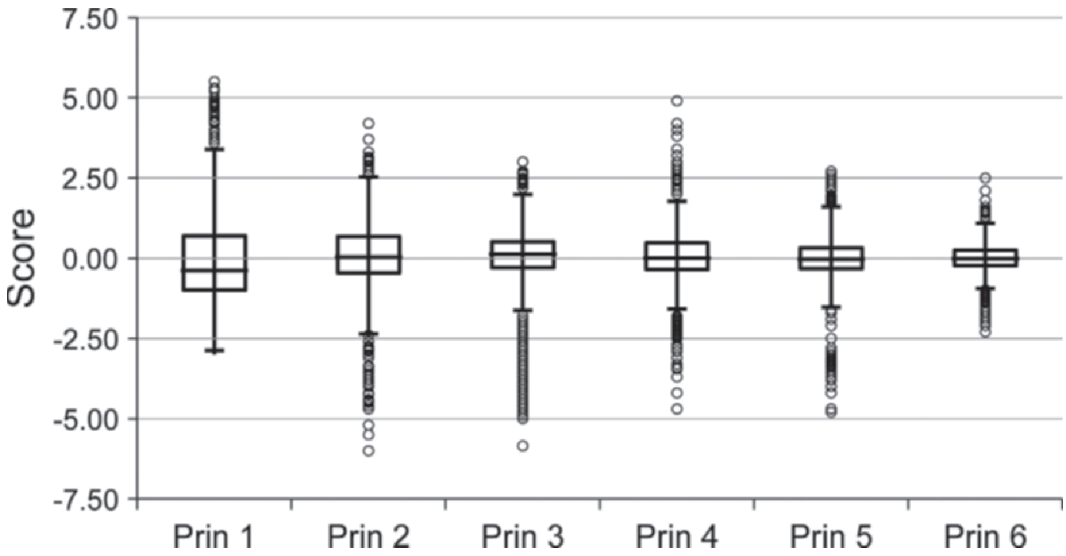

Figure 2. Boxplot of principal components analysis (PCA) scores for distillers dried grains (n = 2,276) by principal component (Prin 1 to Prin 6) after removal of univariate outliers but before removal of outliers identified by steps 2 and 3 of the described procedure. The box represents the interquartile range (IQR); the horizontal line within each box is the median score and the whiskers represent 1.5 times the IQR (which is approximately equal to 2.7 SD units). The circles represent observations that are outside 1.5 × IQR. Observations >3.5 SD units were deemed outliers and removed during step 2.

Table 4. Nutrient concentration statistics (% of DM) and correlations for the broad population of distillers dried grains and 2 subpopulations that were identified using a statistical clustering procedure

Item	Population[1]	Cluster A[2]	Cluster B[3]
No. of samples	1,906	453	1,453
CP			
Mean	29.4	29.3	29.5
SD	1.49	1.79	1.38
Range	24.7 to 40.3	25.2 to 40.3	24.7 to 36.9
NDF			
Mean	35.0	41.4	32.9
SD	4.45	3.32	2.35
Range	26.4 to 52.1	32.3 to 52.1	26.4 to 43.9
NDICP			
Mean	5.00	8.68	3.86
SD	2.47	1.83	1.18
Range	0.76 to 14.70	3.86 to 14.70	0.76 to 10.0
Correlation (r)			
CP and NDF	−0.003	0.266	−0.033
CP and NDICP	0.034	0.122	0.171
NDF and NDICP	0.740	0.100	0.275

[1]Population derived from a data set provided by a commercial laboratory after the data set was processed using the first 2 steps of the outlined statistical procedure.

[2]Based on average nutrient concentrations, cluster A was provisionally identified as samples that were assayed for NDF and neutral detergent-insoluble CP (NDICP) without sodium sulfite.

[3]Based on average nutrient concentrations, cluster B was provisionally identified as samples that were assayed for NDF and NDICP using sodium sulfite.

that could not be provisionally identified. Other feed populations that had 2 or more clusters included many grains and most by-products. For example, wet brewers grains split into 3 clusters (19.5, 29.1, and 51.4% of the total population). These different subpopulations could represent different breweries with different recipes for beer or processing procedures for the spent grains.

In most cases (exceptions described above), major clusters were recombined and formed the final population. This does not necessarily mean that a cluster was not a definable population; it simply means that we did not have enough information regarding the samples to allow us to tentatively or provisionally define the subpopulation.

Effects of Sample Removal on Population Statistics

A primary objective for developing this method was to build a feed database that has accurate estimates of variances for important nutrients and correlations among nutrients [see example in supplemental spreadsheet file (http://dx.doi.org/10.3168/jds.2013-7724) and at http://dairy.osu.edu/resource/OSU%20Dairy%20Pubs.html#FeedComposition]. Accurate standard deviation and correlation estimates are important for determining the optimal sampling frequency of ingredients in a diet (St-Pierre and Cobanov, 2007), the application of stochastic diet formulation (St-Pierre and Harvey, 1986), and in determining the full economic value of feeds while accounting for uncertainty in feed composition (St-Pierre and Harvey, 1986). Across most feeds, the mean nutrient concentration was similar before and after removal of records (Table 5), but the standard deviation was almost always lower after removal of outliers. The reduction in standard deviation varied among feeds and nutrients. For example, the standard deviation for NDF in alfalfa silage decreased 3.5% compared with a 48 and 76% reduction in the standard deviation for canola meal and corn gluten meal CP concentrations, respectively. The standard deviation estimates after removal of outliers were often less than those in NRC (2001).

The correlation matrix changed in many cases following removal of outliers (Table 6). Removing outliers using PCA and clustering should change correlation estimates because the covariance structure of a feed is used by those methods to identify outliers. Correlation matrices provide some understanding as to how multiple nutrient concentrations are expected to vary and deviate from the average nutrient concentration. For example, a corn silage sample can have a high concentration of NDF or a high concentration of starch, but because of expected relationships between the 2 nutrients (Dardenne et al., 1993), a corn silage sample with high concentrations of both NDF and starch is improbable. The error in probability from ignoring the correlation can be calculated for this example, assuming normal and bivariate normal distributions and mean concentrations of 41.9 and 31.2%, with SD = 4.73 and 6.28% for NDF and starch concentrations, respectively.

Table 5. Means and SD for CP and NDF in selected feeds provided by the NRC (2001) and from data provided by commercial laboratories before and after outliers were removed using the described statistical method

	CP, % of DM			NDF, % of DM		
Feed	NRC[1]	Before	After	NRC[1]	Before	After
Alfalfa silage						
No.[2]	8,576	13,276	11,020	8,567	13,276	11,020
Mean	20.0	20.7	20.8	45.7	42.2	41.8
SD	3.0	2.5	2.4	6.5	5.7	5.5
Canola meal						
No.	230	642	515	81	642	515
Mean	37.8	40.5	41.1	29.8	28.2	28.1
SD	1.1	4.64	2.41	6.6	4.24	3.01
Corn gluten meal						
No.	57	127	63	39	127	63
Mean	65.0	46.2	68.1	11.1	21.6	7.2
SD	7.8	22.9	5.4	10.1	15.7	3.3
Corn grain (dry)						
No.	4,845	4,876	3,752	4,729	4,876	3,752
Mean	9.2	8.8	8.5	10.3	11.6	11.3
SD	0.9	1.4	0.57	2.7	2.7	1.2
Cottonseed hulls						
No.	134	197	77	106	197	77
Mean	6.2	16.1	6.4	85.0	60.3	80.8
SD	3.6	8.92	1.91	5.9	18.97	5.49
Corn silage						
No.	1,033	109,190	103,119	1,033	109,190	103,119
Mean	8.8	8.0	7.9	45.0	42.2	41.9
SD	1.2	1.07	0.94	5.3	5.23	4.73
Soybean meal (expellers)						
No.	546	1,228	806	70	335	193
Mean	46.3	46.1	47.1	21.7	17.9	17.5
SD	3.2	5.70	1.87	8.0	7.21	5.26

[1]From the feed composition table in NRC (2001).

[2]No. = number of samples in data set.

The univariate probability that a sample would exceed 44.2% NDF is easily found using a standard normal distribution with a standard score (z-value) equal to $(44.2 - 41.9)/4.73 = 1.4862$; this probability is 0.32. Likewise, the probability that a sample would exceeds 34.2% starch is also equal to 0.32. If the correlation between NDF and starch is ignored (i.e., the nutrients are incorrectly considered independent), then the probability that a corn silage sample would exceed both 44.2% NDF and 34.2% starch would be 10% [i.e., $(0.32 \times 0.32) \times 100$). However, if the correlation between NDF and starch ($r = -0.88$) is accounted for, then the actual probability decreases to 0.19%, which is 50 times less (the calculation of the probability for the bivariate normal distribution requires a numerical integration; see Monahan, 2001, chapter 12). Expressed differently, one would expect only 19 out of 10,000 samples to exceed both 44.2% NDF and 34.1% starch as opposed to 1,000 out of 10,000 if the correlation is ignored during the calculations.

Knowledge of nutrients and their relationships will help understand how NE_L varies from sample to sample. Weiss (1998) showed that alfalfa silage with a constant 35% ADF does not result in a constant NE_L value, but ranges from 1.28 to 1.44 Mcal/kg because of different concentrations of NDF, lignin, and ash, none of which are perfectly correlated with ADF. The correlation matrix of alfalfa silage shows that critical nutrients such as CP, NDF, and ash are not perfectly correlated. Imperfect correlations among nutrients illustrate the problem of using a single nutrient to estimate NE_L or MP, or using a single nutrient to estimate the economic value of a feed.

CONCLUSIONS

A 3-step statistical procedure was developed that can be used to quickly identify outliers and subpopulations of samples within a larger population of feed samples. Results from the simulation studies showed that applying the procedure in cases where it is not needed has negligible effects on the accuracy of the estimates of the means and (co)variances. In addition, the procedure provided accurate estimates of the means, variances, and correlations of nutrients when the database contained a large number of samples from different subpopulations (multivariate normal or log-normal distributions). Multivariate techniques applied to feed

Table 6. Correlation estimates for selected pairs of nutrients in some feeds before and after outliers were removed

	CP:NDF		Ash:NDF		NDF:lignin		Starch:NDF	
Item[1]	Before	After	Before	After	Before	After	Before	After
Alfalfa silage n = 11,020	−0.73	−0.74	−0.16	−0.27	0.52	0.55	−0.01	0.03
Barley silage n = 3,822	−0.12	0.09	0.17	0.23	0.54	0.59	−0.61	−0.65
Canola meal n = 515	−0.32	−0.11	−0.06	0.11	0.27	0.29	−0.03	0.05
Corn silage n = 103,119	0.23	0.20	0.38	0.37	0.76	0.76	−0.88	−0.88
Corn gluten feed n = 347	−0.51	0.24	−0.13	−0.27	0.03	−0.08	−0.14	−0.35
Distillers grains n = 1,906	0.11	0.00	−0.08	−0.03	0.39	0.28	0.08	0.27
Grass hay n = 13,874	−0.69	−0.58	−0.47	−0.30	0.72	0.64	−0.09	−0.02
Rye silage n = 6,719	−0.65	−0.61	−0.27	−0.20	0.62	0.58	−0.16	−0.13
Cottonseed, whole n = 294	−0.36	−0.37	0.03	−0.08	0.37	0.14	−0.43	−0.26

[1]Numbers are the number of samples in the final data set.

populations detected improbable feed samples and subpopulations that were not identified as outliers using univariate techniques. In many cases, the outliers and outlying populations likely were from apparent misidentification of samples. In other instances, subpopulations likely represented definable feeds. Application of this procedure to feed databases provided more accurate estimates for standard deviations and correlations for feed populations, which are important in stochastic programming, in economic valuation of feeds, and in the determination of optimal safety factors.

ACKNOWLEDGMENTS

This project was supported by National Research Initiative Competitive Grant no. 2009-55206-05242 from the USDA National Institute of Food and Agriculture (Washington, DC). We also express our sincere thanks to Ralph Ward of Cumberland Valley Analytical Services Inc. (Hagerstown, MD) and Dave Taysom of Dairyland Laboratories Inc. (Arcadia, WI) for providing the raw data that was essential for this project.

REFERENCES

Chen, S. K., P. Mangiameli, and D. West. 1995. The comparative ability of self-organizing neural networks to define cluster structure. Omega 23:271–279.

Cumberland Valley Analytical Services. 2014. Resources—Lab Procedures. Accessed Apr. 4, 2014. http://www.foragelab.com/Resources/Lab-Procedures.

Dairyland Laboratories. 2014. Methods. Accessed Apr. 4, 2014. http://www.dairylandlabs.net/feed-and-forage/methods.

Dardenne, P., J. Andrieu, Y. Barrière, R. Biston, C. Demarquilly, N. Femenias, M. Lila, P. Maupetit, F. Rivière, and T. Ronsin. 1993. Composition and nutritive value of whole maize plants fed fresh to sheep. Ii. Prediction of the in vivo organic matter digestibility. Ann. Zootech. 42:251–270.

Dong, F. M., and B. A. Rasco. 1987. The neutral detergent fiber, acid detergent fiber, crude fiber, and lignin contents of distillers' dried grains with solubles. J. Food Sci. 52:403–405.

Eriksson, L., P. L. Andersson, E. Johansson, and M. Tysklind. 2006. Megavariate analysis of environmental QSAR data. Part 1—A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). Mol. Divers. 10:169–186.

Fan, X., Á. Felsövályi, S. A. Sivo, and S. C. Keenan. 2002. SAS for Monte Carlo Studies: A Guide for Quantitative Researchers. SAS Inst. Inc., Cary, NC.

Gallo, A., M. Moschini, C. Cerioli, and F. Masoero. 2013. Use of principal component analysis to classify forages and predict their energy content. Animal 7:930–939.

Gnanadesikan, R. 1977. Multidimensional residuals and methods for detecting multivariate outliers. Pages 292–305 in Methods for Statistical Data Analysis of Multivariate Observations. 2nd ed. John Wiley & Sons Inc., New York, NY.

Monahan, J. F. 2001. Numerical Methods of Statistics. Cambridge University Press, New York, NY.

NRC. 2001. Nutrient Requirements of Dairy Cattle. 7th rev. ed. ed. Natl. Acad. Press, Washington, DC.

SAS Institute. 2014. SAS/STAT User's Guide, Version 9.2, 2nd ed. SAS Inst. Inc., Cary, NC.

Shen, Y. 2007. Outlier detection using the smallest kernel principal components. Ph.D. Diss. Temple University, Philadelphia, PA.

St-Pierre, N. R., and B. Cobanov. 2007. A model to determine the optimal sampling schedule of diet components. J. Dairy Sci. 90:5383–5394.

St-Pierre, N. R., and W. R. Harvey. 1986. Incorporation of uncertainty in composition of feeds into least-cost ration models. 2. Joint-chance constrained programming. J. Dairy Sci. 69:3063–3074.

Walczak, B., and D. L. Massart. 1995. Robust principal components regression as a detection tool for outliers. Chemometr. Intell. Lab. Syst. 27:41–54.

Weiss, W. P. 1998. Estimating the available energy content of feeds for dairy cattle. J. Dairy Sci. 81:830–839.

Weiss, W. P., D. Shoemaker, L. R. McBeth, P. Yoder, and N. R. St-Pierre. 2012. Within farm variation in nutrient composition of feeds. Pages 103–140 in Proc. Tri-State Dairy Nutr. Conf., Ft. Wayne, IN. Ohio State Univ., Columbus.

Weiss, W. P., and N. R. St-Pierre. 2009. Impact and management of variability in feed and diet composition. Pages 83–96 in Proc. Tri-State Dairy Nutr. Conf., Ft. Wayne, IN. Ohio State Univ., Columbus.

Wong, M. A. 1982. A hybrid clustering method for identifying high-density clusters. J. Am. Stat. Assoc. 77:841–847.

Xie, L., X. Ye, D. Liu, and Y. Ying. 2008. Application of principal component-radial basis function neural networks (PC-RBFNN) for the detection of water-adulterated bayberry juice by near-infrared spectroscopy. J. Zhejiang Univ. Sci. B 9:982–989.

Yu, P. 2005. Applications of hierarchical cluster analysis (CLA) and principal component analysis (PCA) in feed structure and feed molecular chemistry research using synchrotron-based Fourier transform infrared (FTIR) microspectroscopy. J. Agric. Food Chem. 53:7115–7127.