IDENTIFICATION OF ERRORS IN COTTON FIBER DATA SETS USING BAYESIAN NETWORKS

G F. Sassenrath

J. E. Boggess

Xintong Bi

H. C. Pringle

See next page for additional authors

Recommended Citation

Sassenrath, G F.; Boggess, J. E.; Bi, Xintong; and Pringle, H. C. (2004). "IDENTIFICATION OF ERRORS IN COTTON FIBER DATA SETS USING BAYESIAN NETWORKS," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1167

Author Information

G F. Sassenrath, J. E. Boggess, Xintong Bi, and H. C. Pringle

IDENTIFICATION OF ERRORS IN COTTON FIBER DATA SETS USING BAYESIAN NETWORKS

G.F. Sassenrath[1], *J.E. Boggess*[2], *Xintong Bi*[2], *H.C. Pringle*[3]
[1]USDA-ARS, APTRU, Stoneville, MS; [2]Department of Computer Science and Engineering, Mississippi State University; [3]Delta Research and Extension Center, Stoneville, MS

ABSTRACT

Cotton fiber is graded on a series of parameters based on physiological factors (strength, length, and thickness), lint color, and presence of non-lint matter such as leaves, stems or other foreign materials. Cotton lint is graded by the USDA-AMS after harvest and ginning, and the grade determines the price of the lint. Given the importance of cotton fiber quality to the value of the crop, the spatial variability of cotton fiber properties is of particular interest to researchers and producers in developing management scenarios for optimal profitability. Previous research studies have relied on hand-harvesting the cotton at intervals throughout the field to obtain a measure of the cotton fiber quality and the extent of spatial variability. However, hand-harvested cotton has different qualities than that harvested by machine and ginned in the large-scale production gins. Part of this arises from the difference in efficiency of harvest between machine and humans, and part results from the different gins used for the smaller sample sizes. While these studies have demonstrated the extent of spatial variability of fiber properties, hand-harvesting is not amenable to large-scale or production research efforts. Moreover, the differences in fiber properties limit the extension of the results to the production setting. We have developed a mechanism of sampling cotton from the cotton chute during mechanical harvest. The samples are then ginned on a research gin. This study was undertaken to develop a method of translating these small-scale researcher level results to full-scale production level results. The research reported here is the first step in that effort, and demonstrates the use of Bayesian networks to detect erroneous entries in cotton fiber data sets.

Keywords: Bayesian networks, cotton fiber quality, cotton fiber spatial variability, neural networks

1. INTRODUCTION

Two areas of agronomic research have simultaneously realized the impact of harvesting and ginning methods on the final fiber quality: precision agriculture and breeding. Because both of these areas of study are directed at improving profitability to the end user – the cotton producer – research results need to be comparable to those that a producer would find. In the research setting, scale issues have resulted in common practices of cotton harvesting and ginning that are different than those used by producers. Researchers routinely hand-harvest cotton (Johnson *et al.*, 2002), and then gin the cotton on small research gins. The research gins are equipped with saws that remove the lint from the seed, and then pass the lint through rollers to the collection area. No preprocessing of the seed cotton is performed. Alternatively, producers harvest cotton by machine, and gin the cotton in large production gins. These large gins have ovens for drying the cotton, and leaf and stick machines for removing leaf fragments and other trash from the cotton. After this initial processing, the cotton is ginned to separate the lint and seed, and packaged for shipping.

Prior research has shown the impact of harvest and ginning methods on final fiber properties (Williford *et al.*, 1984; Calhoun *et al.*, 1996). Concern for the differences in harvest and ginning methods on fiber properties has recently surfaced as more small-scale research studies are extrapolating results to the production scale. In precision agriculture, for example, sampling a field for determination of the spatial variability of fiber quality results in samples that are too small for ginning in the larger gins. Spatially-registered samples that are hand- or machine-harvested are ginned on research gins without preprocessing. While these studies give an indication of the extent of spatial variability, the physical features of the cotton fiber from these studies differs in significant ways from that observed by the producers. Because the cotton fiber quality determines the price of the cotton, the value of the cotton varies between these different harvest and ginning method. In the same way, cotton breeders frequently have small test plots of new varieties, and rely on hand-harvesting and research gins for fiber sample preparation.

Current methods of determining the value of the cotton fiber quality and price result in a mathematically nonlinear and discontinuous evaluation relationship (USDA-AMS, 2004). As a consequence, it is difficult to translate the research results to the producer results through standard statistical methods. Moreover, in order to truly examine the spatial variability of the fiber properties, and better determine the complete range of physiological properties, we need to replicate the entire population of samples. We need to find not only the correct mean of the population, but the range of variability as well. With a typical biological system, we anticipate a standard normal (Gaussian) distribution of values when we sample the population. We obtain such a population of values for the various cotton parameters when we sample both the researcher and producer values. Now, the task is to correctly translate that population of cotton fiber properties from the researcher results to the more correct producer values. This study was undertaken to develop a method of translating these small-scale researcher level results to full-scale production level results. The research reported here is the first stage in that study. This study demonstrates the use of Bayesian networks to detect erroneous entries in cotton fiber data sets.

2. METHODS

2.1 Cotton growth and harvest

Cotton was grown in research test plots using standard agricultural practices. The cotton was harvested with a commercial cotton picker (John Deere 699) modified for plot picking. The large cotton samples from the plot picker were subsampled for comparison with the research gin. The remaining cotton was ginned at the USDA-ARS Ginning Lab in Stoneville, MS, using one lint cleaner and one stick machine on the microgin. The small cotton samples subsampled from the large samples were ginned on a 10-saw research gin (Continental Eagle, Co., Memphis, TN). All cotton lint samples were classed at the USDA-AMS Classing Office in Dumas, AR. Cotton discount (or premium) points were determined using the USDA-AMS Spot Cotton Quotations for the South Delta (USDA-AMS, 2004).

2.2 Data presentation and data set development

Cotton was grown and harvested yearly from 2000 – 2003. The spatially registered cotton yield was recorded with the AgLeader Cotton Yield monitor. Data of cotton yield were put into spatially registered maps using ArcView (3.2, ESRI, Redlands, CA). Statistical analysis and plotting of the yield and fiber properties were performed with SigmaPlot (SigmaPlot, 2001, SPSS Inc., Chicago, IL).

Data sets for testing the network were derived from samples ginned in the micro-gin, and consisted of 90 samples from 2001 (data set 2001) and 90 samples from 2002 (data set 2002). Eight feature attributes of the cotton fiber were evaluated. These data constitute the information required to compute the discount subtracted from (or premium added to) the standard value of the cotton sample.

The information provided in the data set varies widely in type and value. For example, while some values are represented as real numbers, other values are discrete numbers. Some values appearing to be integers are actually discontinuous numbers, representing degree of membership in some category, rather than true integers. The data is of high dimension, and calculation of the discount rate is based on look-up tables, rather than a numerical function. An example of the data set is illustrated in Table 1.

In order to create a data set for training the Bayesian Network, artificial erroneous data was created. For each of the 90 entries in each original data set, an erroneous entry was created by randomly selecting an attribute and then changing its original value to a different value between the lower and upper bounds for that attribute such that the new value resulted in a different discount value for the given cotton fiber sample. For example, the lookup table entry for the Uniformity attribute is given in Table 2.

Assuming that the original value of this sample's Uniformity attribute is 81, a change to a new value of 79 or 83 would be acceptable for use in an erroneous data record, since the discount values for both 79 and 83 are different from the discount value for 81. A change to 80 or 82 would not be useful, however, because new Uniformity values of 82 and 80 produce the same discount value as the original Uniformity value of 81.

Using this method of creating artificial data, 90 erroneous data items were produced for each year's data, giving a training set of 90 correct and 90 incorrect items for each year, for a total of 180 items for each of the two years.

2.3 Experimental procedure

A Bayesian Network is a widely-used Machine Learning technique which represents the conditional dependencies of various factors in a statistical relationship by means of a directed acyclic graph. The nodes of the graph represent features of the data space; arcs connecting some (or possibly all) of the nodes are labeled with their conditional probability.

Figure 1 is a representation of a Bayesian Network (adapted from Mitchell, 1997). In this network, the conditional dependencies are indicated by directed arcs (arrows). For example, a forest fire may be dependent upon a campfire or lightning, but it is conditionally independent of thunder.

The conditional dependence probabilities for *campfire* in Figure 1 might look something like that presented in Table 3. In a Bayesian network, a given node may not be connected to some of the other nodes if the effects of these nodes on the overall probability calculated by the network are independent of one another. In these cases, the conditional probabilities do not have to be known or calculated.

In this experiment, a Bayesian Network package created by Christian Borgelt (2004) was utilized with the training data generated as describe above to induce a Bayesian classifier. This classifier was then used to classify the data in the test set.

3. EXPERIMENTAL RESULTS

A range of yield values is apparent from the map of cotton yield (Figure 2). Of particular concern in precision agricultural studies is the extent of variability in crop growth characteristics. To this end, we don't want to simply move, or translate, the research results to the mean for comparison to production level observations. Rather, we want to translate the entire population of fiber quality measurements from one to the other, maintaining a reasonable representation of the variability in the sampled parameters. For a biological system, we expect to have a close approximation to a standard normal distribution of results. We see this normal distribution in the yield data (Figure 3). Deviations in the normal distribution indicate separate populations within the sampled area. In this way we can see that the frequency distribution of yield values gives three separate populations – one low yielding area (red – one standard deviation below the mean), one high yielding area (blue – one standard deviation above the mean), and one intermediate yielding area which contain most of the sampling points (Figure 3). The separate populations from the frequency distribution correspond well with the areas of the map and can be used to separate out the regions of the map based on yield (Figure 4).

The great number of data points in the yield data allows generation of high-density histograms, sufficient to separate out different populations within the overall data set. The limited number of data points of fiber properties may not allow such distinction, but should show a normal distribution. For several of the fiber properties, such distributions are found (Figure 5). The difference between the production and research gin is seen as a shift in the entire population, and an increase in the variability. We want to map that standard distribution of research-level "incorrect" results onto a standard distribution of production-level "correct" results. In this way, we will better be able to explore the population, and map, for example, the spatial variability of the various fiber properties, final fiber quality and value.

The experiment was conducted by using one of the training sets with Borgelt's *bci* software to induce a Bayesian Classifier, and then using Borgelt's *bcx* software to execute the learned classifier on one of the test sets to see how well the classifier performs. Four iterations of this experiment were performed, one for each of the four possible combinations of the two data sets, using each data set alternately as a training set or a test set. The results of the experiment are summarized in Table 4.

4. DISCUSSION

The results are encouraging. The Bayesian Network evidently is able to learn to identify correct from incorrect data quite well when it is trained on and tested on the same year's data.

Although the relationship between the feature values and the cotton discount value is nonlinear and discontinuous, and thus very difficult to learn, the Bayesian Network apparently has deduced a set of statistical rules which allow it to correctly classify the great majority of the data, provided that both the training and testing data come from the same year.

Learning to predict the classification of erroneous data entry is in fact a probabilistic type of learning. The full Bayesian network outperforms the naïve Bayesian network if trained and tested on the same set of data. However, the naïve Bayesian network outperforms the full Bayesian network if trained and tested on different data sets. Full Bayes is better for the current problem, which is simply identifying erroneous data. Naïve Bayesian networks assume that attribute values are conditionally independent, given the classification of the instance. This is not true for the current problem, due to the interaction among the features.

With regard to the greater number of errors when training with one years' data and testing with a different year, it must be observed that cotton lint quality is so dependent on the growing conditions that the quality parameters vary greatly from one year to the next. The particular Bayesian classifier which was learned using one year's data cannot be expected to work very well for correctly assessing a different year, simply because of the differences in the way the cotton grew each year.

FUTURE WORK

Future work will include determining errors in cotton properties due to type of harvesting (machine versus hand) and type of gin (10-saw research [no lint-cleaners or trash removing steps] versus production level gin [additional steps to remove trash and leaves decreases lint strength and length]). If this is successful, it will permit “translating" from results commonly observed by researchers to that commonly observed by producers. In addition, the variability of cotton quality from one year to another is something that will need to be taken into account, if possible.

ACKNOWLEDGEMENTS

The authors wish to thank Mr. Stanley Anthony for generous use of the ginning technology.

REFERENCES

Borgelt, C. http://fuzzy.cs.uni-magdeburg.de/ ~borgelt/software.html#cluster (Current May 23, 2004).

Calhoun, D.S., T.P. Wallace, W.S. Anthony, M.E. Barfield. 1996. Comparison of lint fraction and fiber quality data from hand- vs machine-harvested samples in cotton yield trials. Proceedings of the Beltwide Cotton Conferences. National Cotton Council, Memphis, TN. 1:611-615.

Johnson, R.M., R.G. Downer, J.M. Bradow, P.J. Bauer, and E.J. Sadler. 2002. Variability in cotton fiber yield, fiber quality, and soil properties in a Southeaster coastal plain. Agronomy Journal. 94:1305-1316

Mitchell, T. 1997. Machine Learning, McGraw-Hill Companies, Inc.

Mitchell, T. http://www-2.cs.cmu.edu/afs /cs.cmu.edu/project/theo-3/mlc/hw2 (Current May 23, 2004).

USDA Agricultural Marketing Service. 2004. Market News Report. Cotton Report, Daily Spot Quotations. Available at: http://www.ams.usda.gov/cotton/mncs/index.htm. (Current June 2, 2004).

Williford, J.R., W.R. Meredith, Jr., A.C. Griffin, Jr. 1984. Effect of variety, harvest method, and lint cleaners on cotton quality and value in 1983. Proceedings of the 1984 Beltwide Cotton Production Research Conference. National Cotton Council, Memphis, TN. pp. 114-115.

Figure 1. Example of a Bayesian Network (adapted from Mitchell, 1997)

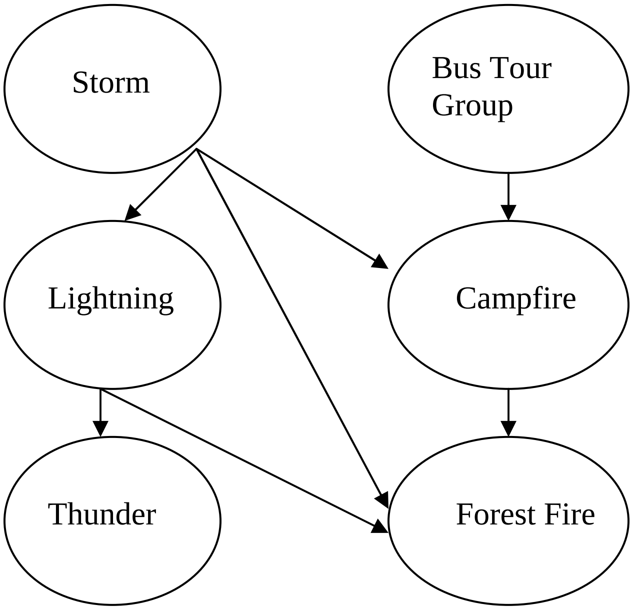


Figure 2. Spatially registered cotton lint yield as determined by a cotton yield monitor.

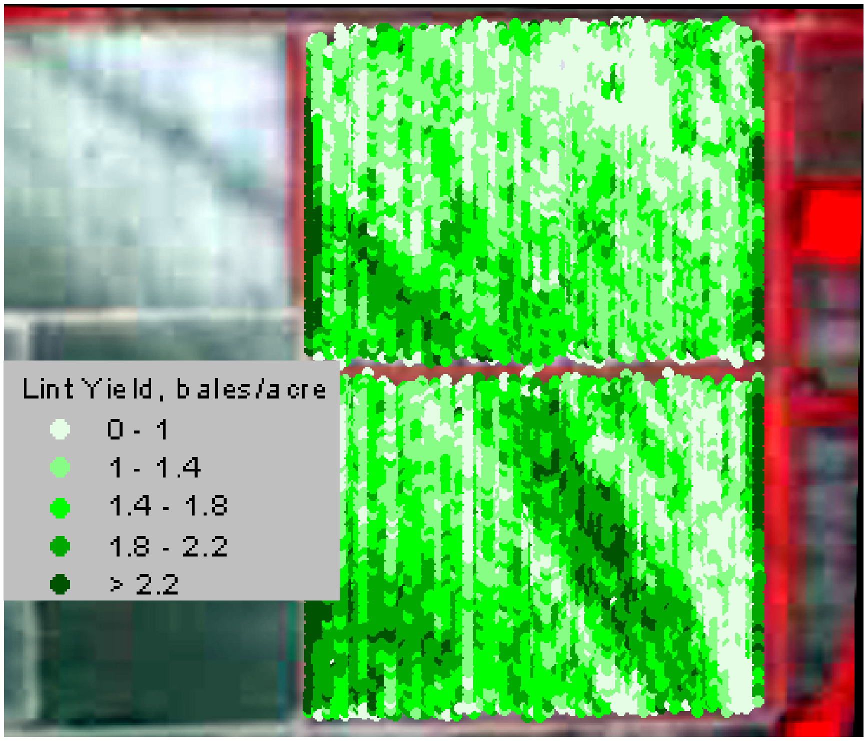


Figure 3. Frequency distribution of cotton lint yield.

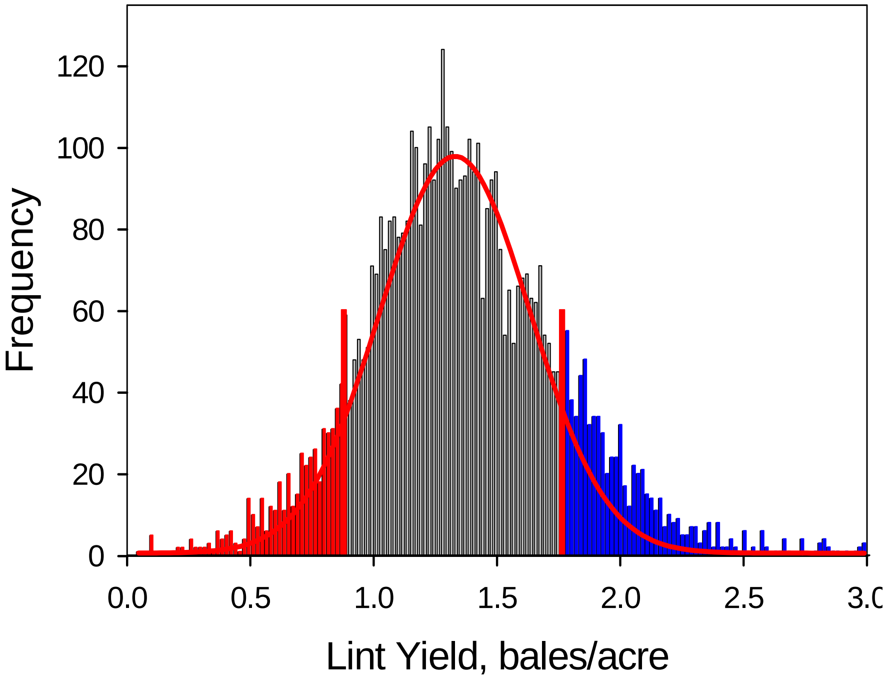


Figure 4. Segregation of populations according to cotton lint yield.

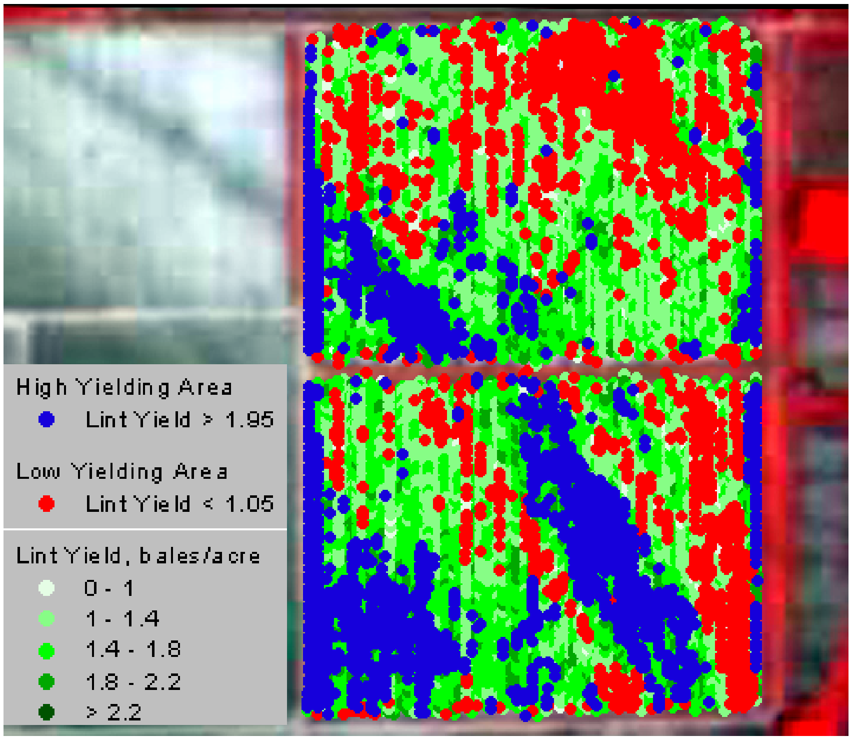
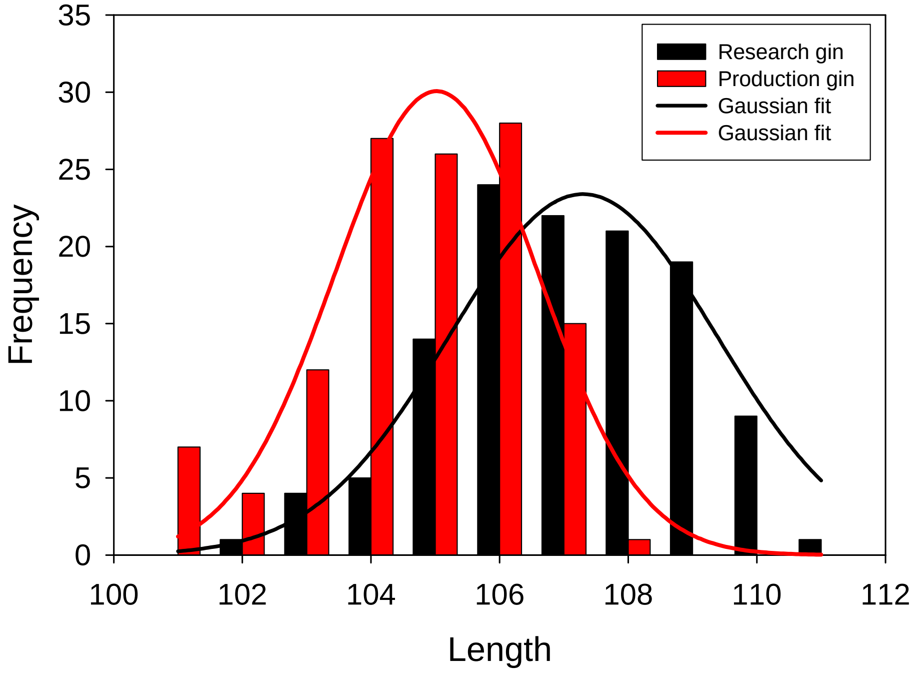


Figure 5. Distribution histograms of length from cotton lint ginned on research and production gins.



Color	Staple	Micronaire	Extraneous	Strength	Leaf	Uniformity	Discount	Correct
42	35	49	0	24.8	3	81	-3.14	1
32	34	48	0	25	3	81	-5.45	1
32	35	49	0	24.8	3	81	-3.14	0
32	40	48	0	25	3	81	-5.45	0

Table 1: Example of Training and Test Data Sets

Uniformity	
Unit	**Discount**
77	-59
78	-49
79	-39
80	0
81	0
82	0
83	23
84	33
85	43
86	53

Table 2: Lookup table entry for the Uniformity attribute (after USDA-AMS Spot Cotton Quotations).

	Storm, BusTourGroup	Storm, ~BusTourGroup	~Storm, BusTourGroup	~Storm, ~BusTourGroup
Campfire	0.1	0.4	0.7	0.2
~Campfire	0.9	0.6	0.3	0.8

Table 3. Example of conditional dependence probabilities for *campfire.*

Training Set	Testing Set	Accuracy Using Full Bayes	Accuracy Using Naïve Bayes
2001	2001	95%	82%
2002	2002	94%	78%
2001	2002	74%	77%
2002	2001	60%	65%

Table 4: Summary of Testing Results with Bayesian Network Classifier