

Cleaning Yield Data

By

Jon Kleinjan, Jiyul Chang, Jim Wilson, Dan Humburg, Gregg Carlson, Dave Clay and Dan Long

Summary

Combine yield monitors are increasingly being used by production agriculturalists across the world. A portion of the data acquired from yield monitors is erroneous. Erroneous data can result from rapid speed changes, extraneous vibration resulting from crossing bumps in the field, not cutting a full header width, erroneous position information and a yield sensor that is not calibrated properly. When using yield monitor data in a decision support system, any data that can be identified as erroneous data must be removed. This guideline discusses sources of problem data and methods of yield monitor file data cleaning.

Introduction

Yield monitor data normally contains some values that are incorrect. For example, when a combine crosses an area of a field that has already been combined with the header down in harvesting position; the yield monitor continues to record yield (or really no yield). This second recording of harvest data at zero yield is clearly in error. Another example includes the case when a field is almost completely harvested and there are harvest swaths that are not the full header width. Blackmore and Moore (1999) demonstrated an approach to remove these narrow finishes. Cleaning the data removes problem values and can improve our ability to explain yield variability. Thylen and Algerbo (2000) reported that a filter used to clean data reduced the standard deviation of the yield data by 25%. In this approach, 10 nearest neighbors excluding data from the same transect are compared to the measured value. The mean yield of the 10 nearest neighbors is calculated. If the measured value falls outside of the bounds of the nearest neighbor mean plus or minus a threshold of acceptance, the measured value is excluded from the data set. From a visual perspective, removing erroneous data may have little impact upon the appearance of yield maps (as is the case in figure 1 and figure 2). However, from an analytical perspective, removing erroneous data may have a significant impact upon our ability to compare the data to other information layers within the decision support system. For example, a remotely sensed near infrared (NIR) image was taken of a 160 acre South Dakota field. The correlation coefficient (our ability to predict corn yield from NIR imagery) between yield and NIR reflectance was 0.42 ($p > 0.05$) for an uncleaned data set (a correlation coefficient of 1 would indicate perfect ability to predict yield from a NIR image). Removing erroneous data, or cleaning, improved the correlation coefficient to 0.61 ($P > 0.05$). Clearly this is a statistically significant improvement. Cleaning data is necessary within a decision support system.

**Moody 1999
Corn yield monitoring data
(Un-cleaned up)**

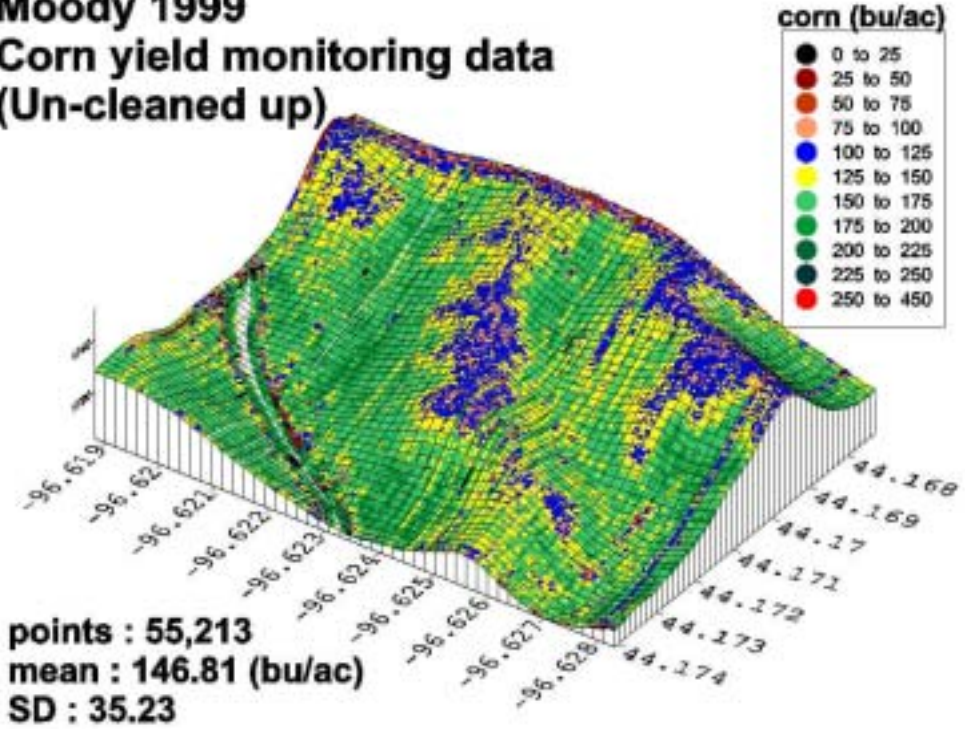


Fig. 1. Unclean corn yield monitored map, superimposed on topographic map.

**Moody 1999
Corn yield monitoring data
(Cleaned up)**

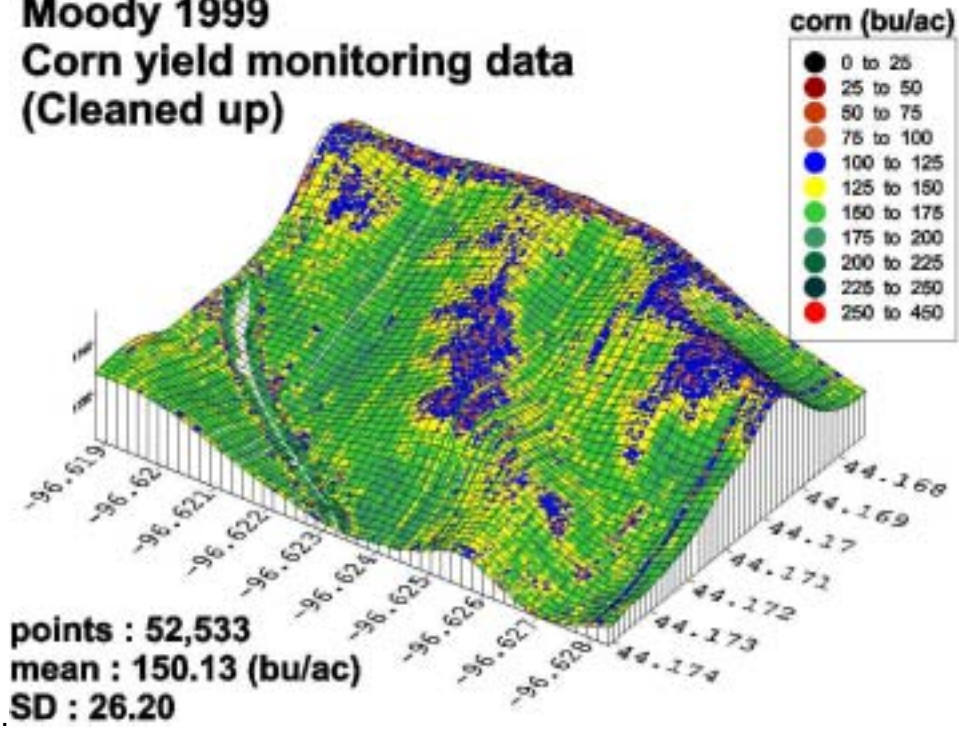


Fig. 2. The same, but cleaned corn yield monitored map superimposed on a topographic map.

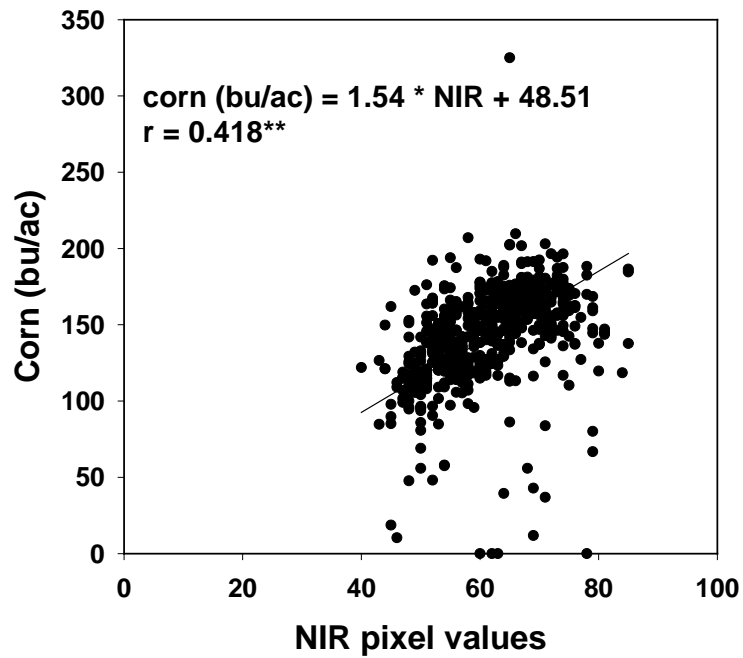


Fig. 3. Regression between NIR band image pixel values taken in Sep.21, 1999 and resampled (581 points) corn yields from un-cleaned yield monitoring data.

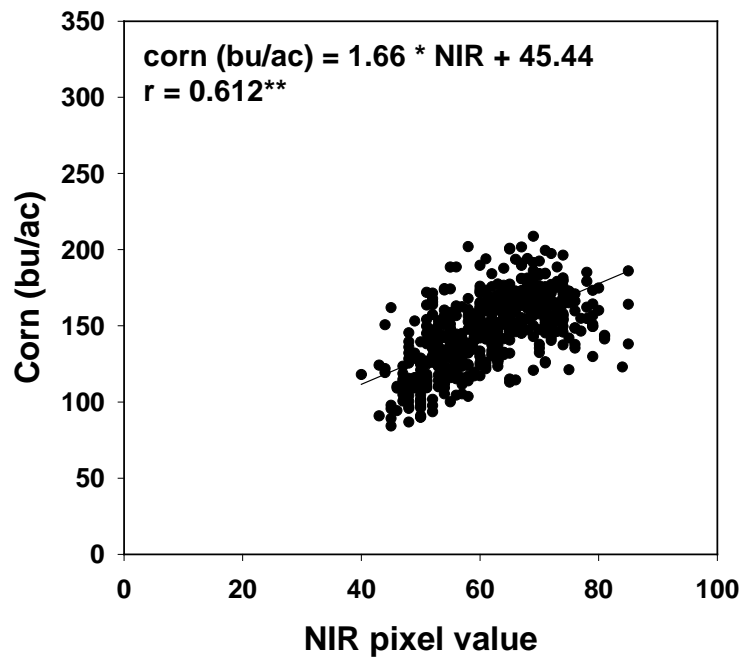


Fig. 4. Regression between NIR band image pixel values taken in Sep.21, 1999 and resampled (581 points) corn yields from cleaned yield monitoring data.

Cleaning

To begin cleaning data, one must export yield data out of the yield monitor system. Each yield monitor system has a different exported file format. We will address here yield monitor data generated by John Deere and Ag Leader monitors.

John Deere. John Deere exports yield data as a comma delimited ASCII file (often denoted as CSV or comma separated variable). Each record or measurement is separated by a carriage return, line feed and so occupies a single line in the file. The complete file format is found in appendix D of the Green Star user manual. The first eight columns of the CSV file are respectively; Longitude (decimal degrees), Latitude (decimal degrees), Flow (lb/sec or dekagrams/sec), Time (GPS time in seconds), Cycles (number of seconds since last reading), Distance traveled since last measurement (inches or millimeters), Swath (inches or millimeters), and moisture (%). John Deere cautions that when exporting data of a field that has had calibration accomplished within that field, the user should use **(File, Preferences, Export, Export yield – moisture with post calibration values)**. Flow and moisture of the entire field will then be exported with the latest calibration data. It should be also noted that the exported data already has already been adjusted for the time delay from the header to the yield and moisture sensors.

Ag Leader. Ag Leader uses two CSV file formats, Basic and Advanced. See Appendix A of the Ag Leader user manual to find specific details. Each record or measurement is separated by a carriage return, line feed and so occupies a single line in the file. The first four components of the basic file format are Longitude (decimal degrees), Latitude (decimal degrees), yield (bu/acre corrected to standard moisture content), and moisture (%). The first eight columns of the advanced CSV file are respectively; Longitude (decimal degrees), Latitude (decimal degrees), Flow (lb/sec), GPS Time (seconds), Cycles (number of seconds since last reading), Distance traveled since last reading (inches), Swath (inches), and moisture (%). Both basic and advanced data sets already have the time delay from the header to the yield and moisture sensors accounted for. Both basic and advanced data have the most current calibration information already incorporated into the yield estimate. Filtering is accomplished in the basic data output but not in the advanced output. (this is why yields calculated by each system are not the same) Ag Leader also has SMS software that will export data to a TXT (text) or CSV file format. This software can be down loaded from Ag Leader site <http://www.agleader.com/sms-download.htm> . When the data file is read by the software, the data is filtered and the time delay is applied.

Any harvest yield monitor system is subject to some unavoidable design limitations. For example, there is a somewhat indeterminate time lag from the time when a kernel of grain is removed from the field by the harvest system's header until when that kernel of grain is accounted for by the yield monitor. The reason for this time delay can be illustrated with an example. Assume that we dye an ear of corn florescent green. We throw the florescent green ear into the header of a combine as it moves through a field. We then watch for the florescent green kernels to come out the clean grain auger into the grain holding tank (which in most cases happens to be quite close to where the load cell of the yield monitoring system and the moisture sensor are located). Typically, we see, depending upon the specific combine being investigated, the first florescent green kernel coming out the clean grain auger about 7-8 seconds after the ear was thrown into the header. The highest concentration of florescent green kernels may come out of the clean grain auger at 12-13 seconds, and the final florescent green kernel at 21-22 seconds. Since the yield monitor is sending a yield measurement to our yield archiving computer data storage system every second, the question that must be answered is, when

was a measurement of our florescent green ear of corn made? Obviously the florescent green ear entered the header at a single point in the field. It entered into the system instantaneously. However the florescent green ear's kernels were measured by the combine yield monitoring system over a period of time which makes it almost impossible to locate the exact point of harvest. This indeterminate time of flow happens because different kernels of corn can take different paths from the header to the combine grain hopper. Since a yield monitoring system actually weighs the mass of grain flowing past a particular point in the combine's flow path, pinpointing the transit delay and hence, the point of origin is difficult. Yield monitors will either use a constant delay time to represent the average transport time through a machine or will allow the operator to input an estimate of the delay time.

Yield is determined by the three equations listed below. The mass from the flow sensor, width of header, velocity of travel, and the moisture content from the moisture sensor are used to calculate corn and soybean yield using the following equations.

$$\begin{aligned} \text{wt of corn} \\ \text{at MC \%} \\ \text{moisture} \\ \text{(lb/bu)} \end{aligned} = \frac{\frac{\text{MC} * 47.32}{100}}{1 - \frac{\text{MC}}{100}} + 47.32 \quad (1)$$

$$\begin{aligned} \text{wt of soybeans} \\ \text{at MC \%} \\ \text{moisture} \\ \text{(lb/bu)} \end{aligned} = \frac{\frac{\text{MC} * 52.2}{100}}{1 - \frac{\text{MC}}{100}} + 52.2 \quad (2)$$

$$\begin{aligned} \text{yield corn} \\ \text{(bu/acre)} \end{aligned} = \frac{\text{mass flow} \\ \text{(lb/sec)}}{\begin{aligned} \text{wt of corn} \\ \text{at MC \%} \\ \text{moisture} \\ \text{(lb/bu)} \end{aligned}} * \frac{1}{\begin{aligned} \text{header width} \\ \text{(inches)} \end{aligned} * \begin{aligned} \text{velocity} \\ \text{(inch/sec)} \end{aligned}} * \frac{144 \text{ inch}^2}{\text{ft}^2} * \frac{43560 \text{ ft}^2}{\text{acre}} \quad (3)$$

The actual cleaning of the data is a multi level process. Each phase of the process is listed below. Because of the complexity of actually doing the cleaning, an Excel spread sheet with Visual Basic macros is available at website (<http://www.>) to clean CSV exported files. Because we wish to maintain the integrity of the original data set, our method is to read the original and write a new corrected file.

Phase I (the combine's header is up). We begin with errors that are easy to identify. If the header is up and data are still being recorded, these data are of little value as no crop is entering the machine. In Ag Leaders advanced export file format, these data are contained in the 9th column and are coded with an integer 1 when the header is down. A line of data that indicates that the header is up is just not written to the new (corrected) file.

Phase II cleaning (the speed is rapidly changing). Recall the discussion of the fluorescent green kernels from above. If we were to repeat this test under different situations, we would find that the time that it takes for kernels to pass the mass flow sensor changes from a light flow of grain to a heavy flow of grain. This creates uncertainty determining when the yield monitor is

actually reading an amount of grain cut by the header. Because of this, the probability of a yield monitor error becomes significantly greater when the speed of travel changes. When there is a velocity change of *GREATER THAN* 15% (in the cleaning routine this number is a variable and has been changed many times. The decision to use 15% is subjective) of initial velocity from one reading to the next, the data is excluded.

If V_i is $> (V_{(i-1)} + .15*V_{(i-1)})$ or If V_i is $< (V_{(i-1)} - .15*V_{(i-1)})$ then

The line of data containing V_i is not written onto the new file

Where

$V_{(i)}$ = the velocity at present time

$V_{(i-1)}$ = the velocity at the previous time step

Phase III cleaning (the speed is slow). A yield monitor is designed to operate at normal harvest speed. When the velocity of the combine approaches or equals 0 mph the area harvested by the header approaches 0. Since a combine can sit still but have grain flowing in the clean gain auger, any mass measured by the yield monitor results in mass divided by 0 area which results in infinite yield. (note that some yield monitor systems do not record data if the velocity is zero) Since normal calibration of most modern combines and their yield monitors occurs at velocities of 3 to 5 mph and since data recorded at velocities below 1 mph is clearly taken at other than normal harvest velocity and thus is suspect, we choose to throw out any data taken at speeds of less than 1 mph.

If V_i is < 1 mph then

The line of data containing V_i is not written onto the new file

Phase IV cleaning (the flow of grain past the yield monitor is low). Yield monitors are designed and calibrated to measure grain flow. When data are taken in a range for which no calibration data has been collected, the potential for error is greater. When calibrating the yield monitor, we try to get both high and low flow rates. It should that high or low flow rate do not necessarily correspond to high or low yield. For example, one can harvest 200 bu per acre at 1 mph and get a moderate flow rate or harvest 100 bu per acre yield at 8 mph and get a very high flow rate. Calibration of the yield monitor is actually made to flow rate, not to yield. If flow rates are outside the range that was calibrated for, the collected flow rate data is less credible. If the flow rate is outside of the range of $limit_{lower}$ to $limit_{upper}$, our acceptable flow limits, consider not using the data.

If $Flow_i < limit_{lower}$ or $flow_i > limit_{upper}$ then

The line of data containing $Flow_i$ is not written onto the new file

Where

$Flow_i$ is the flow (usually lb/sec) registered by the yield monitor at time i

Phase V cleaning (yield exceeds +- 3 standard deviations). After applying the first four criteria to our yield data, it is still possible to have anomalous yield results in our yield monitor file. The fifth stage of our cleaning process involves calculating the Average and Standard Deviation of the yield data as it remains within a particular distance constraint. Within a block that is X_1 ft by X_1 ft, (we have been using $X_1 =$ three header widths, for a combine with a 30 ft header, $X_1 = 90$) we make the assumption that there should be relative homogeneous data. If the data in this area are greater than X_2 (we have been using $X_2 = 3$) standard deviations from the mean of the block, this data becomes suspect and is thrown

If $Flow_i > (AVE_{bk} + X_2*SD_{bk})$ or $Flow_i < (Ave_{bk} - X_2*SD_{bk})$ then

The line of data containing $Flow_i$ is not written onto the new file

Where

AVE_{bk} = the average of flow within the block bk

SD_{bk} = the standard deviation within block bk

Summary

Yield monitoring can provide valuable information regarding the extent and location of variability in a field. Software and decision support systems are being developed to assist producers in managing variability, and maximizing value across varying landscapes. Utilizing those decision support systems requires understanding of yield variability and its causes. Erroneous data and associated artificial variability that results from erroneous data can confound the analysis and result in poor decisions. The steps outlined in this guideline will allow a producer to remove many of the data points within a harvest yield set that are potentially problematic and improve the quality of the subsequent analysis and decision making process.

References

Blackmore, S. and M. Moore. 1999. Remedial Correction of Yield Map Data. Precision Agriculture, 1, 53-66

Thylen, L. and P. A. Algerbo. 2000. An Expert Filter Removing Erroneous Yield Data. Proceedings of the Fifth International Conference on Precision Agriculture. American Society of Agronomy.

Appendix

1 mile/hour = 1.466ft/sec

Several moisture correction examples.

10,000 pounds of corn is harvested at 17.5% moisture.

Equation 1 is used to determine the constant to use to correct a weight of grain at a measured moisture content to bushels of grain at a standard weight and moisture content (for corn, 56 lb/bu at 15.5% moisture).

From equation 1, 1 bu 17.5% moisture corn corrected to 15.5% moisture corn = 57.3575 pounds corn/bu

So 10,000 pounds of 17.5% corn = 174.34 bu

Unfortunately, if grain is very dry, (less than 15.5% moisture) the bushels of grain calculated by this equation may not be the same as the bushels of grain that is paid for by the market. As an example, if 10,000 pounds was harvested, at 12% moisture. The corn is delivered to an elevator, the equation below indicates

From equation 1, 1 bu of 12% moisture corn corrected to 15.5% moisture corn = 53.773 pounds/bu.

So 10000 pounds of 12% corn = 186 bu.

However, the elevator may calculate the bushels of corn based on 56 lb/bu ignoring the actual moisture content. They may conclude that you have delivered 178.6 bu of corn.