# The Basics of Cleaning Yield Monitor Data

*This guide was produced as part of a project jointly funded by the North Central Soybean Research Program and the United Soybean Board. Research Team: William Wiebold and Harlan Palm (University of Missouri); Ken Sudduth and Newell Kitchen (USDA/ARS/University of Missouri); Bill Batchelor (Iowa State University); Kurt Thelen (Michigan State University); David Clay (South Dakota State University) Don Bullock and German Bollero (University of Illinois); Ron Schuler (University of Wisconsin).*

## Introduction

Yield monitors and the ability to map yields have developed significantly during the last decade and have played a large part in the development and adoption of precision farming. While farmers have always known that fields are variable, the yield map provides the ability to quantify and "remember" those variations (Figure 1). Over time, with good record-keeping, these maps can provide a roadmap for directed scouting, analysis and possibly treatment of under-performing areas. They also provide a means to evaluate successes (and failures) in management decisions and practices.

However, yield maps are only as good as the data collection and processing techniques used to create them. And while yield monitoring systems and software have become more user-friendly over the last decade, collection and processing of sensor data are by no means completely automated. It is generally understood that proper installation, calibration and testing of the yield monitoring system are crucial in the data collection process. Yet most people do not

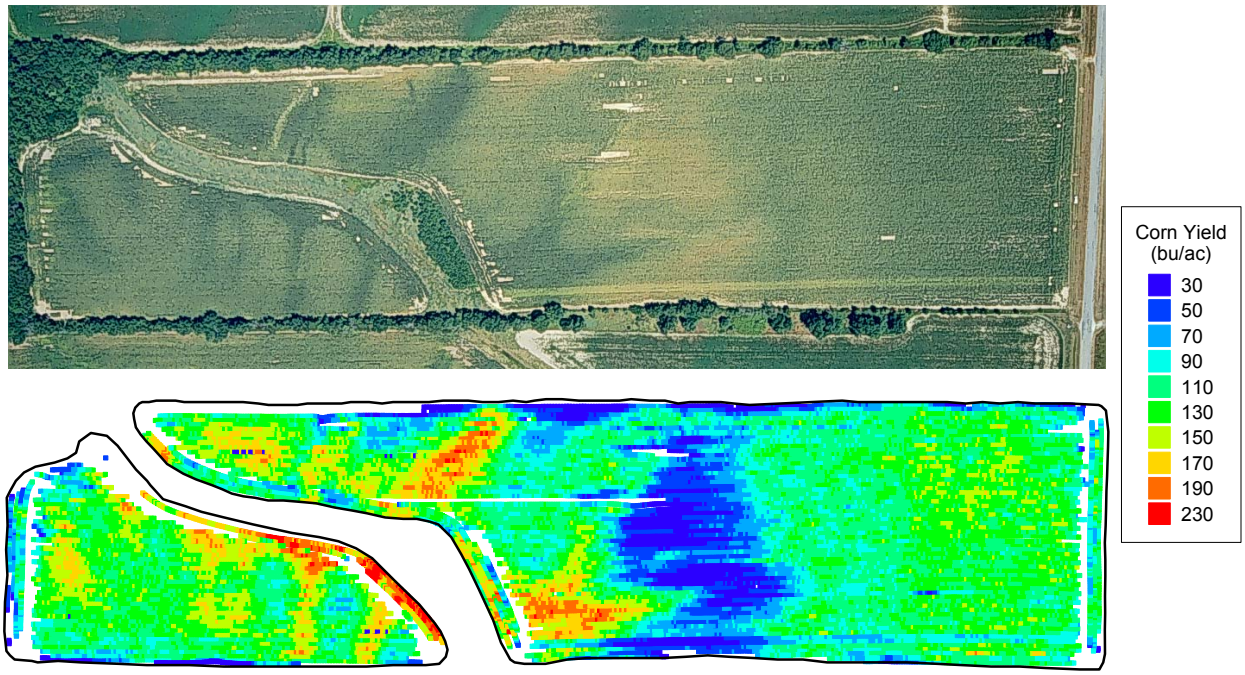


**Figure 1. Yield monitors provide a way to quantify the yield variations producers know exist**

realize the necessity of cleaning the processed yield data in preparation for mapping and/or any further analysis. This guide will try to address the most critical aspects of yield data cleaning.

**Glossary:**

Yield monitoring system - A system consisting of AT LEAST grain flow and speed sensors, along with a data acquisition system to collect and store the data from these sensors. It usually includes a global positioning system (GPS) for geo-referencing the data.

Yield monitor calibration - The process of collecting data consisting of known test samples (i.e. known distance, known weight of grain, etc.) for each sensor in the system, and using these known samples to improve the accuracy of the sensor estimates.

Yield processing - Procedures that merge, align and calculate instantaneous (point) yields from yield monitoring system sensor data.

Yield map - A graphic showing a set of geo-referenced instantaneous point yields, usually with graduated colors or symbols to represent different yield levels.

**What is yield data cleaning?**

Yield data cleaning consists of a variety of procedures performed during and/or after yield processing that optimize yield accuracy and remove problem points from the yield data. This sounds complicated, and it can encompass a wide variety of data issues that might require a book to describe them all. However, the vast majority of problems encountered in the average yield map are due to a few common causes. In this guide, we will investigate the most common, and the most critical aspects of cleaning yield data.

**Why do we need to clean yield data?**

Figure 2 helps demonstrate why we should clean yield data. This 330 foot square (approximately 2.5 acres) of corn yield data was clipped from the within a 160 acre field. There appears to be good yield variation here, and the yield estimates in adjacent transects (which were harvested in opposite directions) seem quite reasonable, indicating that the yield processing was done correctly. There are really no significant problems with this yield map, save one. Notice the circled area in the lower left portion of the map. It appears from the tight spacing of the yield points in this region that the combine had to stop for some reason. When this occurred, one point with a high yield (in red) was recorded. What really happened here is that the combine stopped rather abruptly, while grain flow to the sensor tailed off slowly

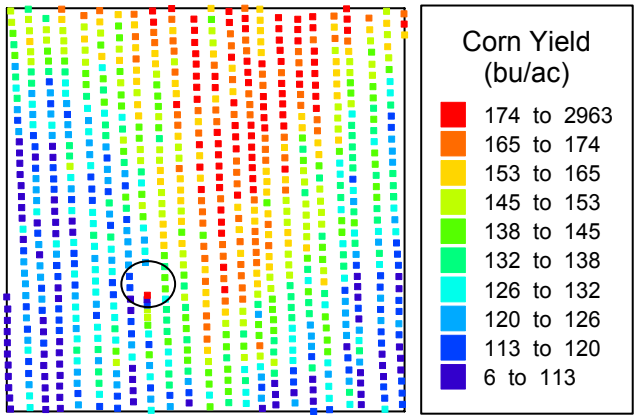


**Figure 2. Example of an error caused by an abrupt change in combine velocity.**

over a period of time. A reasonably low grain flow was recorded at this point (1.27 lb/s) but the combine moved a distance of only 1 cm (.4 in) during this interval. Performing the yield calculation, a yield estimate of 2963 bu/ac was received for this point. Visually, the effect of this single point is minor. However, calculating an average yield for this 2.5 acre region, including this point gives a value of 143.9 bu/ac. Without this single point, a value of 140.6 bu/ac is obtained, for a difference of 3.3 bu/ac. Imagine the possible effects should several dozen such points exist in a yield map.

How necessary is it to find and remove errors in yield data? That depends on how the yield data are to be used. If a yield map is to be used solely as a visual tool to assist an individual in recognizing areas of a field that need attention, then very little yield cleaning may be necessary. Should the user be interested in numerical calculations, such as field or field-region yield averages, then more attention to detail is required. Should the user be interested in on-farm research, comparing data layers of yield to other data layers (i.e. aerial imagery, soil nutrients, elevation , slope, etc.), the greatest of care should be taken to remove all data that might lead to erroneous conclusions.

**What needs to be cleaned?**

The first step in learning how to clean yield data is developing an understanding of the various problems that commonly occur. As discussed in the previous section, abrupt changes in combine velocity can cause significant problems for numerical analyses of yield data, yet are very difficult to locate visually. Other error types can be much easier to detect.

For example, one of the most common errors seen in yield maps is the selection of an incorrect combine delay time. This value represents the time it takes for grain entering the header to reach the grain flow sensor. If the grain took 12 seconds to reach the grain flow sensor, we would like to place the position of that sensed grain at the location occupied by the combine 12 seconds earlier. This delay time is usually selected and applied inside the yield processing software, and can vary significantly (especially across crops, grain moisture percentages, combines and combine settings). The selection of an incorrect delay time can be quite obvious (Figure 3). On the left map, the yield data was processed with a delay time of 9 seconds, while the right map shows the results of the data processed with a 14 second delay time. The distinct

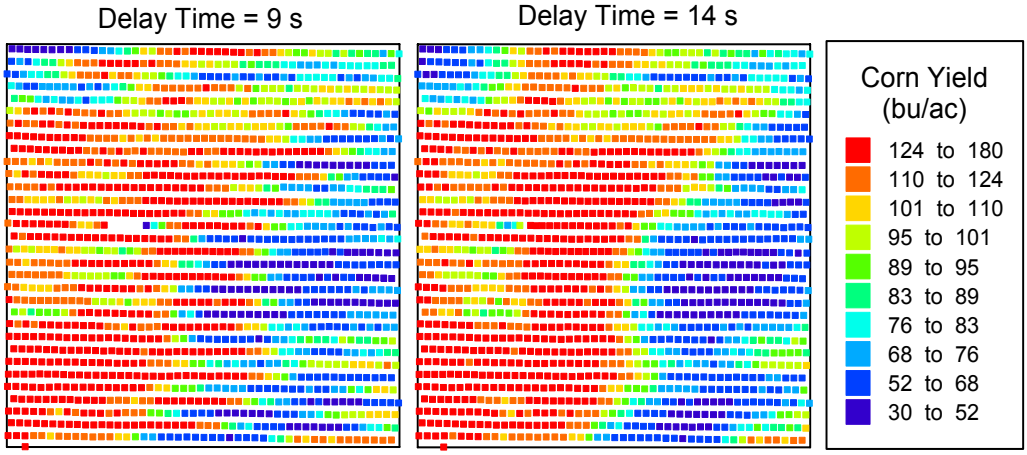


**Figure 3. Example of the visual problems created by an incorrect delay time.**

sawtooth pattern demonstrated by the left map is a clear indication that the delay time has not been set properly, and that the resulting yield data are suspect.

Other errors that occur commonly and are visually easy to detect include low yielding strips due to swaths of less than full header width and positional errors (i.e. GPS wander). Figure 4 shows two more yield maps of a 325 foot square size to demonstrate. The two ellipses in the left map highlight two narrow partial swaths that remained after the soybean field had been harvested in lands. Note that both are very low yielding, and that both pinch out in the middle of the field (to the left). Even if great care has been taken by the operator to accurately estimate swath width on the yield monitor, it is generally better to remove at least the very narrow regions of these partial swaths from yield maps. The right figure shows several transects, all of which should be

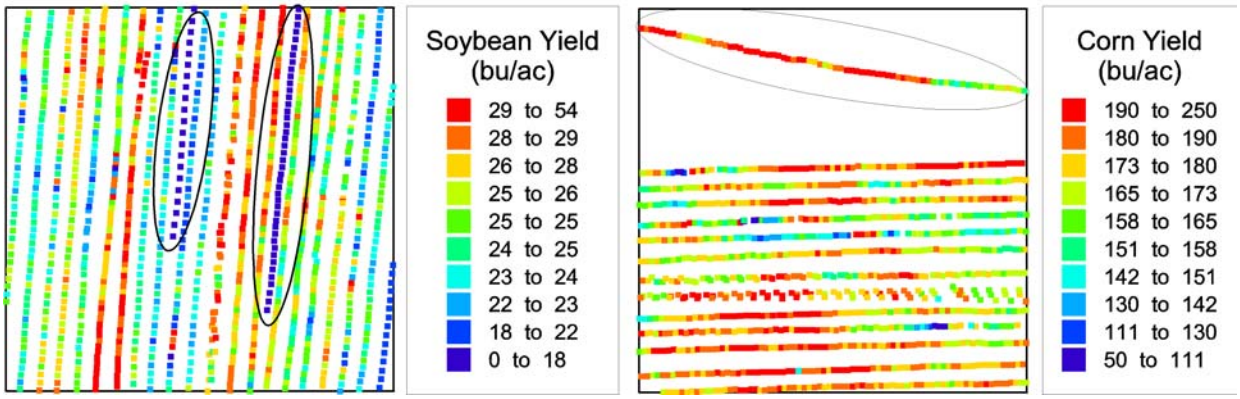


**Figure 4. Yields maps illustrating examples of inaccurate swath width (left) and loss of GPS differential signal.**

evenly spaced and parallel to one another. An ellipse is drawn around the single transect where the GPS differential correction signal has been lost, and positional accuracy has been significantly degraded. Again, unless means exist to reposition this data fairly accurately, the best course of action is generally to delete the suspect data. As a general rule, it is better to remove any point or transect where serious questions about the quality of the yield or position data exist, assuming the remaining yield data are relatively dense.

Another major problem that can be experienced, and one that is often overlooked, is the ramping up/down of flow rates at the edge of the crop. As we've already discussed, the delay time for a given yield data set is relatively constant, given that crop and machinery conditions do not change significantly. However, that delay time is for a combine operating at a reasonable capacity in the middle of the field. When first entering the crop, a combine may require a considerably longer time period before a steady stream of grain is achieved at the flow sensor (due to storage capacity in the combine threshing system). When a combine exits the crop, the grain flow over time may be quite different than on entry, and some grain may continue to flow long after the combine delay time has been reached. The net effect of these different ramping rates is to make the yield map appear very "ragged" at the edge of the crop (i.e. Figure 5, left). Some people adjust the combine time delay to try to minimize this effect. However, as we have already seen, an incorrect combine delay time can strongly affect the regions of the yield map that do not lie near the crop edge (Figure 3). Generally, the best solution is to trim back the data near the crop edge (at least) until the differences between transects entering the crop and those leaving the crop are not obvious (Figure 5, right). The data in these regions have already been

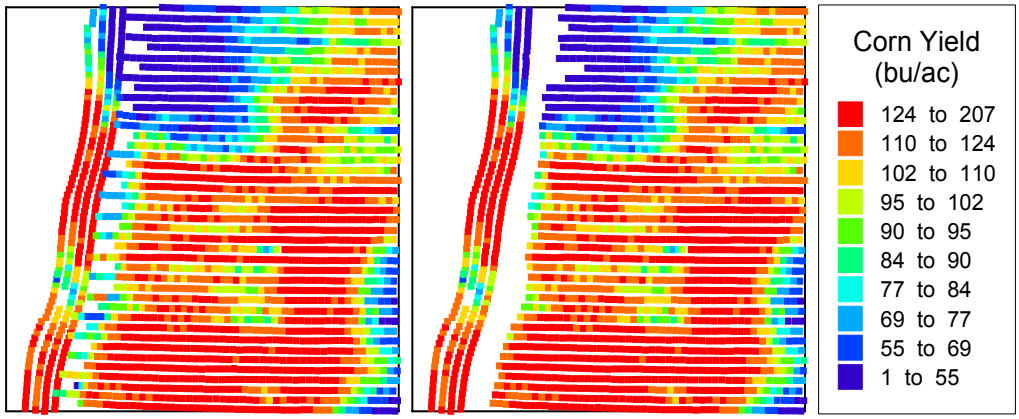


**Figure 5. Example of the effects of combine ramping at the crop edge.**

somewhat compromised by the ramping of the combine, so it is better to remove some of these data points than to affect those in the rest of the field.

**How do I clean the yield data?**

Many of the errors discussed above can be addressed inside the yield processing software provided by yield monitoring system manufacturers. For example, combine delay time, and additional delays for entering and leaving the crop can be set by the yield processing software before the yield data is calculated. However, the process for determining the optimal values for the various settings generally requires visual comparison of several different maps. This can require a significant amount of time and effort if the resulting maps cannot be quickly updated and accurately displayed inside the software.

Furthermore, some errors simply cannot be addressed by traditional yield processing software, and must be edited out through visual selection and removal. Traditionally, geographic information system (GIS) packages (i.e. ArcView, SSToolbox, etc.) have been used to perform these (and many other) types of tasks. While a good deal of time must be invested in order to learn to use one of these software packages, they provide a very flexible tool that can assist with many different types of analysis tasks.

A simpler alternative for yield processing and cleaning is currently under development. The software is available through the Missouri Precision Ag Center web site at (http://www.fse.missouri.edu/mpac/). The software will accept data exported from commercial yield monitoring software systems and allow a variety of numerical filters to be set. The yield data are then processed using those settings, the yield map is updated, and the user is able to see what points have been removed by each filter. Most importantly, it allows the user to visually interact with the map, selecting points, transects or whole areas of data to edit, remove or export.

Also available is a data cleaning program through a web site at South Dakota State University (http://plantsci.sdstate.edu/precisionfarm/Publications.htm). This software is an Excel worksheet with embedded macros as the program. It can be used to remove erroneous yield monitor points from the yield monitor files.

**Points of caution**

- Proper calibration of the yield monitoring system is critical to provide accurate yield data.
- Incorrect combine delay times will affect visual analysis of the yield map, possibly very significantly.
- Abrupt velocity changes and ramping at the crop edge are the factors most likely to affect numerical analysis of the yield data.
- The list of possible yield errors covered in this guide should in no way be considered exhaustive. Many more exist (e.g. moisture sensor issues, corrections, delay times, etc.), but the most common error types seen in yield maps have been considered.