Institute of Agriculture and Natural Resources
EC2005

Improving Yield Map Quality by Reducing Errors through Yield Data File Post-Processing

Joe D. Luck, Extension Specialist, Precision Agriculture Engineer
Nathan Mueller, Extension Educator
John P. Fulton, Associate Professor, The Ohio State University

Introduction

Yield monitor data is certainly one of the most valuable pieces of information that is gathered throughout the year. It can allow producers to estimate profitability, evaluate management decisions, and develop recommendations for the upcoming year. If this information is to be used to its fullest potential, ensuring that the yield data represents accurate estimates of crop performance is critical. However, yield monitor data typically contains some errors. While errors are generally a very small percentage of the data gathered, they can influence the final results.

Common physically-measured errors include:

1. Header cut-width (or harvest width)
2. Header position
3. Lag time (or flow delay) settings
4. Travel distance measurements

Soon after yield monitoring systems became commercially available, researchers quickly began to develop methods to improve the quality of those datasets. Different procedures (some real-time and others post-harvest) were developed as early as the late 1990s to solve many of these issues.

The goal of this publication is to help end users understand why post-processing or "cleaning" yield data may be important for their operations by showing examples of common errors and providing suggested best management practices (BMPs) for reducing them within their datasets.

A list of abbreviations used in this article are: Best management practices (BMPs), file format for files using comma-separated values (.csv), farm management information systems (FMIS), global position systems (GPS), inverse distance weighted (IDW), kriging (KRG), prescription (Rx), file format for files using text (.txt), Spatial Management Software (SMS), and United States Department of Agriculture (USDA).

Why Post-Processing (or Cleaning) Yield Data is Important

At the time yield data collection became mainstream, many farm management information systems (FMIS), including software comparable to Ag Leader's SMS or John Deere's Apex, were difficult to use for most customers. Most users would generate yield maps for viewing; however further analysis using the data was not widespread. Many prescription (Rx) maps were generated by manually creating zones that didn't require a high level of accuracy when viewing yield maps. General trends across a field were considered when generating these management zones on such maps.

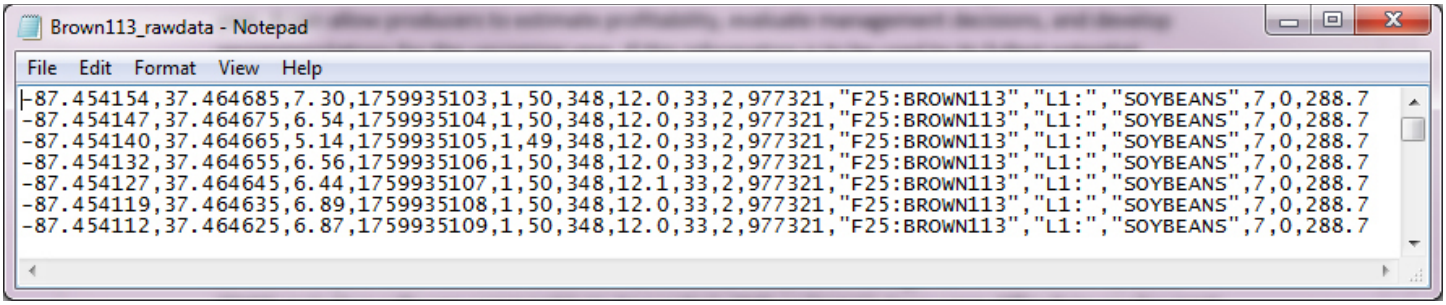

Figure 1: Example of comma separated data (in Ag Leader Advanced .txt format) used to estimate yield and create the yield map.

Since that time, some have continued their efforts to create tools for post-processing yield errors (i.e., cleaning) from datasets. Recall that each yield data point consists of a series of values determined by either sensors on the combine or by manual operator entry. *Figure 1* shows a text (.txt) file format example of the data (separated by commas) used to calculate yield and create the yield map.

Three pieces of information are all that is necessary to create the traditional yield map (*Figure 2*): GPS coordinates (longitude and latitude) and the yield estimate are used to create the yield map which is generally displayed as points when viewed by the user. Knowing that the crop harvested was soybeans in a dryland field, it should be evident that yield values up to 181 bu/ac are not realistic and some erroneous data are included in this dataset shown in *Figure 2*. Harvest or cut width errors resulting in low yield estimates are also noticeable along the southeast with some lag time setting errors as the combine exits the headland in the cut crop edge (shown in red in *Figure 2*). Some yield data points with excessively high yields are also highlighted in *Figure 2* in white.

More recently, automated data analysis has become popular in FMIS software packages that use grid- or contour-based yield maps for delineating management zones, creating Rx maps or quantifying yield in general. The process of creating the grid or contour yield maps is known as interpolation and essentially converts the point yield data into a continuous surface of estimated yield values. The two most commonly used forms of interpolation are Kriging (KRG) and Inverse Distance Weighted (IDW); both options are offered in most FMIS software. In general, due to the density of yield data points, either KRG or IDW can be used to effectively create a grid or contour yield map.

Yield Error Impacts on Zone Yield Values

Header Cut-Width Errors – In one sense, both KRG and IDW interpolation methods mask many of the errors that have been discussed due to the averaging between points when creating the grid or contour yield map. Consider the yield data shown in *Figure 2* when the IDW method is applied to create a gridded yield map as shown in *Figure 3 (A)*. The raw data from *Figure 2* indicates yield data point values ranging from 6.0 to 181.0 bu/ac whereas with IDW gridded yield estimates (50 ft square grids), range from 31.2 to 61 bu/ac. So it appears that the quality of the yield data has been improved, however, in reality the invalid data points shown in *Figure 2* have been averaged in with legitimate yield information to create the map in *Figure 3 (A)*. *Figure 3 (B)* illustrates a contour based yield map created from the raw data shown in *Figure 2*.

Today, FMIS software can automatically import yield data and within a few steps, create either grid- or contour-based yield maps like the ones shown in *Figure 3*. While it's not readily clear that errors exist in the *Figure 3* maps, if yield data cleaning is performed and the two maps are compared, errors become more obvious. The *Figure 2* raw data were cleaned (post-processed) using Yield Editor software, and grid- and contour-based yield maps were created using the same methods as for *Figure 3*. The resulting cleaned yield maps are shown in *Figure 4*. Comparing the *Figure 3* and *Figure 4* yield maps, it is obvious where many of the errors occurred along the east and west end rows (lag time settings) and along the southeast point rows (cut-width errors).

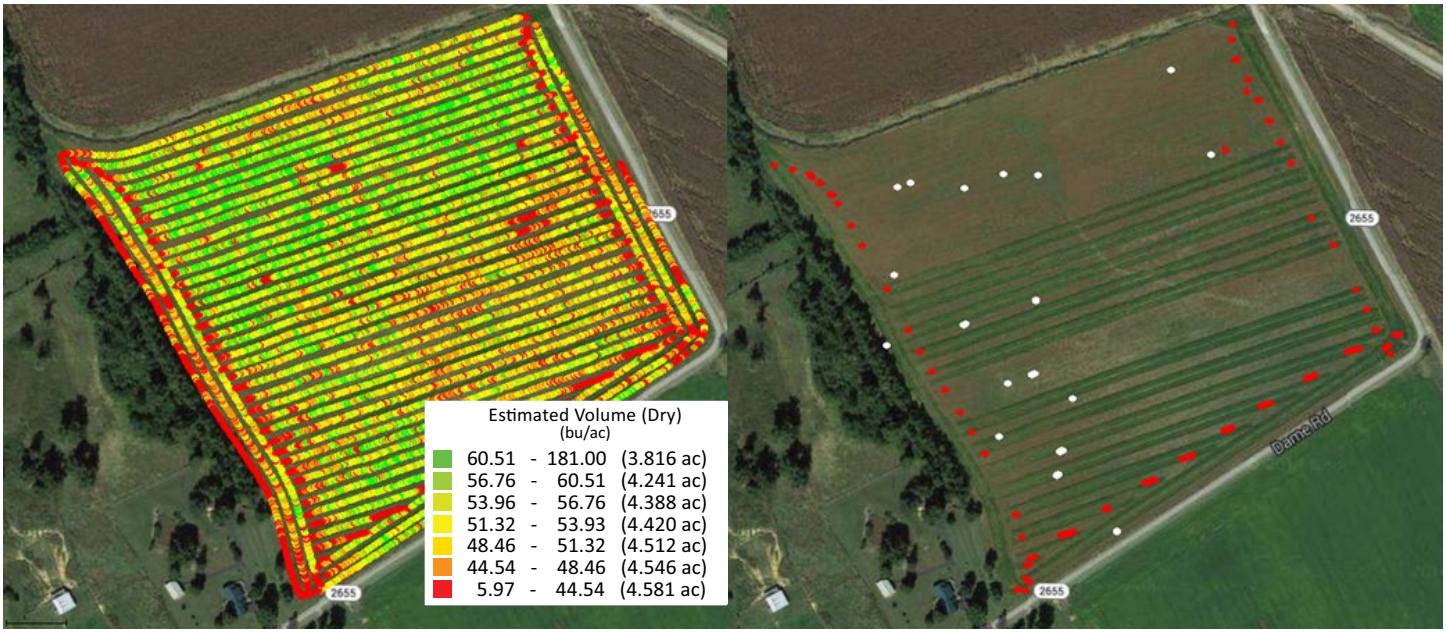

Figure 2: Traditional view of yield map (left) consisting of GPS latitude, longitude, and yield estimate data points. Yield data errors are shown at right, over-estimates (white, >80 bu/ac) generally due to quick deceleration with under-estimates from cut-width, lag time, or header setting errors (red, < 30 bu/ac) around edges of the field.

Figure 3: Grid-based (50 ft square) yield map (A) and contour-based yield map (B) using raw data points displayed in Figure 2.

Figure 4: Grid-based (50 ft square) yield map (A) and contour-based yield map (B) using clean data points from Figure 2.

Table 1: Summary statistics comparing grid- and contour-based maps created from raw and clean yield data.

Yield Statistic	*Raw Point Yield Data File (bu/ac)*	*Raw Yield Data Grid Map (bu/ac)*	*Clean Yield Data Grid Map (bu/ac)*	*Raw Yield Data Contour Map (bu/ac)*	*Clean Yield Data Contour Map (bu/ac)*
Minimum	6.0	31.2	36.9	32.2	38.0
Maximum	181.0	60.3	61.0	56.4	60.3
Average	51.7	51.7	53.0	51.6	52.8
Harvested Volume	1,577 (bu)	1,660 (bu)	1,701 (bu)	1,653 (bu)	1,693 (bu)

Having the maps converted into either grids or zones ties yield values to measurable areas within the field. This allows for further comparison between raw and cleaned yield data. *Table 1* contains a summary of minimum, maximum, and average yield values from the raw and clean yield data for both map styles (i.e., grid and contour). The field depicted in *Figures 3 to 5* was approximately 30 ac in size.

While some of these average values may not be alarming, viewing the actual differences between a gridded yield map from raw and clean data show errors more clearly. *Figure 5* shows the result of comparing a grid map of raw yield data from a similar map using clean data. In most locations, the raw data resulted in low yield estimates, cleaning the data brought estimated yield values higher. *Figure 5* also illustrates that some grid estimates were regularly off by -10% or greater. In a few areas, yield estimates were high, but for the most part, yield was not over-estimated in the raw data.

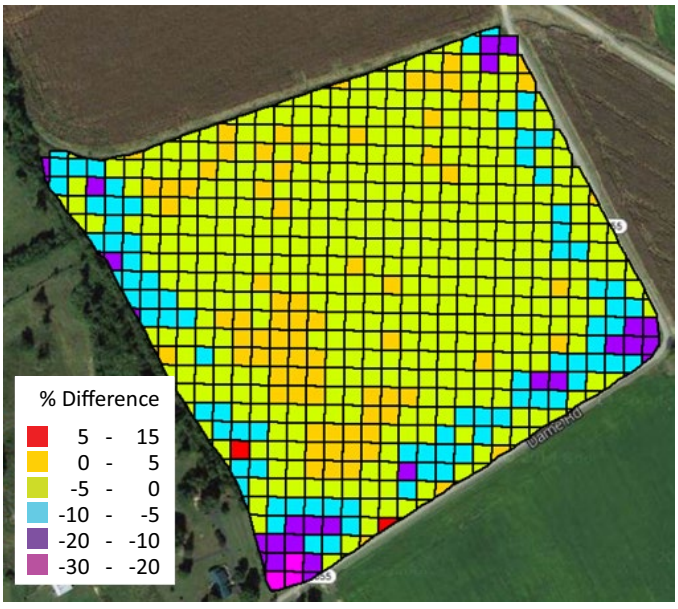

Figure 5: Differences (%) between clean and raw gridded yield data (raw grid data subtracted from clean grid data divided by clean grid data).

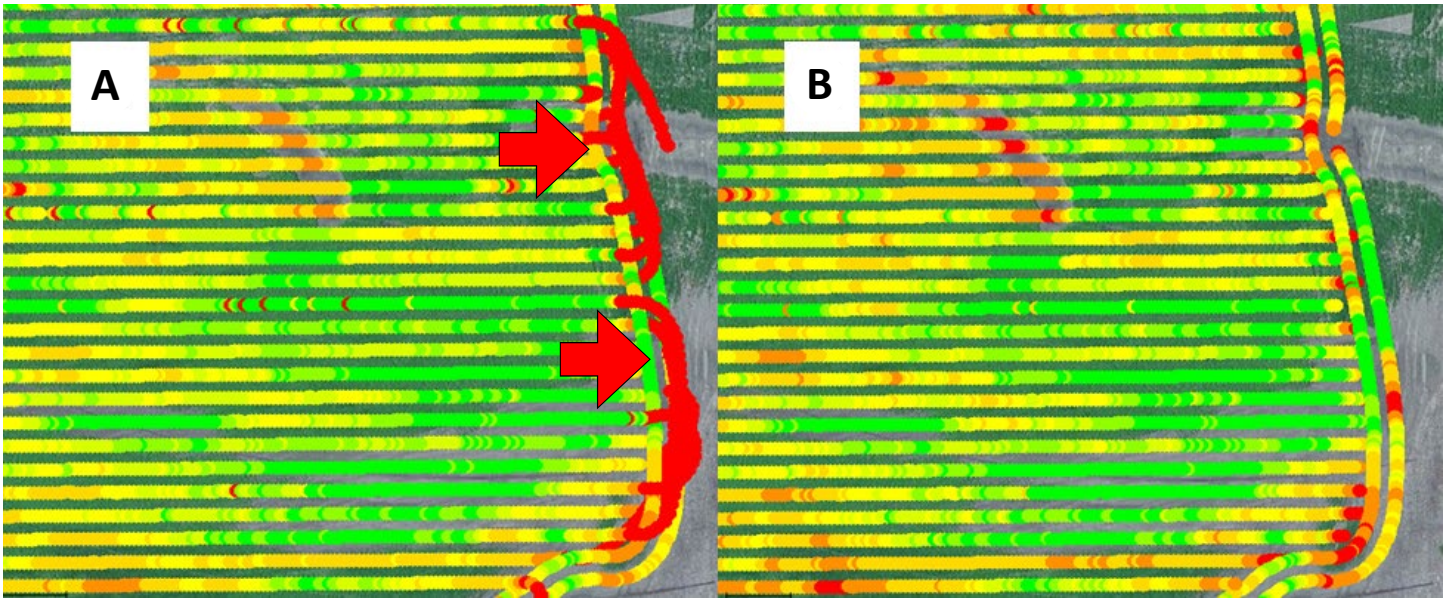

Figure 6: Header position sensor errors in raw yield data (A) and clean yield data after removal of errors (B).

The problem is that when trying to quantify yield values within zones, these errors are not consistently in the same location from year to year. *Figure 5* also highlights the fact that errors in yield data may not always skew the data in one direction; values may be estimated high or low depending on the nature of the error. The average yield for the study field shown was around 50 bu/ac; errors of the magnitude shown in *Figure 5* were in many cases greater than 10%. With automated data processing available today in most FMIS packages, creating a fertilizer Rx map with a yield map that contains errors would allow those errors to contribute directly to fertilizer estimates, ultimately resulting in over- or under-fertilizing.

Header Position Sensor Errors

Observing some of these errors commonly seen in point yield data maps may help users identify locations where incorrect data have been logged. Fields with irregular shapes are notorious for having cut-width errors, but even square fields can show the effects of poor data quality. *Figure 6 (A)* shows a field where the header position setting was either input incorrectly, or proper header control was not observed by the operator (i.e., header only down when harvesting). Point yield data are shown in *Figure 6 (B)* after the header position errors were removed. Cleaning this yield data and comparing grid values (50 ft square) exposed ***errors greater than 50%*** between datasets (*Figure 7*). The main problem was that many data points were collected while the header was down and not harvesting (i.e., no grain flow through the clean grain elevator). The effect of adding many of these invalid data points only increases the magnitude of the error in comparison to the cut-width error analysis previously shown.

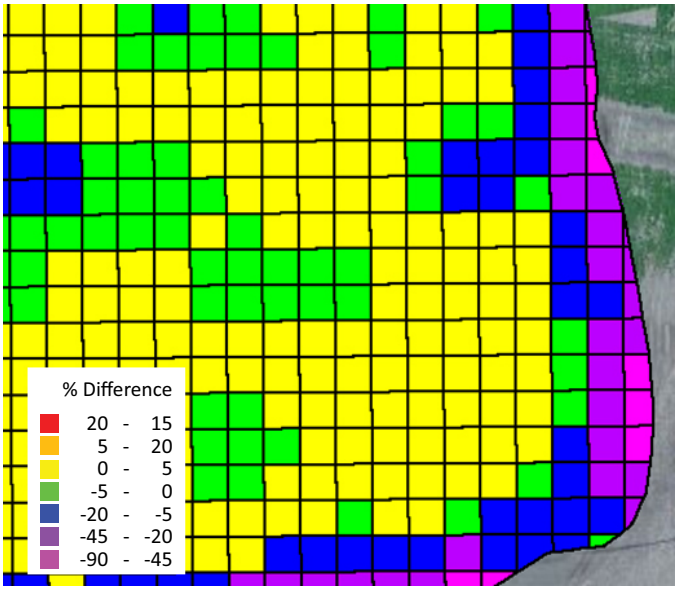

Figure 7: Differences (in %) between clean and raw gridded (50 ft square) yield data (raw grid data subtracted from clean grid data divided by clean grid data).

Table 2: Summary statistics comparing grid- and contour-based maps created from raw and clean yield data.

Yield Statistic	*Raw Point Yield Data File (bu/ac)*	*Raw Yield Data Grid Map (bu/ac)*	*Clean Yield Data Grid Map (bu/ac)*	*Raw Yield Data Contour Map (bu/ac)*	*Clean Yield Data Contour Map (bu/ac)*
Minimum	0	0.0	1.5	3.9	10.4
Maximum	1,785	291.0	262.7	234.6	242.7
Average	159.8	159.2	168.0	157.8	167.5
Harvested Volume	-	45,957 (bu)	48,522 (bu)	45,578 (bu)	48,354 (bu)

Table 2 contains a summary of minimum, maximum, and average yield values from the raw and clean yield data for both map styles (i.e., grid and contour). The field depicted in *Figures 6 and 7* was approximately 300 ac in size.

Lag Time Settings

All yield monitoring systems come with an option to adjust lag time or flow delay settings. This setting allows us to correct for the time needed for grain to pass through the machine as it enters the header and contact the mass flow sensor in the clean grain elevator. For most machines, a lag time setting of 10 to 15 seconds is appropriate but it can depend on the amount of grain passing through the threshing system. *Figure 8* illustrates a field where the lag time may have been a bit longer than actually needed. Data points are shifted several seconds back as the combine enters the crop; as it exits, data points are not recorded where there should be crop left. Another potential yield monitoring system issue could be GPS antenna offset settings. For a harvester, the distance from the harvest point to the GPS antenna on the cab should be entered in the in-cab display. As with previous errors, having several low estimates for yield data points along the east edge of the field in *Figure 8* would result in lower grid or contour values in the final yield map.

Travel distance measurements

Errors in travel distance measurement are often unavoidable many times during harvest operations. For the most part, speeding up is a more gradual process and less noticeable (often contributing to lower than expected yield estimates) than slowing down abruptly. Sudden stops are often necessary when a problem is encountered. For example, when a sink hole or plugged header is noticed, the operator must bring the machine to a complete stop as quickly as possible. More often than not, sudden stops result in a very high estimate of yield because a very short travel distance is recorded by the system while grain is passing over the mass flow sensor. This is unavoidable even with an accurate lag time setting. As shown in *Figure 9*, one or two points were logged (upper left) that contained yield estimates of over 1,500 bu/ac.

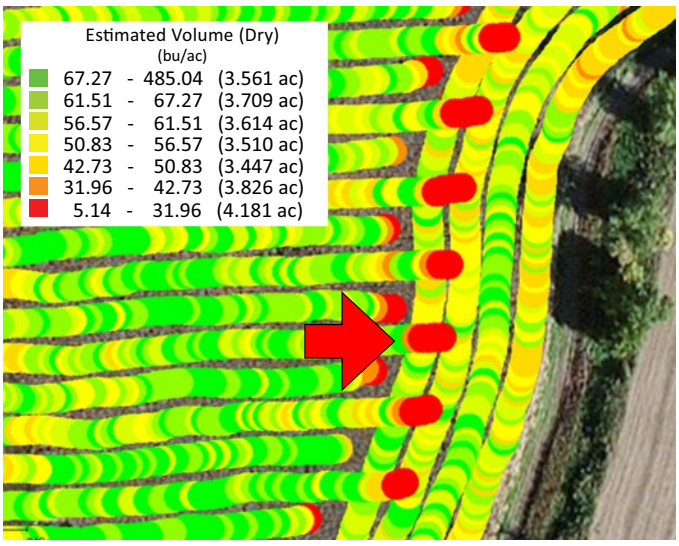

Figure 8: Effects of incorrect lag time settings (and potential GPS offset issue) illustrating data shifted too far along passes.

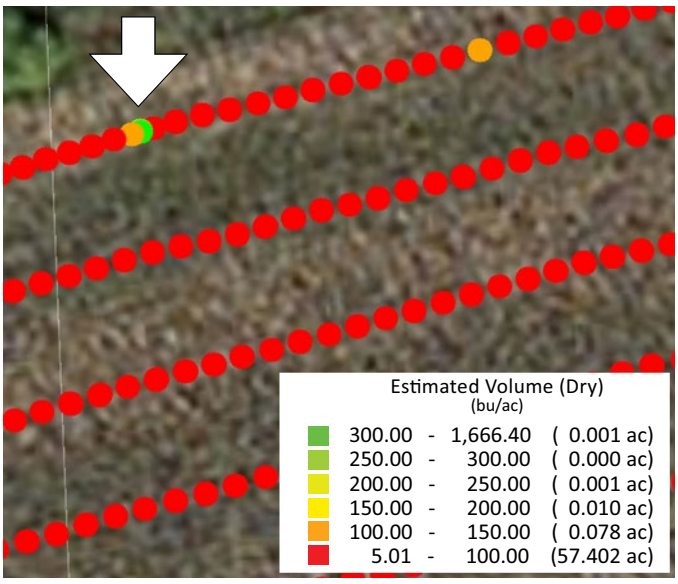

Figure 9: Effects of travel distance error measurement where a quick stop (at arrow) results in a high (1,666 bu/ac) yield data point estimate.

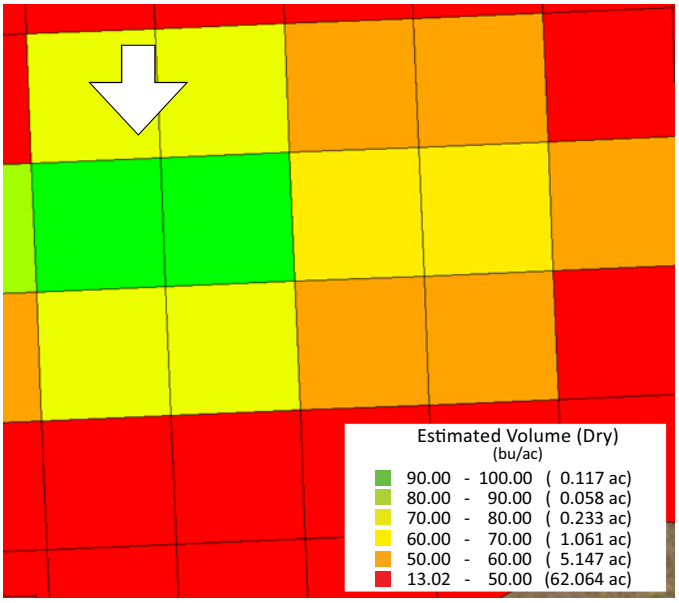

Figure 10: Grid yield map identifying location (at arrow) where an excessively high yield data point has affected grid values around it.

When these large point data values are converted to grid or contour maps through the interpolation process, they have an effect on the final yield map values. For the field in *Figure 9*, the grid yield map (50 ft square) values shown in *Figure 10* were artificially inflated for grid cells around the point in question. In this example, typical grid values for this field location were around 60 bu/ac; whereas the impact of the invalid data point caused grid values to exceed 95 bu/ac. A major problem with such errors is that they do not typically occur in the same field location every year and therefore may affect yearly comparisons to a greater extent that some other errors.

Potential Data Cleaning Software for Post-Processing

First and foremost, every raw dataset should be saved or backed up immediately after it has been downloaded from the yield monitor. Examples of data backup include saving data files to an external hard drive, DVD, cloud, or other form of secure storage device. This ensures that the original set of yield data is secure and can be accessed if something happens to the data during post-processing.

Some form of FMIS software is required for downloading yield data for mapping and analysis. During any yield data cleaning process, data points are deleted from the file. This is one reason that backing up yield information is so important. While many would view deletion of data points as counter-productive; the majority of software users will create an interpolated map for further analysis. Since yield data points are generally collected at a higher density, removal of a few points is generally corrected by the interpolation process. In the end, removing these invalid data points is a priority if more accurate data maps are desired.

Manual yield data point editing is generally offered by most FMIS packages. During this process, individual or groups of points may be manually selected by the user and deleted. While this process can be time consuming, in some cases it can be an effective way of removing incorrect data. Most FMIS packages allow users to export data into a comma separated (.csv) or text (.txt) file format. These files can then be imported into different software programs for editing. Microsoft Excel is one spreadsheet editing program that is capable of importing, editing, and saving yield data files. During the editing process, the data can be sorted and questionable points can be removed.

Another software tool developed by the USDA specifically for yield data cleaning is Yield Editor version 2.0 (USDA, 2014). This program is free to download and comes with a user's guide that has suggestions for yield data editing. While the program only accepts SMS Advanced Export files (.txt format) or GreenStar (.txt format) files, it is a very useful tool for removing many errors. The Yield Editor program allows for manual filter or threshold selection by the user for eliminating points or an automated system can be enabled to eliminate data points when files are uploaded. Another exciting benefit of the Yield Editor software is the automated cut-width cleaning process which can detect and eliminate most of these errors without involvement by the user (Sudduth et al., 2012). Once files have been cleaned, they can be exported (.txt format) and imported back into an FMIS package for further use. *Figure 11* illustrates an example of the Yield Editor interface where the user is able to select which filters to use for data cleaning and specify threshold values for those filters.

Methods of Yield Data Cleaning

Different methods exist for cleaning invalid yield data points; one focuses on physically-measured parameters (e.g., distance traveled, cut-width, moisture content) while the second involves statistical data interpretation, and a third uses minimum and maximum yield thresholds. It is our opinion that eliminating errors based on physical measurements should be the primary focus for those interested in yield data cleaning. For example, travel distance measurements of less than 1 ft or greater than 12 ft at a one second logging interval would correspond to speeds below 0.7 mph and greater than 8.2 mph, respectively. In most cases, harvesting while exceeding those speeds would be questionable. Setting a filter to eliminate points that exceed these thresholds would be simple in both Excel and Yield Editor. Moisture sensor readings that exceed 33% or fall below 10% are another example of acceptable thresholds to eliminate points where sensor readings are likely to create errors.

Statistical analysis methods involve a comparison between each yield data point and those points that are nearby. In most situations, a search distance is used to select yield data around the point in question. Average and standard deviation is calculated among those points. If the suspect point exceeds the average (plus or minus a multiple of standard deviations), it may be deleted; however, if it falls within the acceptable range it remains in the dataset. The third option method involves determining minimum and maximum values, 100 to 300 bu/ac for instance, and all points that exceed those values would be deleted.

In most cases, eliminating data points based on physical measurements will actually cover many points identified by both the statistical and minimum to maximum methods. The danger with using these two methods is that valid data points may be removed, a frequent occurrence with the minimum to maximum threshold method. When local clusters or groupings of data points are identified as errors, users should be careful to ensure that these groups are actually errors. For the most part, errors that result in very high yield estimates occur sporadically across a field. Errors that cause low yield estimates generally occur in point rows or as the combine exits headlands to begin harvesting the crop.

There is no ideal method for cleaning yield data errors using post-processing methods. Yield Editor does allow the user to save filtering (or cleaning) configurations and apply them to subsequent datasets which can save some time. The benefit is that if an acceptable filtering configuration is found for one or two fields for the combine during soybean harvest, that configuration can be saved and applied to the remaining fields from that season. When cleaning yield data points, all maps should be verified after the files have been processed to ensure that the remaining data are accurate. Users may also choose to compare the number of data points removed to verify that the cleaning process was not too aggressive.

Potential Downstream Impacts from Poor Data

In recent years, tools have become available to quickly process precision agriculture data for evaluating field

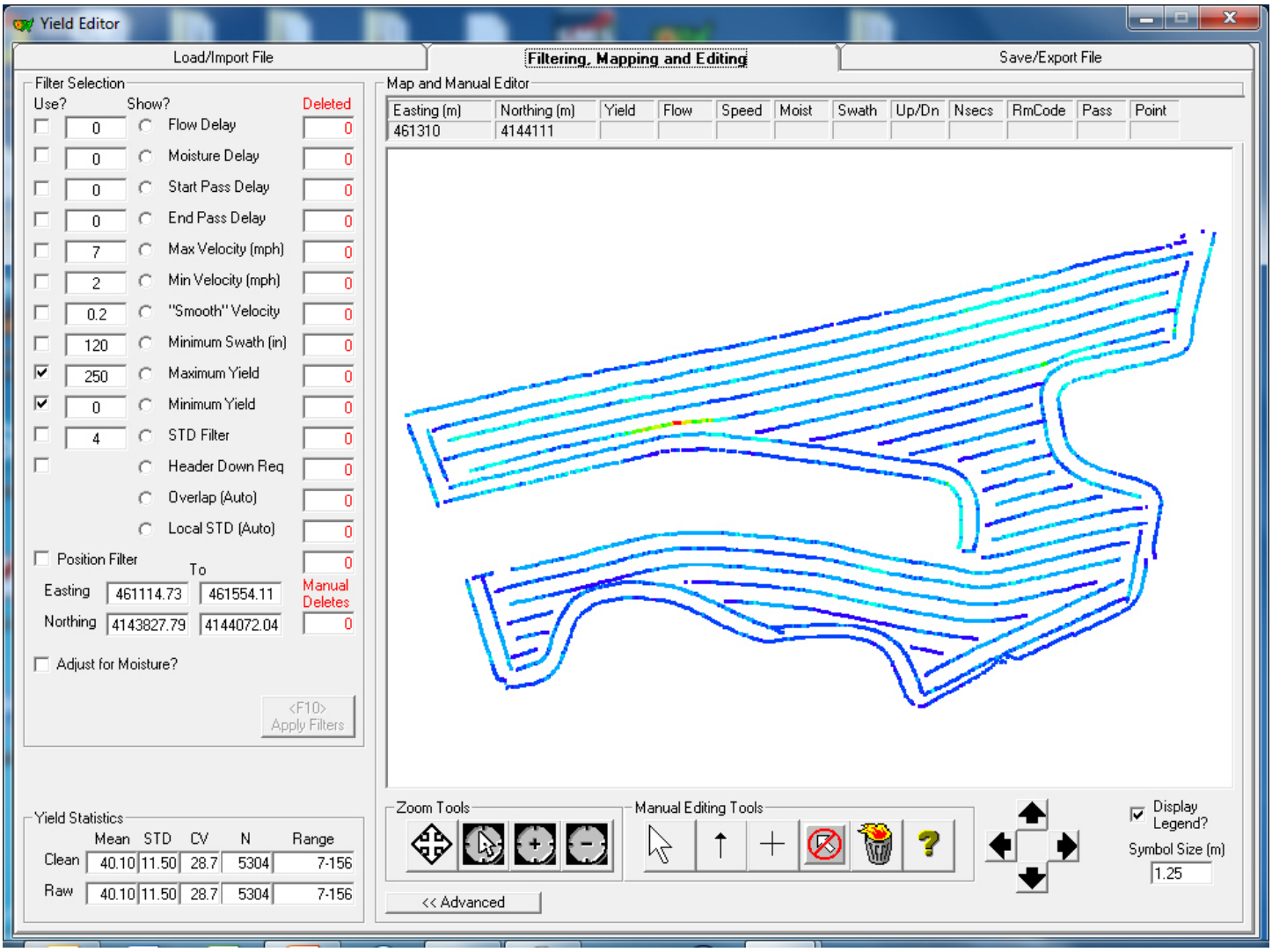

Figure 11: Example of yield data cleaning process using Yield Editor. Filter settings can be viewed along the left side of the screen with a map of current yield data points (deleted points may also be viewed).

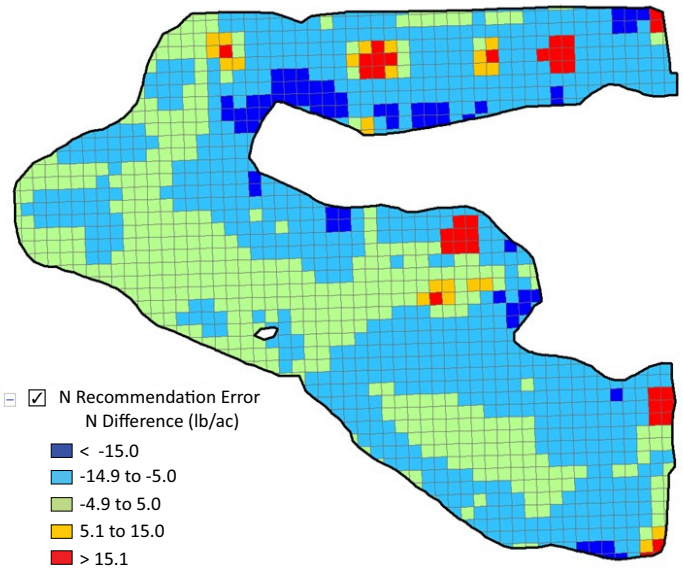

Figure 12: Map (50 ft grid) showing the potential differences in a N prescription map when using raw yield data versus cleaned yield data. In many instances, predictions of N can exceed 15 lb/ac.

management practices and creating prescription maps for product application. One example is using yield monitor data to create "expected yield" maps for use in nitrogen (N) recommendations. Yield monitor data (from one or multiple years) for a corn crop can be used in an equation to determine N needed in the upcoming year using processes similar to those shown by Shapiro et al., 2008. To illustrate the effects of poor yield data in this example, two spatial N recommendation maps were created, the first using raw yield data and the second using cleaned yield data. All other variables in the N recommendation equation (Shapiro et al., 2008) were held constant. After creating the two N (lb/ac) prescription maps, application rate values for the raw yield data map were subtracted from the clean yield data map. The result is shown in *Figure 12*. Even when the data are interpolated to a 50 ft square grid, the effects of the poor yield data are still evident. In many cases, N recommendations may be off by over 15 lb/ac (high and low) at different field locations. The result would be over and under application to many areas of the field. This highlights only one example of how poor data used in a well-accepted process (University N recommendation formula) could return poor results in the form of N application rate prescriptions.

Summary

As we move forward with agricultural field data collection and usage, it's important to remember that errors are going to be unavoidable in our datasets. Minimizing how these errors influence our data analysis by implementing BMPs will be very important in the future. Data collected and analyzed can be easily automated throughout the growing season as well as during winter planning activities. Putting good data into the yearly crop management examination process is critical to ensuring the information is accurate and useful. This article discusses many of the problems associated with creating yield maps from raw field data and potential methods for correcting many errors through post-processing (or cleaning) that data. Farm data management personnel should try to ensure that BMPs are consistently followed as much as possible throughout the data analysis process. Key points to remember are:

- A variety of sources can cause errors in yield data; while they are not as noticeable in grid- or contour-based yield maps, errors do impact the accuracy of these maps and may negatively affect any future analysis based on them.
- Different software packages may be used to clean raw yield data and remove many errors; most FMIS packages allow manual point-by-point deletion, Microsoft Excel and Yield Editor (from USDA) are also options.
- When cleaning yield data with any system, physical parameters should be the primary focus for eliminating yield points in our opinion. In most cases, these values will remove any points that statistical or minimum-maximum thresholds would delete as well.
- It is a good idea to check maps to ensure data cleaning was successful and ensure that neither too few nor too many points were removed.

References

Shapiro, C.A., R.B. Ferguson, G.W. Hergert, and C.S. Wortmann. 2008. Fertilizer Suggestions for Corn. The Board of Regents of the University of Nebraska. Available online at: *http://ianrpubs.unl.edu/live/ec117/build/ec117.pdf.*

Sudduth, K.A., S.T. Drummond, and D.B. Myers. 2012. Yield Editor 2.0: Software for Automated Removal of Yield Map Errors. In Proceedings of the 2012 ASABE Annual International Meeting. Available online at: *http://extension.missouri.edu/sare/documents/ASABEYieldEditor2012.pdf.*

USDA (United States Department of Agriculture). 2014. Software download for Yield Editor 2.0. Available online at: *http://www.ars.usda.gov/services/software/download.htm?softwareid=370.*