Data Integrity

By Joe Luck, Laura Thompson, Yeyin Shi and Nathan Mueller

TECH TOOLSHED

GOAL For data users to understand how errors during data collection or processing (including field study setup) may affect any analysis results and, ultimately, decision-making processes. Information regarding options for reducing errors through post-processing are also discussed.

VALUE STATEMENT Using accurate data when conducting field analyses or developing prescription maps is critical for ensuring positive future on-farm decisions and recommendations. For example, using historical yield data containing errors to drive nitrogen management decisions could lead to artificially high or low fertilizer recommendations, resulting in reduced yields or profits in subsequent years. Further, how the quality of those datasets might affect future decisions or analysis results is of primary concern as we continue down the path of data-driven agriculture.

Yield Monitor Data Quality

Georeferenced yield monitor data is considered one of the most valuable datasets collected throughout the growing season. Over the course of two decades, yield monitor data has progressed from simple printed map creation to an input for prescription map development to a dependent variable in on-farm research analyses. The effects of yield monitor errors can be extreme in some cases. Without visual inspection of the final yield map, errors may propagate through the analysis process and affect product prescription maps if they are used in this fashion.

Errors can negatively affect the crop management system when yield data is used for developing prescription maps for nutrients. In Figure 1, yield errors are visible along the headland areas where the header sensor was not engaged (indicating the system was not harvesting) in the raw yield data map. Those errors were removed via an automated post-processing software, Yield Editor, available free from the U.S. Department of Agriculture (USDA), shown in the clean yield data map (USDA, 2014). In both cases, N prescription (Rx) maps were generated (using the UNL nitrogen algorithm; Shapiro et al., 2008) where raw and cleaned yield data were used as the 'expected yield' inputs. In this worst-case scenario, certain areas of the field would have received lower nitrogen rates – in some cases over 75 lbs. N/ac less – with the raw yield data compared to the clean yield data. In the end, if some type of post-processing (either manually or automatically) had not been completed, errors like these could have affected future management and may have propagated errors into future growing seasons. Specifically related to yield monitor data, several options exist for post-processing errors using different software as discussed in Nebraska Extension Circular: EC2005 (Luck et al., 2015).

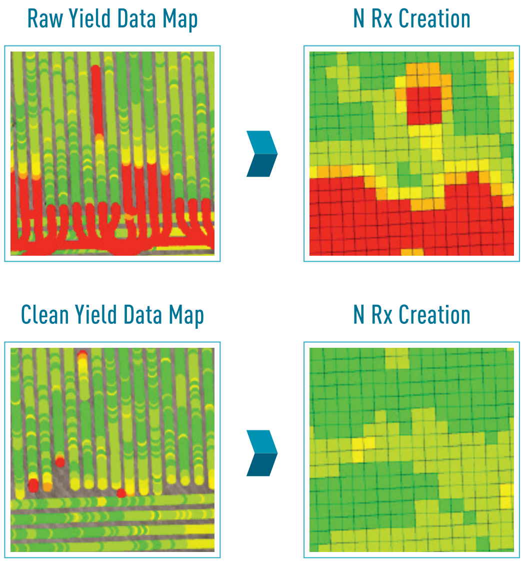

Raw Yield Data Map
N Rx Creation
Clean Yield Data Map
N Rx Creation

Figure 1. Illustration of how errors in yield data may propagate through to prescription maps when raw yield data (top) are used versus cleaned yield data (bottom).

Target Rate (Mass) (lb/ac)
🟩 150.00 - 198.61 (47.77 ac)
🟩 125.00 - 150.00 (30.53 ac)
🟨 100.00 - 125.00 (14.51 ac)
🟧 75.00 - 100.00 (13.05 ac)
🟥 -5.25 - 75.00 (26.75 ac)

Remotely-Sensed Data

As mentioned in the Data Sources section, remote-sensing data, including imagery obtained from satellites, manned aircrafts, drones and other data from canopy sensors mounted on ground implements have been providing useful information for growers for decades. Data types include imagery and point measurements. Point measurements can sometimes be collected by active reflectance sensors which have their own modulated light sources and independent with environmental lighting. Imagery are usually collected by passive reflectance sensors or cameras which measure the canopy reflectance of sunlight. Because of this, measurements from passive sensors are subject to environmental lighting change. Stronger incident light results in larger measurement values and vice versa.

The way to compensate the environmental lighting change on the measurements is to calibrate the sensor data using either calibration sensors or calibration targets.
We recommend:

- *Collecting data on sunny days with constant lighting conditions. For satellite or manned aircraft images, a cloudy day would result in cloud cover in the image (Figure 2)*
- *Use the same sensor/camera for a field throughout the season if possible*
- *Choose the multispectral camera with a downwelling light sensor (DLS) or an incident light sensor (ILS) installed at the top of the drone to monitor the environmental lighting, and use the environmental lighting data to correct the values in the final map generated*
- *Set up or select constant objects or calibration targets in or near the field for each data collection as references for post calibration*

Some other considerations are image spatial resolution and the quality of image stitching. Image spatial resolution usually needs to be fine enough to see the details of interest. For early season stand count, high spatial resolution is required, while for in-season nitrogen application, lower spatial resolution can be used as long as it is enough for the implement's application resolution. Assessment of data and imagery post-collection typically must be done visually which may be needed to understand field variability. Any distortion or misalignment can be discovered which is due to errors in flight path planning and/or image stitching in post processing.

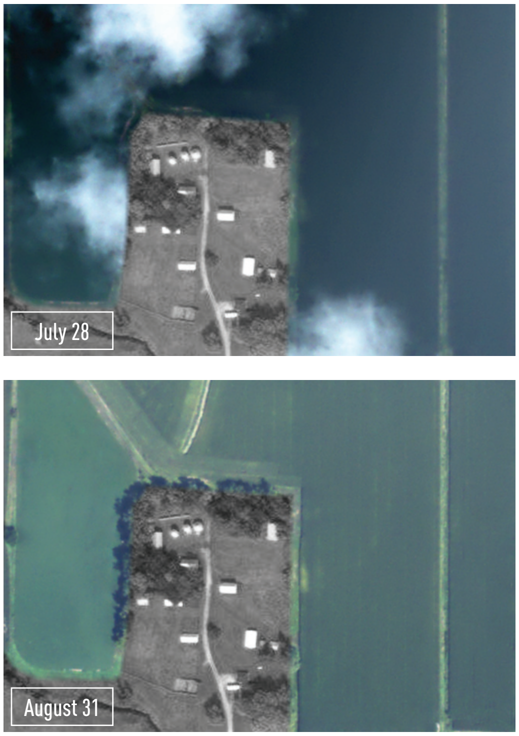


Figure 2. Example of how poor lighting conditions resulting from cloud cover may affect aerial remote-sensed imagery.

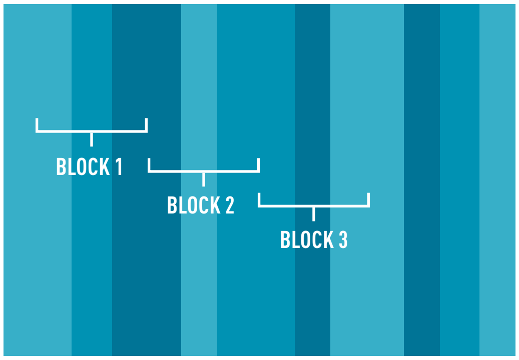
Experimental Design and Analysis of Field Studies

Appropriate statistical design of trials for on-farm research is a critical step in the planning and data collection process. Key elements of an appropriately designed field study include randomization, replication and blocking of the treatments included. In general, the number of replications and, thus, the area of the field dedicated to each treatment, should be similar among those treatments. Figure 3 illustrates three seeding rates applied across a field which were replicated in six blocks.

Notice that among the blocks, the seeding rates are randomized. In other words, the order is not consistent across each block. Mixing up the order of the seeding-rate treatments reduces the impact of any geographic differences that may occur across the field (e.g., yield variability due to constant slope).

Regarding treatment selection, care should be taken to choose large enough differences in target application rates

Figure 3: Illustration of three seeding-rate treatments that have been randomized and replicated within six blocks across a field.



Target Rate (ksds/ac*)

■ 32 ■ 36 ■ 40

*ksds/ac = 1,000 seeds per acre

such that a response from these treatments will be seen. For the study in Figure 3, 4 ksds/ac* was the difference between treatments selected to ensure the planter could generate this separation and greater effect from seed rates could be noticed. Further scrutiny of this example study data shows how poor data quality may affect the outcome of an analysis for which seeding rate was optimal for this field. Table 1 shows average yield values for each treatment from the raw yield dataset as well as a dataset (clean) once errors were removed. Differences in both yield averages and standard deviation of yield are evident. More importantly, once marginal return was calculated for the clean yield data, the results indicate the lower seeding rate was most economical.

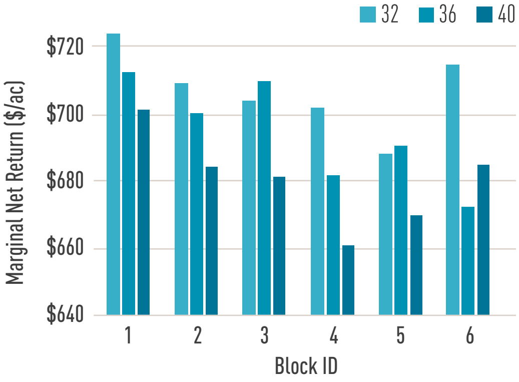
Table 1. Analysis summary differences for seeding rate study when raw versus clean yield data were used.**

	RAW YIELD DATA			CLEAN YIELD DATA		
Target Seed Population (ksds/ac)	Avg. Yield (bu/ac)	Avg. Yield St. Dev. (bu/ac)	Marginal Net Return ($/ac)	Avg. Yield (bu/ac)	Avg. Yield St. Dev. (bu/ac)	Marginal Net Return ($/ac)
32	237	27	690^{A}	241	18	707^{A}
36	242	30	691^{A}	243	20	695^{B}
40	239	34	663^{B}	244	18	680^{C}

**Letters for MNR indicate statistical significance in differences for raw or clean data (alpha = 0.1)

Viewing individual treatment strip Marginal Net Return (MNR) values illustrates the critical need for replication within field studies (Figure 4). The data in Figure 4 shows such instances where without replication, incorrect conclusions may have been drawn. For instance, had only three rate strips from block #3 been studied, the producer would have concluded that 36 ksds/ac was the optimal rate. However, averaging across multiple blocks improves the power of the study, and as shown in Table 3, 32 ksds/ac was the optimal economic seeding rate for this field based on the statistical analysis.

Figure 4. Individual MNR data per treatment strips within blocks for clean yield dataset.



Clean yield data: 4 out of 6 blocks indicated 32 ksds/ac would have resulted in higher MNR.

Resources

Luck, J.D., J.P. Fulton, and N.M. Mueller. 2015. Improving Yield Map Quality by Reducing Errors through Yield Data File Post-Processing. Institute of Agriculture and Natural Resources, University of Nebraska-Lincoln, EC2005. http://extensionpublications.unl.edu/assets/pdf/ec2005.pdf

Shapiro, C.A., R.B. Ferguson, G.W. Hergert, C.S. Wortmann, and D.T. Walters. 2008. Fertilizer Suggestions for Corn. Institute of Agriculture and Natural Resources, University of Nebraska-Lincoln, EC117. Available at: http://extensionpublications.unl.edu/assets/pdf/ec117.pdf.

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

For more information and links to additional resources, visit www.unitedsoybean.org/techtoolshed

Technical editing for this publication was led by Joe Luck, Ph.D., Laura Thompson, Yeyin Shi, Ph.D., and Nathan Mueller, Ph.D., University of Nebraska-Lincoln.

The United Soybean Board neither recommends nor discourages the implementation of any advice contained herein, and is not liable for the use or misuse of the information provided.