Kansas State University Department Of Agricultural Economics Extension Publication 07/03/2019

Distributed Ledger Technology Applied to Farm Data: Tracking Yield Monitor Data Changes With Blockchain

Terry W. Griffin (twgriffin@ksu.edu) – K-State Department of Agricultural Economics
Keith Harris (kdharris@ksu.edu) – K-State Department of Agricultural Economics
Jason Ward (jason.ward@ncsu.edu), NC-State University Department of Biological and Agricultural Engineering

July 2019

Abstract

This call to action proposes manufacturers of yield monitor sensors and developers of farm data systems integrates distributed ledger technology into yield monitor data inseparable from one another. The Analyst's Problem is defined as describing uncertainties of unknown data quality due to potentially being adversely affected by previous data management. Complexities such as the Analyst's Problem have been a barrier to implementing Big Data into reality and impeding the development of farm data communities. Applying distributed ledger technology to farm data is of interest to the developers of farm data communities, engineers and manufacturers developing crop sensors, and other agriculturalists interested in digital technology across the agribusiness sector.

Introduction

Heightened interest in farm data has arisen from commercialization of digital agricultural technology, aka precision agriculture, by investors outside of agriculture, agricultural manufacturers, service providers, farmers, and policy makers. Literally billions of venture capital dollars have been invested in farm data; however, data are prone to inherent problems and shortcomings that ushers in the notion of "Big Data" (Coble et al., 2018; Griffin et al., 2016). Farm data are known to be messy, i.e. exhibit limited veracity, especially data from combine yield monitors and other crop sensors.

Since the advent of precision agriculture, analysts assigned with managing and analyzing farm data have desired more information regarding how yield monitor data have been handled before reaching the analyst's computer. Analysts often question whether the data measurements originated from sensors that were properly calibrated, what observations were already been deleted, and if any relocation of data points occurred to adjust for flow delays. Therefore, the Analyst's Problem is in search of a solution.

Recently, substantial discussion on how sequential distributed ledger technology (DLT) impacts agriculture have been abuzz in the media and academia (Sylvester, 2019). Distributed ledgers are commonly known as "blockchain" although differences exist. A distributed ledger is a way of producing consensus about the facts that are necessary for commerce to function. Ledgers are the basic transactional recording technology at the heart of all modern business. Most conversations focus on merchandising and banking transactions, agricultural inputs, and production especially for traceability regarding food safety. The coordination of distributed ledgers impacts the value proposition of Big Data especially with respect to traceability, trust, and data quality (Coble et al., 2018). Blockchain has been applied to geospatial data management for geographical information systems (GIS) (Altaweel, 2019). Most agricultural technology specialists conceptually understand how DLT benefits supply chains on both sides of the farmgate, however many agriculturalists struggle with how DLT works. Data quality assurance will be necessary for farm data communities to be operational in the long-run.

Distributed ledger technology is applied not only to agricultural products, i.e. bags of seed arriving on a farm or bushels of corn leaving the farmgate, but to its controversial co-product, farm data especially yield monitor data. Distributed ledgers may be the solution to the Analyst's Problem of tracking who has done what to farm data. A vision for how DLT may conceptually be applied to farm data rather than specific mechanisms are described here. Concepts of applying DLT to yield monitor data are likely obvious to agriculturalists who have dealt with the Analyst's Problem or anyone familiar with farm data. This call to action could apply to many forms of farm data, however yield monitor data are the focus as it affects manufacturers and farm data software developers as detailed below.

Farm data

Farm data is comprised of many possible layers. Soil chemical (e.g. pH, levels of available nutrients) and physical characteristics (e.g. texture, topography) have been collected even before the advent of precision agriculture. Imagery from satellites, manned and unmanned aircraft, and other remote sensing platforms provide data on reflectance visible or invisible to humans. Imagery has been available for many decades but only recently attracted attention of researchers and industry due to low cost providers and attractive platforms. As-applied information on rates, products, and other agricultural inputs such as seed, fertilizer, and crop protection chemicals are feasible given introduction of automated controllers. Since global navigation satellite systems (GNSS, formally referred to as Global Positioning System or GPS when specific to the US Department of Defense system) became available for civilian use, site-specific yield data have been collected with combines equipped with instantaneous yield monitor sensors. Farm data layers provide opportunity for improved decision making and participation in farm data communities, i.e. Big Data. Opportunity exists to enhance data quality by proper data handling or cheat the system to fit desired narrative. Site-specific yield monitor data are the focus below.

The Analyst's Problem

Publicly available tools have been developed to reduce farm data veracity; however, these tools also provide an opportunity to increase data quality uncertainty. Sudduth et al. (2012) proposed rigorous tools, e.g. USDA ARS Yield Editor, to clean yield monitor data by flagging potentially erroneously measured observations and relocate data to more appropriate locations. Griffin et al. (2007) suggested best management practices to manage and clean yield monitor data with Sudduth's proposed tools. Utilization of tools such as USDA ARS Yield Editor increase the quality of yield monitor data, but only if proper processes are correctly performed. However, no system of documenting changes has been implemented that are inseparable from the data. Uncertainty exists by downstream users of the data coming from a farmer's field regarding if correct processes were performed, i.e. the Analyst's Problem.

The Analyst's Problem presents the troubling issue of dealing with farm data previously modified with unknown procedures. Although discerning analysts request original yield monitor data files, these data often arrive at the analyst's computer as delimited text or shape files. For instance, a data analyst may receive yield monitor data from a farmer or crop advisor but have no information on how that data were previously manipulated. Data may have been subjected to rigorous data cleaning protocols or potentially modified with nefarious intentions. In many cases, the data may have been modified by multiple upstream players who each believed they were improving data quality by removing observations (but without the benefit of knowing what had already been deleted). The farmer, and the farmer's advisors including crop advisor, crop consultants, fertilizer applicator, sales agronomist, or others who have access to soil nutrient analysis, as-applied, and yield data may all be considered upstream players. Field laborers operating equipment have also been known to impact yield monitor data quality. USDA NRCS technical service providers (TSP) (USDA NRCS) have been tasked with the development of comprehensive nutrient management plans (NMPs) based on data from soil nutrient analysis and site-specific yield monitor data. The discerning analyst has very little confidence any further analyses would be reliable given prior data management uncertainties.

The Analyst's Problem is analogous to the "whisper game" (or "telephone game") that many people played in elementary school; where children line up and whisper a sentence that they were just given. The last child announces what they heard and compares to the original statement; rarely would these two statements match when playing the whisper game. Moral to the story: if analysts do not know how the data were manipulated by potentially every upstream person along the line (e.g. farm employee, farm operator, crop consultant, technical service provider, sales agronomist, others), then they are not likely to trust any results based on analysis of that data.

Status quo best management practices regarding yield monitor data

The current *status quo* best management practices for processing yield monitor data have not substantially changed in at least a decade (Griffin et al., 2007). Currently, yield monitor data can be accessed via logged or controller area network (CAN) data; however, processes are not drastically dissimilar between the two. Logged data may originate in one of several proprietary file formats (e.g. *gsd from John Deere GS1, RCD from John Deere GS3, *yld or *.ilf from AgLeader) essentially one or more for each brand and generation of yield monitor or combine harvester. Yield monitor data via CAN (Rohrer et al., 2019; Ward, 2004) may be subjected to most of the same data cleaning parameters as logged data but also offers the added benefit of possibly being already loaded in cloud-based platforms.

Best management practices for logged yield monitor data have included using field operator data modules (FODM) provided by equipment manufacturers to process proprietary file formats (Griffin et al., 2007). FODMs have been integrated into desktop and cloud-based software; and at least one software tool dedicated to this procedure (specifically FOViewer). Most farm software capable of opening proprietary file formats are capable of exporting data in the format required by USDA ARS Yield Editor (Sudduth et al., 2012). Workshops have been offered by Land Grant Universities and state colleges to farmers, their advisors, and others interested in farm data (Griffin 2018). Farm mapping software may impose filtering procedures that removes yield monitor observations; the recommendation is for analysts to set software to not remove or filter any data so that the analyst retains full control.

The USDA ARS Yield Editor tool provides the analyst with a range of filtering operations. In all, 11 parameters can be set to relocate observations to proper positions and flag observations that may have been erroneously measured. These parameters can be set manually by the analyst's expertise or the analyst may accept results from automated cleaning routines. Once the analyst is satisfied with parameter settings, the remaining cleaned (and deleted) data points can be exported from USDA ARS Yield Editor in a delimited text file including coordinates and yield. Once exported, these data can be imported into farm mapping software, sophisticated geographical information systems (GIS), or online platforms.

Desktop and cloud-based farm software tools perform some sort of data cleaning when data are imported. Users of these services can manually remove yield observations that are believed to be erroneous. For traceability purposes, service providers retain local copies of 1) original yield data coming into the system and 2) one or more cleaned versions of data for use in specific purposes (data are sometimes cleaned with certain procedures for specific purpose). Although storing copies of the original data are standard operating procedures for any analyst, it is also the failsafe in the event questions arise regarding how data were manipulated. In many cases, the "cleaned" data would be discarded (or at least set to the side) when questions arise, then the updated cleaning procedure would be applied to (hopefully) the original data.

Regardless of whether data originate from logged or CAN processes or manipulated on desktop or cloud computing systems, uncertainties regarding how data were managed remains. One possible solution to the Analyst's Problem is presented next.

Distributed ledger technology applied to the Analyst's Problem

One possible solution to the Analyst's Problem of uncertain data quality from prior data manipulation could be distributed ledger technology (DLT). Distributed ledger technology may be the system that ensures data have not been inappropriately manipulated or at the very least documents what changes have been made by specific individuals.

Distributed ledger technology allows tracking of who manipulated yield monitor data and how that person manipulated that data. More than one person may be tracked along the process that may begin as early as calibration of the yield monitor. Although determining if yield monitor was calibrated after the fact may be impossible, DLT could be applied to data collected in previous growing seasons assuming proprietary file formats from the yield monitor are available. At the very least the tracking would occur from the time the sensor measures grain flow to the current data analysis. Each analyst accessing data would also be recorded as potential users of data. Data flagged for deletion by USDA ARS Yield Editor (or similar) would be retained such that current analyst could undo previous cleaning procedures without resorting to the original data (similar to the failsafe standard operating procedures used by many cloud-based software platforms). Regardless of whether the analyst uses USDA ARS Yield Editor or a commercial farm software tool, the need for distributed ledger type of tracking of who performed what manipulation to the data are needed in the agricultural industry.

Distributed ledgers need to be linked to the data being recorded at the yield monitor such that these are inseparable. Distributed ledgers preferably exist before data measurement at the sensor so that calibration information is included. Calibration, or thereof, has been a farm data issue since the commercialization of the yield monitor. Improper, or lack of sufficiently frequent, calibration has caused analysts and farmers to question the validity of data and resulting decision recommendations. Knowledge of how and when yield monitors were calibrated enhances confidence of data analysis and decision making.

As on-vehicle generated data finds more applications beyond the farm gate, the discussion shifts from purely data quality to data authenticity. Downstream users such as reporting, insurance, regulatory compliance, and nutrient management, will all need some method to verify data. Existing methods such as a hash function or checksum can be used to compare two discrete files for differences. Algorithmic techniques can describe file contents but do not describe what was changed, when, and by whom. So the Analyst's Problem remains and the type of data manipulation continues to lack description.

Incentives to cheat the system: farm data example

Distributed ledger technology was developed to prevent moral hazard. Incentives exist to cheat most systems, and farm data are no exception. Regulatory and compliance requirements are anticipated that incentivize farmers, crop consultants, technical service providers, and input suppliers to manipulate soil nutrient, as-applied fertility, and yield data in their favor. This is especially true near environmentally sensitive areas where overapplication of nutrients have been suggested as nonpoint source pollution.

As digital agriculture becomes more common, pay-for-performance marketing models are expected to become more prevalent. Farmers and input suppliers may benefit from removing selected data points or entire passes of yield monitor data to avoid being at a relative disadvantage. As farm data becomes more valuable, farm employees may desire to cover up anomalies caused by improper equipment operation. These are only a few of innumerable incentives for individual players along the farm data chain to attempt to cheat the system. Many other examples could easily be described by several different individuals.

Limitations of distributed ledger technology applied to Analyst's Problem

Distributed ledger technology requires internet connectivity. In many crop-producing locations in the United States and other developed nations, limited wireless internet connectivity is available (Mark et al., 2016; Whitacre et al., 2014). Wireless connectivity tends to improve as service providers update technology, therefore this limitation may soon resolve itself. However, connectivity impedes DLT at locations where farm data are being collected by yield monitors.

Similar to other aspects of farm data communities, the number of users on the network impacts the feasibility of the system. Scalability is likely a major issue in real-life implementation of DLT for farm data issues. Larger-scale implementation of DLT concerns both technological scalability and social scalability, i.e. number and types of users. This is analogous to the value of farm data communities as a function of the number of growers, farms, fields, and acres in the system (Miller et al., 2018).

A robust DLT for agricultural data must include the ability to track multiple versions of the same file even as file names, data structures, and contents change. The same yield monitor file could find applications with multiple different end users who apply different data quality practices or downstream analyses. Even after the data content changes or is aggregated, the originating file should be clearly identified. Even the definition of data manipulation will need clarification. For example, changing the spatial projection of a data file will not alter any contents other than the spatial coordinates and may not require a new line in the ledger. A similar approach will be needed for file format conversion. Even if the ledger is not tracking all versions of a file, the original data file will need a unique identifier so that the starting point of the data can be verified. A similar question can be raised about data exported from the terminal versus yield data logged from the CAN network over the same field. Both datasets fundamentally describe the same system but the numeric scale, numeric values, and spatial scale may be different as the terminal file will employ data aggregation or smoothing to create a consistent one-point-per-second, or similar, data file.

Summary and moving forward

Distributed ledger technology could be applied to farm data within the farmgate and beyond. Specifically, a need exists to track how yield monitor data have been managed, manipulated, and cleaned including calibration of the yield monitor sensor. The Analyst’s Problem has been presented along with how DLT may one possible solution. Substantial opportunities exist for tracking how farm data have been manipulated. These opportunities may impact the overall value of Big Data in agriculture.

In addition to changes made to yield data, the DLT documents specific individual players who make changes including the person(s) involved in calibrating the monitor and operating the combine during grain harvest operations. Individual operators may be required to log into the yield monitor system during these mechanical operations, a process that would ultimately benefit farm data communities. These capabilities would allow recourse in the event of misappropriation of data or some sort of suspected interference.

Although most readily applying to yield monitor data, DLT concepts are applicable to a broad range of farm data including as-applied data and remotely sensed data (e.g. satellite and aerial imagery). Tracking changes to as-applied fertilizer and grid soil nutrient analyses may be important to compliance reporting, e.g. nutrient management plans and applications in environmentally sensitive areas. Tracking of changes to imagery may stimulate interest from insurance underwriters concerned about players attempting to cheat the system. Although enhancing farm data systems by implementing DLT is obvious to most analysts in agriculture, the opportunity to implement these systems are wide open.

Acknowledgements

We appreciate informal discussions or reviews from colleagues. Gary Wagner, a Minnesota farmer who has long been considered an innovator of precision technologies, and an educator of farm data practices to college students, provided invaluable feedback and spirited conversation on the blockchain concept applied to yield data.

References

Altaweel, M. 2019. GIS and Blockchain. GIS Industry, GIS Lounge. 19 April 2019.

Coble, K., Ferrell, S.L., Mishra, A., and Griffin, T.W. 2018. Big Data in Agriculture: A Challenge for the Future. *Applied Economics Perspectives and Policy*, 40(1): 79–96.

Griffin, T. W., Brown, J. P., & Lowenberg-DeBoer, J. (2007). Yield Monitor Data Analysis Protocol: A Primer in the Management and Analysis of Precision Agriculture Data. Purdue University.

Griffin, T.W. 2018. Why yield monitor data cleaning workshops are offered. Blog 5 November 2018.

Griffin, T.W., T.B. Mark, S. Ferrell, T. Janzen, G. Ibendahl, J.D. Bennett, J.L. Maurer, and A. Shanoyan. 2016. Big Data Considerations for Rural Property Professionals. *Journal of American Society of Farm Managers and Rural Appraisers*, 79:167-180

Mark, T.B., Griffin, T.W. and Whitacre, B. 2016. The Role of Wireless Broadband Connectivity on 'Big Data' and the Agricultural Industry in the United States and Australia. Special Edition: The Value of Big Data in Agriculture: Inputs, Farming and Processing. *International Food and Agribusiness Management Review*, 19(A):43-56

Miller, N.J., Griffin, T.W., Goeringer, L.P., Ellixson, A. and Shanoyan, A. 2018. Estimating Value, Damages, and Remedies when Farm Data are Misappropriated. *Choices*. 4th Quarter 2018 33(4)

Rohrer, R.A., Pitla, S.K., Luck, J.D. 2019. Tractor CAN bus interface tools and application development for real-time data analysis, *Computers and Electronics in Agriculture*, 163

Sudduth, K.A., Drummond, S.T., & Myers, D.B. (2012). Yield Editor 2.0: Software for automated removal of yield map errors. Paper presented at the 2012 American Society of Agricultural and Biological Engineers Annual International Meeting. ASABE Paper 121338343. St. Joseph, MI: ASABE.

Sylvester, G. 2019. e-Agriculture in Action: Blockchain for agriculture: Opportunities and challenges. Food and Agriculture Organization of the United Nations.

Ward, J.K. 2004. Development of an On-vehicle Controller Area Network for Data Acquisition. Lexington, Ky.: [s.n.], 2004. Print.

Whitacre, B.E., Mark, T.B., and Griffin, T.W. 2014. How Connected are Our Farms? *Choices, The Magazine of Food, Farm, and Resource Issues*. 29(3) 9 pp