Field Crops Research 221 (2018) 130–141

Contents lists available at ScienceDirect

Field Crops Research

journal homepage: www.elsevier.com/locate/fcr

Sifting and winnowing: Analysis of farmer field data for soybean in the US North-Central region

Spyridon Mourtzinis[a,*], Juan I. Rattalino Edreira[b], Patricio Grassini[b], Adam C. Roth[a], Shaun N. Casteel[c], Ignacio A. Ciampitti[d], Hans J. Kandel[e], Peter M. Kyveryga[f], Mark A. Licht[g], Laura E. Lindsey[h], Daren S. Mueller[i], Emerson D. Nafziger[j], Seth L. Naeve[k], Jordan Stanley[e], Michael J. Staton[l], Shawn P. Conley[a]

[a] Department of Agronomy, University of Wisconsin-Madison, Madison, WI 53706, USA
[b] Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68583-0915, USA
[c] Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA
[d] Department of Agronomy, Kansas State University, Manhattan, KS 66506, USA
[e] Department of Plant Sciences, North Dakota State University, Fargo, ND 58108-6050, USA
[f] Iowa Soybean Association, Ankeny, IA 50023, USA
[g] Department of Agronomy, Iowa State University, Ames, IA 50011-1010, USA
[h] Department of Horticulture and Crop Science, The Ohio State University, Columbus, OH 43210, USA
[i] Department of Plant Pathology and Microbiology, Iowa State University, Ames, IA 50011, USA
[j] Department of Crop Sciences, University of Illinois, Urbana, IL 61801, USA
[k] Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108, USA
[l] Department of Crop and Soil Sciences, Michigan State University, East Lansing, MI 48824, USA

ARTICLE INFO

Keywords:
Soybean
Survey data
On-farm yield
Statistical technique
Conditional inference tree

ABSTRACT

Field trials are commonly used to estimate the effects of different factors on crop yields. In the present study, we followed an alternative approach to identify factors that explain field-to-field yield variation, which consisted of farmer survey data, a spatial framework, and multiple statistical procedures. This approach was used to identify management factors with strongest association with on-farm soybean yield variation in the US North Central (NC) region. Field survey data, including yield and management information, were collected over two crop growing seasons (2014 and 2015) from rainfed and irrigated soybean fields (total of 3568 field-year observations). Fields were grouped into technology extrapolation domains (TEDs) that accounted for soil and climate variation and 9 TEDs were selected based on the number of fields needed to detect yield differences due to management as determined using power analysis. Average yield ranged from 2.5 to 5 Mg ha^{-1} across TEDs, with field yield distributions in half of the domains having a distributional peak that was close to maximum yields. Conditional inference trees analysis was chosen among 26 statistical procedures as the approach that best combines ability to detect and rank factors (and their interactions) with greatest influence on on-farm yield and relatively easy interpretation of results. Survey data from *ca.* 150 fields in each of the nine TEDs allowed us to identify key management factors influencing yields for an agricultural area that includes *ca.* 7 million ha sown with soybean. In five of the nine TEDs, highest yields were observed in early-sown fields. Other factors explaining on-farm yield variation were maturity group, and in-season foliar fungicide and/or insecticide application, but, in some cases, their influence on yield depended upon sowing date and water regime. While the approach proposed here cannot establish cause-effect relationships conclusively, it can certainly provide a focus to replicated field experiments in relation to which management factors to investigate. We believe that future agronomic studies based on farmer survey data can greatly benefit from *ex-ante* identification of most important TEDs (relative to crop area and production) as well as determination of minimum number of farmer survey data that needs to be collected from each of them based on expected yield differences and variability. The approach is generic enough to be applied in other crop producing regions as long as farmer data and associated climate and soil databases are available.

* Corresponding author.
E-mail address: mourtzinis@wisc.edu (S. Mourtzinis).

https://doi.org/10.1016/j.fcr.2018.02.024
Received 18 November 2017; Received in revised form 22 February 2018; Accepted 23 February 2018

1. Introduction

Average crop yields will need to increase substantially during the next 33 years to meet expected food demand increase while avoiding massive expansion of cropland area (Tilman et al., 2011; Alexandratos and Bruinsma, 2012; Grassini et al., 2013). This challenge can be achieved by increasing the rate at which best management practices are identified and adopted for a particular soil-climate context. Replicated field experiments are used in agricultural research to test new technologies and management practices. In these experiments, researchers selectively manipulate a production factor and, by comparing final yield against the yield of a "control" treatment, the magnitude of the yield response and its economic profitability are assessed. A limitation of this approach is that it often examines the effect of management practices at a small number of sites and years due to practical constraints (*e.g.*, costs, logistics, *etc*.). Hence, extrapolation of their findings is typically confined to a narrow range of environments. Likewise, field experiments cannot test the effect of a large number of production factors (and their interactions) on yield due to the large number of plots that would be needed. And, finally, the management selected as "background" for these experiments (*e.g.*, sowing date, tillage method) will also influence crop responses to a given technology or management. Given these limitations, it is relevant to search for alternative, cost-effective approaches that provide an indication of the management practices that perform best for a given climate-soil context.

Farmer survey data can be utilized as a cost-effective source of information to identify yield constraints and fine-tune management practices so that these yield limitations can be ameliorated or eliminated (*e.g.*, Calvino and Sadras, 2002; Sadras et al., 2002; Lobell et al., 2005; Tittonell et al., 2008). An advantage of using farmer data is that it allows examination of opportunities for yield increase within the range of current management practices that are both cost-effective and logistically feasible in farmer fields. Another advantage of using farmer data is that, if surveyed fields are properly contextualized relative to their biophysical environment, it is possible to explore and quantify management × environment interactions (Rattalino Edreira et al., 2017). Such assessment would allow identification of suites of management practices that perform best for a given environment and provide a focus to traditional, costly field experiments so that they can target those management practices with the most likely impact on crop productivity and input-use efficiency.

Statistical analysis of farmer self-reported data poses challenges that need to be addressed to make meaningful and unbiased inferences. For example, in field experiments, different levels of a given management or input are assigned to experimental units. These experimental units are carefully selected based on their similarity, in order to avoid confounding factors influencing yield and to minimize the error variance. Each treatment level is applied to several experimental units ('replicates') to obtain an estimate of average yield and its variation. In contrast, farmer data do not follow an experimental design and lack random allocation of experimental units and replication. Variation in soil, weather, and management practices across fields results in minimal control over error variance. Several management practices (or inputs) may be applied simultaneously, leading to multi-collinearity, making interpretation of results more challenging (Hastie et al., 2001). Additionally, it may be the case that a given management practice does not appear to be significantly associated with yield simply because that practice has already been widely adopted across fields (*e.g.*, cultivars with herbicide-resistance traits). Despite all these limitations, farmer data have the potential to give an indication of the most important yield-limiting factors in a given region, which can, in turn, then be tested in more detailed field trials to experimentally confirm cause-effect relationships.

We argue here that proper analysis of farmer field data, when evaluating the influence of management factors on yield, requires: (i) a biophysical spatial framework to cluster fields into groups with relatively similar climate and soil, (ii) use of appropriate statistical methods that can handle the nuances associated with the structure of farmer survey data and to identify management interactions, and (iii) a deep agronomic knowledge and understanding of the cropping system context to interpret results and translate them into practical recommendations. Application of a spatial framework to identify causes of yield gaps has been addressed in a previous study (Rattalino Edreira et al., 2017). A major limitation of this previous study, as well as other studies looking into the causes of yield gaps (*e.g.*, Mercau et al., 2001, Sadras et al., 2002; Grassini et al., 2011, 2015; Silva et al., 2016), is that the analysis was limited to a comparison of management practices between high- versus low-yield fields or regressions between yield and individual or multiple management practices for a given climate-soil domain, without an explicit attempt to rank the importance of each management practice based on its influence on yield and to identify interactions.

In the present study, we addressed the second requirement listed above, that is, the use of a proper statistical technique to identify and rank management factors (and their interactions) influencing soybean yield in farmer fields. We focused on soybean fields in the North Central US region, which accounts for *ca.* 85% of US soybean production and *ca.* 30% of global production (FAOSTAT, 2016; USDA-NASS, 2016). The objective of this study was to utilize self-reported farmer data and multiple statistical techniques, together with a spatial framework, to identify the management practices with greatest influence on rainfed and irrigated soybean yields across diverse climate and soil conditions.

2. Materials and methods

2.1. Database description

Soybean yield and management practices data were collected from 3568 fields sown with soybean in 2014 and 2015 across 10 states in the US NC region: Iowa (IA), Illinois (IL), Indiana (IN), Kansas (KS), Michigan (MI), Minnesota (MN), Ohio (OH), North Dakota (ND), Nebraska (NE), and Wisconsin (WI) (Fig. 1). Detailed description of the database is provided elsewhere (Rattalino Edreira et al., 2017). The majority of surveyed fields were non-irrigated, except in Nebraska, where there were both rainfed (34%) and irrigated fields (66%) located within the same region. Maize was the predominant prior crop (88% of total fields). Average regional yield represents *ca.* 22 (rainfed) and 13% (irrigated) of the estimated yield potential, indicating a relatively small (but still exploitable) room for increasing farmer yields through fine tuning of current management practices (Rattalino Edreira et al., 2017).

Farmers reported data on field location, average yield (adjusted to 13% moisture content), and management practices, including sowing date, seeding rate, row spacing, variety name, tillage method, drainage system, total irrigation amount (for irrigated crops), seed treatment, fertilizer inputs, lime, manure, and pesticides (Table 1). Farmers also reported incidence of other field adversities such as pests, diseases, weeds, iron deficiency chlorosis, hail, waterlogging, and frost. Data were subjected to quality control to remove erroneous entries. Likewise, fields subjected to unmanageable field adversities (*e.g.*, hail, frost, flooding) leading to substantial yield losses were excluded from the analysis. To do this, fields reported as affected by any of the aforementioned adversities were grouped within regions with similar soil and climate (see Section 2.2), and we excluded those that fall below the 25th percentile of the yield data distribution within each region-year. To summarize, we excluded data from fields affected by unmanageable adversities *and* that fell below the 25th percentile of the yield distribution in each climate-soil domain; these data were excluded from all the statistical analyses, as well as tables and figures presented here. We did not exclude fields that suffered from drought, heat stress, temporary waterlogging, or disease, insect or weed pressure. After quality control, the database contained data from a total of 3216 fields sown with soybean in 2014 and 2015 (92% of total surveyed fields). Fields were