IFPRI Discussion Paper 02332

April 2025

Detecting Cumulative Effects of Inputs Within the Flexible Production Function Framework Through LASSO Shrinkage Estimation

Implications for Potassium Fertilizer Use in India

Hiroyuki Takeshima
Avinash Kishore

Innovation Policy and Scaling Unit
Development Strategies and Governance Unit

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

The International Food Policy Research Institute (IFPRI), a CGIAR Research Center established in 1975, provides research-based policy solutions to sustainably reduce poverty and end hunger and malnutrition. IFPRI's strategic research aims to foster a climate-resilient and sustainable food supply; promote healthy diets and nutrition for all; build inclusive and efficient markets, trade systems, and food industries; transform agricultural and rural economies; and strengthen institutions and governance. Gender is integrated in all the Institute's work. Partnerships, communications, capacity strengthening, and data and knowledge management are essential components to translate IFPRI's research from action to impact. The Institute's regional and country programs play a critical role in responding to demand for food policy research and in delivering holistic support for country-led development. IFPRI collaborates with partners around the world.

AUTHORS

Hiroyuki Takeshima (H.takeshima@cgiar.org) is a Senior Research Fellow of the Innovation Policy and Scaling Unit of the International Food Policy Research Institute (IFPRI), Washington, DC.

Avinash Kishore (a.kishore@cgiar.org) is a Senior Research Fellow with IFPRI's Development Strategy and Governance Unit, New Delhi, India.

Abstract

Despite recognition of the potentially significant cumulative effects of input use on annual crop output—such as the effect of applying inorganic fertilizer in one year on crop output in the subsequent year—real-world evidence from smallholder farmers' fields in lower-income countries remains scarce. We narrow this knowledge gap using unique district-level and farm-household-level annual panel datasets in India. We start with flexible translog production functions, which are well-suited for identifying cumulative effects in farmers' actual production environments. We then apply shrinkage methods (LASSO and GMM-LASSO) to approximate the production function with reduced parameter dimensions, addressing various challenges such as multicollinearity among multiple inputs, including the same inputs from the current and previous years, and potential endogeneity in inputs. Our results indicate that, throughout the shrinkage process, potassium remains a key predictor of outputs, while other inputs (land, labor, capital, irrigation, and other fertilizer nutrients) drop out. More important, the cumulative quantity of potassium from both the previous and current years is a consistently more critical determinant of production than the quantity of potassium from the current year alone, demonstrating the potassium's significant cumulative effects. These patterns hold at both the district and farm levels across diverse agroecologies and cropping systems. Furthermore, the dynamic panel data analyses suggest that farmers' use of potassium in the current year is significantly negatively affected by its use in the previous year, potentially stabilizing outputs across years. Our results support earlier agronomic findings suggesting that the cumulative effects of potassium may be relevant across wider geographic regions than previously thought.

Keywords: Multiyear translog production function; machine learning; LASSO and GMM-LASSO; potassium; district-level panel data; farm household–level panel data; India

Acknowledgments

The authors thank the Tata Cornell Institute (TCI) and the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) for maintaining and making the data publicly available, which enabled this study. The authors are also grateful to David Spielman for his constructive comments. This study was supported by the United States Agency for International Development (USAID)–funded project Cereal Systems Initiative in South Asia (CSISA), the CGIAR Research Program on Transforming Agrifood Systems in South Asia (TAFSSA), and the CGIAR Sustainable Farming Science Program. The opinions expressed here belong to the authors and do not necessarily reflect those of USAID or CGIAR. The authors are responsible for all remaining errors.

1 Background

The cumulative effects of agricultural inputs on production, with implications for the demand dynamics, have long been recognized in the agricultural economics literature (e.g., Moser & Barrett 2006). For example, farming experiences acquired from previous years can affect labor productivity in the current year (e.g., Berger 2001). Other inputs can also exhibit cumulative effects through their effects on land quality, including the use of agricultural equipment that can either improve soil aeration or cause land compaction, or irrigation that affects hydrological conditions. The changes in land quality can, in turn, alter land productivity in subsequent seasons. Inorganic fertilizer is another potentially important input that can have cumulative effects on production, not only in the current year but also in subsequent years, through the buildup of soil nutrients and lagged effects (e.g., Meena et al., 2019; Wang et al., 2024).

The cumulative effects of fertilizer on agricultural production can be substantial. Nonetheless, knowledge gaps regarding their significance remain, especially in smallholder production systems in low- and middle-income countries. While the potential cumulative effects of specific fertilizer nutrients have been suggested anecdotally or through more focused agronomic studies (e.g., Askegaard et al. 2004; Johnson 1986; Kihara et al. 2016; Meena et al., 2019; Wang et al., 2024), they have less commonly been examined at representative field levels that reflect smallholder farmers' actual production conditions in low- and middle-income countries. Addressing this knowledge gap is important because the significance of cumulative effects at the field level influences farmers' demand for fertilizer in a given year and determines the effectiveness of policies such as fertilizer subsidies, which are often the primary policy instruments to increase yields and output.

While flexible production function forms, including translog production functions, have been increasingly used in agronomic studies (e.g., Nayak et al., 2022b' Quilty et al., 2014; Qiao et al., 2021), addressing the knowledge gap within the production function framework has been challenging for various reasons. Reasonably long panel data on agricultural production, which are required for such analyses, are generally scarce. Estimation is complicated by significant multicollinearity among specific fertilizer nutrients (e.g., N, P, and K), as well as among fertilizer nutrients and other inputs. In addition, there is similar multicollinearity between input use in the current year and previous years, as well as potential endogeneity of input variables.

We narrow this knowledge gap using annual panel data of agricultural production at both the district level—the District-Level Database for Indian Agriculture and Allied Sectors (DLD data hereafter) (ICRISAT & TCI 2024)—and the household level—the Village Dynamics in South Asia (VDSA) project of the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT 2024) (VDSA data hereafter) from India. These datasets enable us to obtain unique evidence of the lagged and cumulative effects of fertilizer. In doing so, we apply the Least Absolute Shrinkage and Selection Operator (LASSO) method, often considered a machine learning approach, to address multicollinearity among input variables and robustly compare the current-year and lagged effects of nitrogen (N), phosphorus (P), and potassium (K). Furthermore, we apply the Generalized Method of Moments (GMM)-LASSO method by Caner (2009) and Shi (2016) to further address endogeneity in farm household–level data (VDSA data).

Our results suggest that the aggregate quantity of K from both the current and previous years is a consistently more important determinant of production than the quantity of K from the current year alone. In addition, using Arellano and Bond's (1991) dynamic panel linear probability model (LPM) and Wooldridge's (2005) panel Tobit model, we find that K use is more strongly affected by its dynamic relations with the previous year's use, while its response to price changes is weak and slow.

Although our results highlight the cumulative effects of K, our analytical approach also considers the whole set of inputs, such as land, labor, capital, and other expenses. Therefore, we arrive at our results not because we focus specifically on fertilizer, but because fertilizer (especially K in the case of India) proves to be the input for which cumulative effects matter (unlike those for other inputs). In other words, our results show that cumulative effects are significant only for K, and similar cumulative effects are somewhat limited for other inputs and other fertilizer nutrients, such as N. Analyzing the cumulative effects within the production framework, rather than the reduced form fertilizer response framework, also prevents us from erroneously concluding that the cumulative effects of K are significant, when in fact, the true effects are occurring through other inputs that may be highly correlated with K.

India is a suitable case to investigate our research questions. Inorganic fertilizer and its primary nutrients (N, P, and K) have become increasingly critical inputs in agricultural production in India. In 2022, India accounted for approximately 20 percent of global N and P use and 5 percent of K use in agriculture (FAO 2024).[1] Furthermore, India decontrolled prices of P and K fertilizers in 2011, potentially causing significant spatiotemporal variations in the quantity of fertilizer use. Combined with the LASSO approach, these variations can be exploited to identify their effects on production.

Our study contributes to various strands of the literature. We add to the literature on the production function in contexts where more medium-term inputs potentially have significant effects on production, such as in the cultivation of perennial crops (e.g., Chand 1994), the roles of infrastructure and research and development (R&D) (Fan & Pardey 1997), natural resource stock (e.g., Fuglie et al. 2021), or the roles of education on human capital outcomes (e.g., Bernal 2008). We do this by offering an empirical approach to test the effects of accumulated inputs versus non-accumulated inputs. We also contribute to studies highlighting the accumulated or lagged effects of certain fertilizer nutrients (e.g., Johnson 1986; Askegaard et al. 2004; Kihara et al. 2016) by providing formal evidence at field levels. In addition, we contribute to studies assessing fertilizer responses in South Asia (e.g., Kishore et al. 2021; Rashid et al. 2013; Takeshima et al. 2017) by providing farm household–level evidence from India. Last, methodologically, our study adds to the growing literature on model selections (Tibshirani 1996;) and their applications to agronomical contexts (e.g., Ahrens et al. 2020; Arshad et al. 2023; Goncharov et al. 2023; Mondal et al. 2021; Mourtzinis et al. 2018; Nayak et al. 2022a) by applying the LASSO approach to robustly compare the predictive power of the cumulative inputs with single same-year inputs alone.

[1] In 2022, fertilizer use per area in India also reached 120 kg/ha of N, 47 kg/ha of P (approximately double the global- average), and 10 kg/ha of K (approaching half of the global average).

The remainder of this paper is structured as follows. Section 2 describes the data, Section 3 discusses the empirical approaches, and Section 4 summarizes descriptive statistics. Section 5 presents the results, and Section 6 concludes.

2 Data

Our primary datasets consist of district-level panel data (DLD data) and household-level panel data (VDSA data) from India.

Compiled collaboratively by the Tata-Cornel Institute for Agriculture and Nutrition (TCI) and the ICRISAT, the DLD data contain information on annual agricultural production and use of major inputs, including area cultivated, agricultural workers, agricultural credit used for the purchased services and inputs (e.g., machines, agrochemicals), the share of land irrigated, and fertilizer use (including N, P, and K) aggregated at district levels between 2002 and 2016. The DLD data capture the production quantity and price data of major grains, legumes, root crops (potato), oil crops, sugarcane, cotton, fruits, and vegetables, which collectively account for more than 90 percent of total crop production value (FAO 2024). Through this data, we compute total production values for each district and use them as our dependent variables in the production function models discussed in the next section.[2] We start with 496 districts for which all production and inputs are available, representing approximately 85 percent of all districts in India. All observations are reported annually, except for the data on district-level agricultural workers, which is reported only for 2001 and 2011. Because the agricultural worker population tends to follow a more stable trend compared to other production factors, we interpolated annual data from 2002 and 2010 using the annualized growth rate calculated from 2001 and 2011, then extrapolated through 2016. We also treat approximately 5 percent of year-district observations as missing due to the absence of information on some inputs. After dropping observations with missing information, we focus on unbalanced panel samples of 7,054 observations from 496 districts from 2002 through 2016.

VDSA data are farm household-level data compiled by ICRISAT. The sampling frame of VDSA data consists of 18 purposively selected villages in 5 southern and western Indian states (Andhra Pradesh, Gujarat, Karnataka, Madhya Pradesh, and Maharashtra). Based on the census of households in each of these villages, households were classified into four groups according to farm size and land ownership. A predetermined number of households were then randomly selected and interviewed annually between 2010 and 2014. The survey details are described on the VDSA project website.[3] After dropping observations with missing information, we focus on the 3,041 panel observations in the VDSA data.

In addition to DLD and VDSA data, we obtained historical rainfall and drought data from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) (Funk et al. 2015) and the Standardized Precipitation Evapotranspiration Index (SPEI) (Vicente-Serrano et al. 2010), respectively; nighttime luminosity data from NOAA (2022); and soil property data from the Food and Agriculture Organization of the United Nations (FAO et al. 2012). For each survey year, we extracted these data for each DLD as average values within DLD boundaries and for

[2] For certain commodities without price data in DLD (such as potatoes, fruits, and vegetables), we calculate prices by dividing the production values by the quantities reported in FAO (2024) and impute production values for these commodities applying the calculated price to the DLD-reported production quantities.
[3] http://vdsa.icrisat.ac.in/vdsa-microdoc.aspx

each village location in the VDSA data (soil data are time-invariant and extracted as constant across years).

3 Empirical model

3.1 Flexible translog production function

For both DLD data and VDSA data, we start with the following translog production function:

$$\ln Y_{ht} = \alpha + \sum_{X} \beta_X \ln X_{ht} + \frac{1}{2} \sum_{X} \sum_{X^*} \beta_{XX} \ln X_{ht} \ln X^*_{ht} + \beta_S S_{ht} + f_h + \varepsilon_{ht} \quad (1)$$

(h = district j for DLD data, farm household i for VDSA data)

in which Y_{ht} is the total value of crop production in year t. X_{ht} is a vector of input use, including land, labor, capital,[4] and irrigation. X_{ht} also includes the quantity of each of three fertilizer nutrients, N (N_{ht}), P (P_{ht}), and K (K_{ht}). X^*_{ht} is an alias for X_{ht}. Notation S_{ht} is a vector of other exogenous time-variant variables, f_h is time-invariant fixed effects for panel h, α and β are estimated parameters, and ε_{ht} is an idiosyncratic error term.

For DLD data, the control variables S_{jt} include the district-average nighttime light index (a proxy for the level of overall economic development), as well as weather shocks proxied by absolute z scores of annual total rainfall and annual average drought index (SPEI) relative to their historical means and standard variations during the pre-sample period (1980–2000). S_{jt} also includes the year dummy and its interactive terms with state dummy variables to account for any year-specific statewide shocks on production. For VDSA data, S_{it} includes the year dummy and its interaction with the village dummy variable to account for any year-specific village-level shocks.

Equation (1) is a standard translog production function without including lagged effects of inputs. In our empirical analyses, we consider expanding (1) by adding lagged input values, so that

[4] In our analyses with DLD data, capital is proxied by the total value of agricultural credit taken out in the district each year. In India, direct institutional credit from cooperatives, scheduled commercial banks, and regional rural banks reach the individual farmers directly, while indirect loans are issued by cooperatives, scheduled commercial banks, regional rural banks, rural electrification corporations, etc., to institutions and organizations that indirectly serve farmers' interests (Haque & Goyal 2021). A significant majority of these credits are used to purchase inputs such as machines and payment for these services. By specifying key input variables in production function as (1), we assume that most agricultural inputs other than fertilizer, land and irrigation costs are financed through agricultural credit.

For VDSA data, we measure capital through two separate variables: (1) horsepower of the equipment multiplied by the duration (hours) they are used, which proxies the level of productive agricultural assets, and (2) spending on all purchased inputs (including hired labor), except fertilizer.

$$\ln Y_{ht} = \alpha + \sum_X \beta_{X1} \ln X_{ht} + \frac{1}{2}\sum_X \sum_{X^*} \beta_{X11} \ln X_{ht} \ln X^*_{ht}$$
$$+ \sum_X \beta_{X0} \ln X_{h,t-1} + \frac{1}{2}\sum_X \sum_{X^*} \beta_{X00} \ln X_{h,t-1} \ln X^*_{h,t-1} + \frac{1}{2}\sum_X \sum_{X^*} \beta_{X10} \ln X_{ht} \ln X_{h,t-1}$$
$$+ \sum_X \beta_{XC} \ln X_{htc} + \frac{1}{2}\sum_X \sum_{X^*} \beta_{XC1} \ln X_{ht} \ln X_{htc} + \frac{1}{2}\sum_X \sum_{X^*} \beta_{XC0} \ln X_{h,t-1} \ln X_{htc} \quad (2)$$
$$+ \frac{1}{2}\sum_X \sum_{X^*} \beta_{XCC} \ln X_{htc} \ln X^*_{htc} + \beta_S S_{ht} + f_h + \varepsilon_{ht}$$

(h = district j for DLD data, farm household i for VDSA data)

in which notation X_{htc} is an average of X_{ht} in t and $t-1$, that is:

$$X_{htc} = \frac{X_{ht} + X_{h,t-1}}{2}, \quad (3)$$

with c denoting "cumulative."

We define that a certain input has cumulative effects on output if both β_{X1} and β_{X0} are statistically significant, or if β_{XC} is statistically significant. The functional form (2) is deliberately expanded so that it serves as an initial model specification for the LASSO process to narrow down the most influential variables among X_{ht}, $X_{h,t-1}$, and X_{htc}. Conceptually, X_{htc} and its interaction terms, $X_{ht}X_{htc}$ and $X_{h,t-1}X_{htc}$, are redundant since they are subsumed in X_{ht}, $X_{h,t-1}$, and $X_{ht}X_{h,t-1}$. Nonetheless, X_{htc} is included for empirical practicality, primarily to correctly detect cumulative effects when X_{ht} and $X_{h,t-1}$ are highly correlated. If a certain input exhibits a cumulative effect, both X_{ht} and $X_{h,t-1}$ (as well as their interaction terms with other common inputs) have significant coefficients in true production function. However, if there is significant multicollinearity between X_{ht} and $X_{h,t-1}$, the LASSO process may select only one of X_{ht} or $X_{h,t-1}$, forcing us to incorrectly conclude that cumulative effects are absent. If X_{htc} is also included in the model and approximates well the cumulative use of the input in t and $t-1$, the LASSO process may select X_{htc} (while dropping X_{ht} and $X_{h,t-1}$), detecting the evidence of cumulative effects.[5]

3.2 LASSO applied for the estimation of production function

A major challenge in the estimation of full translog production function (2) is the severe multicollinearity among input variables X_{ht} as well as $X_{h,t-1}$ and X_{htc}. Such multicollinearity can arise when many inputs are significant complements or substitutes, and/or when the quantity of inputs used shows little variation between t and $t-1$.

[5] It is still possible that the true model has both terms $\beta_X \ln X_{ht}$ and $\beta_{XC} \ln X_{htc}$ as significant for the same inputs. In such a case, the aggregated effect of $\ln X_{ht}$ on output $\ln Y_{ht}$ is $\beta_X + \frac{\beta_{XC}}{2}$, while it is $\frac{\beta_{XC}}{2}$ for $\ln X_{h,t-1}$.

LASSO approaches are particularly effective in selecting the most influential variables X_{ht} among many highly collinear variables.

3.2.1 LASSO applied when all variables are exogenous (DLD data)

In the agricultural production economics literature, input variables aggregated at administrative levels, such as districts, are more likely to be exogenous (e.g., Gong 2018; Meng et al. 2024) because endogenous responses at more disaggregated levels (such as farm household) tend to be averaged out at aggregate levels. In the estimation of (2) using DLD data, we therefore apply standard LASSO approaches, which are widely available for models consisting of exogenous variables only.

LASSO approaches, originally developed by Tibshirani (1996), essentially penalize the number of explanatory variables over the overall fit of the model and identify parsimonious model specifications by dropping many variables that are deemed only marginally influential or not influential (either unconditionally, or conditionally in the presence of other highly collinear variables). Among various LASSO estimators, we primarily apply theory-driven ("rigorous") LASSO approaches (RLASSO hereafter). The RLASSO has been found to be generally reliable among different classes of LASSO estimators, particularly in facilitating causal inferences when there are many control variables and in avoiding false positives (falsely identifying variables as influential when they are not) (e.g., Ahrens et al. 2020). Appendix A provides additional details of the concepts underlying the overall LASSO approaches, as well as the RLASSO approach.

The basic setup of the LASSO estimator applied to Equation (2) for the DLD data is

$$\widehat{\beta_X}(\lambda) = argmin \frac{1}{n}\sum_{j=1}^{n} \varepsilon_{jt}^2 + \frac{\lambda}{n}\sum_{X}^{p} \psi_X |\beta_X| \qquad (4)$$

$$\psi_X = \sqrt{\frac{1}{n}\sum_h (\ln X)^2} \text{ for each of } X_{jt},\ X_{j,t-1} \text{ and } X_{jtc}$$

in which λ is a penalization (also called "turning") parameter that determines how much the number of estimated parameters is penalized in selecting the model. A larger (smaller) value for λ implies more (less) penalization of the number of parameters, and therefore leads to selecting fewer (more) parameters in the model. Notation p is the number of input parameters selected among X_{jt}, $X_{j,t-1}$, and X_{jtc}. Notation n represents the number of observations. Notation ψ_X is the "loading" parameter, which is set specific to each variable $X(= X_{jt}, X_{j,t-1}$ and $X_{jtc})$ so that parameter βs are independent of the unit of each X. Note that in our application of LASSO (4) to production function (2), we always keep $\beta_S S_{jt}$, intercept and household fixed effects (meaning β_S for all S_{jt}, as well as intercept α, f_j) in the regression, and our LASSO approach focuses exclusively on identifying additional key weather shock parameters.

3.2.2 GMM-LASSO

Applying the LASSO approach to the estimation of (2) for VDSA data requires addressing the fact that, unlike DLD data, many X_{it}, $X_{i,t-1}$, and X_{itc} variables are potentially endogenous. We therefore apply the GMM-LASSO framework proposed by Caner (2009) and operationalized by Shi (2016). (Appendix A describes the GMM-LASSO method in detail.)

Conceptually, GMM-LASSO works similarly to LASSO; it modifies the criterion function under the standard GMM (which is minimized to solve for parameters under the standard GMM) by adding the penalty terms on the number of included parameters so that the solution process also favors more parsimonious models with a fewer number of parameters in the GMM estimates. Specifically, as is described in Appendix A, GMM-LASSO applies the relevant version of two commonly used criteria: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

In GMM-LASSO, we use instrumental variables (IVs) that are considered to exogenously affect the unit costs of each input. Specifically, our IVs for GMM-LASSO applied to VDSA data (denoted as Z_{it} in Appendix A) include the size of farmland owned, household size, value of total agricultural equipment owned, wages (for adult male workers for land preparation), the village-level shares of samples using irrigation, and the prices for urea, DAP, and MOP, respectively.

Both LASSO applied to DLD data, and GMM-LASSO applied to VDSA data, provide more parsimonious approximations of full translog function (2), depending on different penalization methods used.

One of the challenges in VDSA data is that—unlike in DLD data where quantities are aggregated at district levels—significant shares of inputs, particularly K, contain values censored at 0, which complicates their natural log transformations in (1). In our primary GMM-LASSO for VDSA, following a long tradition in the economic literature (e.g., Jacoby 1993; Takeshima 2017; Williams 1937), we add 1 to all X_{it} so that censored observations are also included in the analyses. In the Results section, we show that our results are robust when adding different values, such as 0.001.

3.3 Dynamics of fertilizer nutrient use at the farm household level (VDSA data)

As we show in the Results section, analyses of Equation (2) using the LASSO approach suggest significant cumulative effects of K (K_{ht}). We then conduct supplementary analyses using VDSA data to show that the dependence on significant cumulative effects has crucial implications for the use of K—specifically, the dynamic mechanisms that determine its use, which extend beyond the static mechanism often assumed for fertilizer in the literature.

3.3.1 Arellano and Bond (1991) dynamic panel LPM

We first assess the dynamics of K use decision at the extensive margin (i.e., whether or not to use K in a particular year). Unlike intensive margins, dynamics at the extension margin can be modelled with a less-restrictive assumption and more commonly used methods. We model the decision at the extensive margin as an LPM framework and estimate it within the well-established Arellano and Bond (1991) dynamic panel LPM, as previously done in the literature (e.g., Hendricks et al. 2014; Kerr et al. 2014; López & Yadav 2010; Takeshima et al. 2023). Specifically, we estimate

$$Q_{it}^{*} = \gamma + \gamma_{Q} Q_{i,t-1}^{*} + \gamma_{\Pi} \Pi_{it} + \gamma_{S} S_{it} + \gamma_{Z} Z_{it} + h_{i}^{*} + \upsilon_{it} \quad (5)$$

(for each of N and K)

in which Q_{it}^{*} is a binary variable taking the value of 1 if the farm household i used fertilizer nutrient in year t (for each of N and K), and 0 otherwise. Π_{it} is the price of each nutrient faced

by farmer i in t. The parameter h_i^* is unobserved farm household fixed effects. As is well-known, the lagged dependent variable can be endogenous due to its correlation with the compound disturbance term $h_i^* + \upsilon_{it}$, a common challenge in a dynamic panel model (Nickell 1981). The Arellano and Bond (1991) estimator addresses this endogeneity by instrumenting the lagged dependent variable by Π_{it}, S_{it} and the transformation of lagged variables in earlier t as additional IVs.

3.3.2 Wooldridge's (2005) dynamic panel Tobit model

We also assess the dynamics of fertilizer use accounting for the intensive margins by applying Wooldridge's (2005) dynamic panel random-effects Tobit model, which requires generally stronger assumptions than the extensive margin (5) but still offers useful insights. Specifically, we estimate

$$Q_{it} = \gamma + \gamma_Q Q_{i,t-1} + \gamma_{Q0} Q_{i,0} + \gamma_\Pi \Pi_{it} + \gamma_S S_{it} + \gamma_Z Z_{it} + \tilde{h}_i + \upsilon_{it} \quad (6)$$

in which $Q_{it} \geq 0$ is the quantity of fertilizer nutrients used (each of N and K), and Π_{it} is the corresponding planting-season price. Notation $\tilde{h}_i$ are time-invariant random effects, which, unlike time-invariant fixed effects, can be included in the Tobit regression (Wooldridge 2005). The term $Q_{i,0}$ is the function based on the quantity of each type of fertilizer nutrient (N, K) used in the initial survey years. Notation γ represents estimated parameters, and υ_{it} represents idiosyncratic errors.

A key assumption in Wooldridge's (2005) dynamic panel Tobit model is that the inclusion of $Q_{i,0}$ in the explanatory variable can provide a consistent estimate of the coefficient γ_Q, which is not required in (5). To strengthen the validity of this assumption in our analyses, we approximate $Q_{i,0}$ by the average values of Q_{it} over the first 3 years (2010–2012), and we focus the analyses on the last 3 years (2012–2014). $Q_{i,0}$ based on the average over the 3 initial years (2010–2012) is likely to be more stable and serves as a more reliable approximation of the initial condition than $Q_{i,0}$ based solely on the single initial year (2010), which can be more susceptible to year-specific idiosyncratic shocks. Past studies often used multiple-years pre-sample means of the dependent variable to approximate initial conditions in a similar dynamic panel data framework (e.g., Blazsek & Escribano 2016; Blundell et al. 1999, 2002; Máñez et al. 2020; Sanchis Llopis et al. 2024). Past studies also applied Wooldridge's (2005) dynamic panel models to three rounds of panel data (e.g., Alem & Demeke 2020; Gebru et al. 2019).

4 Descriptive statistics

4.1 General production characteristics

Table 1 and Table ***2*** summarize the descriptive statistics of production, input, and other relevant variables used in the analyses of DLD data and VDSA data, respectively. As Table 1 shows, during the 2002–2016 period, the district typically produced 7.3 million rupees (current rupee) of outputs annually for all the major commodities (described in Section 2) combined and cultivated about 261,000 ha of land, about half of which were irrigated. About 6.6 million rupees of agricultural credit were taken out in total to purchase various services and inputs (mostly machines, mechanization services, and agrochemicals). District-level use of N, P, and K are on average about 27.6, 11.2, and 4.5 million tons each year, respectively. Significant variations

occur across districts, with standard deviations typically in the same order of magnitude as the mean (i.e., coefficient of variation around 1).

As Table 2 shows, typical farm households in VDSA data produced 84,000 and 34,000 rupees' worth of agricultural products annually, at the mean and median, respectively, during the 2010–2014 period. Most households are smallholders who cultivated 1 to 2 hectares of land, using 300 to 400 person-hours of labor and purchasing inputs (other than fertilizer) worth 20,000 to 40,000 rupees. About 30 percent of the sample households used irrigation. The use of machines in terms of horsepower hours is more heterogeneous, with 257 and 12 horsepower-hours at the mean and median, respectively. In terms of exogenous characteristics, the household typically owned 1.25 hectares of land (suggesting that they leave some of the land fallow), had a household size of five, and owned agricultural equipment worth approximately 8,000 rupees at the median. Households faced wage and mechanization service fees of around 21 and 450 rupees per hour, respectively. Consistent with nationwide trends (Kishore et al. 2021), the price of N had remained generally stable, at around 6 to 8 rupees per kilogram, while prices of P and K increased significantly, from around 11 to 25 rupees per kilogram and around 5 or 6 to 18 rupees per kilogram, respectively.

4.2 Stationarity of DLD data variables

For reasonably long panel data like DLD data (with a span of 16 years), unit root in output and input variables can lead to spurious regression for (1). In the panel data context, the presence of stationarity for at least some panel districts j for all variables avoids such spurious regression. We test this using the Fisher panel data unit root test, which allows the unit-root coefficient to vary across panels (instead of imposing a common unit-root coefficient for all panels) and is suitable when the number of panels is finite (Baltagi 2013; Choi 2001; Hlouskova & Wagner 2006), such as ours, which consists of a reasonably small number of panels (496 districts).

Table 3 presents the results. Following Choi (2001), we use inverse normal Z statistics that offer the best trade-off between size and power, and we also control for the trend variable t. The p values indicate that, for each output and input variable, we reject the null hypothesis that all panels contain a unit root, suggesting that at least some panels are stationary.

5 Results

5.1 District-level evidence based on DLD data

Table 4 presents the LASSO-selected input parameters for the full translog production function (2) and shows how the set of key parameters narrows as the penalty weight (λ) on the number of variables increases. Table 5 presents the estimated coefficients of RLASSO-selected input variables, estimated through the Post-Double-Selection LASSO method, which allows causal interpretations of LASSO-selected input variables, as well as simple LASSO-selected regressions for different λs. At the penalization parameter $\lambda = 400$, a total of seven variables are selected. At $\lambda = 690.65$ (the level of λ selected by RLASSO), only three key variables remain, $N_{jt}A_{jt}$ $K_{jtc}A_{jt}$, and $A_{jt}W_{jtc}$, in which N, A, K, and W are the quantity of N, land, K, and capital, respectively. Importantly, we are *not* as interested in comparing the LASSO-selection of

variables across different λ.[6] Rather, we are interested in identifying which variables are consistently selected, particularly whether variables containing the lagged value of K(K_{jtc}) are selected more consistently than those containing only the non-lagged value of K (K_{jt}). Note that, as shown in Table 4, variables containing K_{jtc} ($K_{jtc}A_{jt}$, together with $K_{jtc}K_{jtc}$ and $K_{jtc}N_{jt}$) are always selected, regardless of the level of parsimony, while variables containing K_{jt} are not. Furthermore, at least one of these variables containing K_{jtc} remains consistently statistically significant, indicating its importance as a key predictor of production in the translog production function (2). Figure 1 illustrates how these variables ($N_{jt}A_{jt}$, $K_{jtc}A_{jt}$, and $A_{jt}W_{jtc}$, bold lines) remain significant as the models become more parsimonious through the RLASSO methods and less significant input variables are eliminated (set to 0). These results suggest that K_{jtc}, which includes a lagged quantity of K in $t-1$, is consistently a better determinant of production than $K_{jt,}$ which only includes K in t, within the flexible translog production function specification (2) for DLD data. Put differently, results are consistent with the argument that K has more cumulative effects in production that persist across years (at least beyond the current year).

In Table 5, bottom row, the sample average returns-to-scale based on RLASSO-selected results computed using Kim (1992) illustrate that our RLASSO-approximated results are still reasonable. Results indicate that the sample average returns-to-scale is generally around 1 (i.e., constant returns-to-scale), suggesting that our RLASSO-selected result is a reasonable approximation of a standard production function.

Figure 2 in Appendix B illustrates similar plots as Figure 1 for all inputs, contrasting results for variables consisting of X_{jt} and X_{jtc}, respectively. As discussed above, the plots for K (bottom row) confirm that variables consisting of X_{it0} persistently remain in the model, while variables consisting of X_{it} are excluded during the LASSO process. Interestingly, this pattern for K is in clear contrast to most other inputs, for which variables consisting of X_{it} generally remain more persistently in the model compared to variables consisting of X_{it0}. This pattern also holds for other major fertilizer nutrients like N, which is consistent with the argument that, unlike K, N has less-cumulative effects on production.

5.1.1 Robustness across heterogeneous production systems

Importantly, results for Table 5 describe the average conditions across India, which remains our primary interest due to its representativeness. Nonetheless, given significant heterogeneity, patterns in some subregions or subsystems may differ significantly or even be opposite in nature from the nationwide evidence presented in Table 5. We therefore provide further evidence for subsamples split across major agroecological conditions and cropping systems within India and replicate the LASSO estimation procedure to each subsample.

Soil types

Table C1 in Appendix C summarizes the results of subsample LASSO estimation, based on the subsamples split at the median values of key soil properties at the district level, namely soil organic contents, alkalinity (pH), and drainage levels. These factors have been found

[6] Nonetheless, the value of λ selected (690.65) by RLASSO suggests that the most parsimonious model in Table 4 and Table 5 satisfies certain desired properties in the LASSO procedure (see Appendix A).

particularly important in determining the effectiveness of certain fertilizers, such as K, in India (e.g., Wakeel & Ishfaq 2022). Results indicate that patterns are generally similar to those presented in Table 5 across districts with different soil organic contents and drainage properties, as well as districts with above-median soil alkalinity. Variables containing K_{jtc} are more consistently selected than those containing only the non-lagged value of K (K_{jt}). While no variable containing K_{jtc} is selected by LASSO in districts with below-median soil alkalinity, neither are the variables containing K_{jt}; in other words, evidence does not contradict Table 5, which shows that K_{jt} is not more influential than K_{jtc}.

Cropping systems

Table C2 summarizes a similar subsample LASSO estimation based on average cropping systems during the study period: (1) grain production districts (sample districts with higher area shares of grains, above the sample median) and other districts; (2) districts with higher-than-medium area shares of pulses/legumes and other areas; and (3) districts with higher-than-medium area shares of fruits/vegetables combined and other areas. We again observe generally robust patterns consistent with our main findings in Table 5, namely that variables consisting of cumulative K inputs are often selected as the most influential. For example, $K_{jtc}A_{jt}$ in districts with lower grain area shares and higher fruit/vegetable area shares, as well as in districts with both higher and lower pulse/legume area shares, are often selected. Similarly, $K_{jtc}N_{jt}$ in districts with higher grain area shares and lower fruit/vegetable area shares is also frequently selected, while variables consisting of current-year K inputs (K_{jt}) are not. These results suggest that the importance of K's cumulative effects, as shown in Table 5, holds broadly across diverse cropping systems.

5.2 Farm household–level evidence (VDSA data based on GMM-LASSO)

Similarly, Table 6 summarizes the panel GMM regression results for the translog production function (2) using GMM-LASSO selected variables based on AIC and BIC. As expected, GMM-LASSO based on BIC selected fewer parameters than that based on AIC, which is consistent with the application of AIC and BIC in other general contexts. Similar to the case with DLD data (Table 4 and Table ***5***), it is noteworthy that more variables containing K_{itc} tend to remain as influential as variables consisting of K_{it}. Furthermore, diagnostic statistics indicate that, for the models shown in Table 6, IVs provide sufficient identification power (i.e., the null hypothesis of under-identification is rejected) but also do not cause overidentification, suggesting that results are unbiased and consistent from the GMM standpoint.

As was described in 3.2.2, we also check the robustness of results when using $\ln(X + 0.001)$ instead of $\ln(X + 1)$, with X being the corresponding input quantities including K (Table C3 in Appendix C). Results are largely consistent with those shown in Table 6.

5.3 Dynamics of potassium fertilizer use

The significant cumulative effects of certain fertilizer nutrients like K compared to other major nutrients like N, as shown in Table 5 and Table 6, can have major implications for the demand for these nutrients. Table 7 and Table ***8*** show indicative evidence based on Equations (5) and (6).

5.3.1 Arellano and Bond (1991) dynamic panel linear probability model

Table 7 shows the results of Equations (5) based on the Arellano and Bond (1991) dynamic panel LPM. For both N and K, we present the results of one-step GMM (which is less efficient but more robust) and two-step GMM (which is more efficient, assuming the estimated weighting matrix is correct). The rows for Arellano and Bond autocorrelation tests indicate the absence of autocorrelation with the third lag. We therefore use $Q_{i,t-3}$ and earlier lags as additional IVs. The bottom row indicates that using all lags of Q_{it} up to $Q_{i,t-3}$ satisfies the orthogonality conditions of all IVs.

The results indicate that the likelihood of using K in t is significantly negatively affected by $t-1$ under both one- and two-step GMM methods, consistent with the findings in Table 5 and Table ***6***. It is the aggregate quantity of K in t and $t-1$ that influences production, suggesting that, given the declining marginal returns and conditional on other exogenous shocks, quantities in t and $t-1$ may substitute for each other. Importantly, these results do not hold for N, in which the use in $t-1$ does not significantly affect the likelihood of use in t. The contrasting results for N are again consistent with the findings in Table 5 and Table ***6***, where cumulative effects are generally absent for N, unlike for K.

5.3.2 Wooldridge (2005) panel dynamic Tobit model

Table 8 summarizes the results of Wooldridge's (2005) panel dynamic Tobit equation (6), estimated separately for N and K. To highlight the implications for the demand response to price changes, particularly its lagged nature, we also estimate additional models using the average price over the last 4 years (between *t* and *t* – 3).

Similar to Table 7, results in Table 8 suggest differential patterns consistent with those in Table 5 and Table ***6***. Specifically, we find that the quantity of K used in current year (*t*) is statistically significantly negatively affected by the quantity used in the previous year (*t* – 1). This negative relationship indicates that K in the previous and current years are partly substitutes, consistent with the previously mentioned findings of significant cumulative effects for K. Furthermore, such dynamic relations are statistically insignificant for N, which again aligns with the earlier findings of the absence of cumulative effects for N.

Furthermore, partly due to the significantly strong dynamic effects, when its intensive margin is also considered, K use is generally inelastic with respect to its price, not only in the same year but also in averages including the previous few years. Again, these patterns differ for N, whose quantity used is more responsive to price in both the current year *t* and average prices from the previous few years. While Table 8 refers only to the VDSA samples and may not be readily generalizable to India as a whole, the weak or slow responses of K use to its price are consistent with the observations in Kishore et al. (2021), based on the national trends of relatively stable K consumption following the price increase since 2011.

6 Discussions

The results discussed in the previous section have important implications across various aspects.

6.1 Parsimonious characterization of complex underlying production function

Many previous agronomic studies using the LASSO approach focused mainly on analyzing the yield determinants (Mourtzinis et al., 2018; Mondal et al., 2020; Nayak et al., 2022b; Goncharov et al., 2023; Arshad et al., 2023). While focusing on yields allows for a potentially more delicate analysis of yield determinants, applying LASSO to the flexible translog production function, as we have done, provides unique but also complementary insights on previous studies using LASSO. This is particularly true in contexts where changes in land use from year to year can be significant due to factors including worker availability, which can vary across different years, among other considerations. Similarly, extending the LASSO beyond yields is important when there are complex interactive effects between different production factors (either complementary or substitutive). Our results underscore this point. The LASSO approach suggests that a relatively parsimonious specification (with a limited number of key inputs with high predicting power) may be identified, potentially due to substantial multicollinearity among production inputs, as farmers may use various inputs in relatively constant proportions.

6.2 Cumulative effects of potassium in nonexperimental settings

Our results suggest that K plays significant roles, particularly through its cumulative effects, in agricultural production at the field level under farmers' actual production practices. These findings complement rich evidence in experimental agronomic settings and shed light on the relative importance of different agricultural inputs and their potential cumulative effects in nonexperimental settings. These settings may more closely reflect the production environments in which average farmers operate, considering their specific agroecological conditions and resource endowments. Importantly, by aggregating across multiple crops, our analyses also offer a more representative picture of the agronomic characteristics in India and at the farm household level.

Our findings have important implications not only at the single crop level but also for the overall agriculture sector and all crop production activities within the household. It is also important to note that the cumulative effects of K broadly hold across the different levels of input use. As shown, significant cumulative effects of K are identified within the translog production function framework, which explicitly accounts for the use of other production factors, as well as the complex interactive effects among these other production factors on output. This suggests that the identified cumulative effects of K are less likely to be due to the failure of the model to correctly separate their effects from the other inputs that happen to be highly correlated with K use, which can sometimes be difficult to differentiate in nonexperimental settings. Furthermore, the fact that the cumulative effects of K remain significant in the parsimonious representation of production function by LASSO suggests that these effects are substantial, with implications for overall production, even when compared to the other production inputs.

6.3 Consistent importance of the cumulative effects of K across agroecology and cropping systems despite heterogeneity

The results also suggest that the relative importance of K, as well as its cumulative effects in crop production, may hold relatively consistently across diverse agroecology (based on soil characteristics) and cropping systems within India. In other words, earlier agronomic results

indicating these cumulative effects of K may be relevant across a wider geographic range than previously thought.

At the same time, the set of other inputs found among LASSO-selected K variables varies across these ecologies and cropping systems, suggesting that the marginal effects of K can still vary, depending on the use of other inputs. For example, the significantly positive coefficients for $\boldsymbol{K_{jtc}A_{jt}}$ in the most parsimonious models in Table 5 suggest that nationwide, the output elasticity with respect to cumulative K use is larger (and positive) if more land is used in production. However, within the specific VDSA samples studied (Table 6), the coefficients for $\boldsymbol{K_{itc}A_{it}}$ are significantly negative, suggesting that the output elasticity with respect to cumulative K use is larger (and positive) if less land is used in production. These differences suggest that the joint effects of cumulative K use and land can vary significantly at the local levels.

Similarly, in Table C2, the output elasticity of cumulative K use varies more in response to the size of the land used in areas with higher shares under fruits/vegetables and lower shares under grains, as $\boldsymbol{K_{jtc}A_{jt}}$ is selected by LASSO as the key predictor of output. However, the same elasticity varies more in response to N quantity applied in areas with more grains and less fruits/vegetables, as $K_{jtc}N_{jt}$ is selected by LASSO. These differences suggest that the joint effects of cumulative K and other inputs, such as land or N, can be significantly heterogeneous across cropping systems.

Importantly, however, regardless of the direction of these interactive effects, outputs remain more responsive to cumulative K than to the current year's application of K alone. This consistency in the relative importance of cumulative K across a broad production environment within India is the central significance of our findings.

6.4 Implications of dynamic patterns of potassium use by farmers

Our analyses of the dynamic patterns of K application by farmers further suggest that the identified cumulative effects of K are also causing the current year's K application to depend significantly on the previous year's application. The relative importance of K and its cumulative effects is significant enough to affect farmers' actual K use in dynamic ways. Farmers may learn of these cumulative effects from extension services or their own experiences.

This pattern offers insights into how agronomic technological characteristics translate into actual agricultural outputs. Specifically, compared to the scenarios *without* farmers' adjustments in K use, agricultural production in India is more stabilized between the current year and the prior year. This is because of the following mechanisms: Cumulative K use for the previous year and the current year overall has positive effects on outputs; farmers who used more K in the prior year use less K in the current year, leading to more stable cumulative K inputs and thus more stable outputs across years.

6.5 Policy implications for potassium use in agricultural production growth in India

Beyond its cumulative effects, the generally significant predictive power of K in a production function, relative to other inputs, supports the hypothesis that relatively low K use (compared to N and P) in India, as described in the background section, is one of the key constraints to agricultural production in the country.

Our results have important policy implications. First, promoting increased K use could increase agricultural production in India, given K's significant roles in production, as well as the relatively low levels of current K application in India compared to other nutrients and the rest of the world. Second, in developing improved fertilizer dosage recommendations for farmers—an essential aspect of fertilizer policies in South Asia (e.g., Beg et al., 2024)—the potential cumulative effects of K should be more carefully examined at specific localities and incorporated into these recommendations. In doing so, monitoring and keeping a record of previous years' K use (as in DLD data and perhaps even at lower administrative levels) remains crucial, as the recommended dosage may depend on previous use.

Third and last, regarding price policies, price stabilization efforts by the Indian government may be more justified for certain types of fertilizer, such as Muriate of Potash (MOP)—the major K fertilizer in India—compared to non-K fertilizers like urea or DAP. This is because, given the inelastic demand for K (compared to N fertilizer) in India, price fluctuations can lead to higher farm income fluctuation, while inelastic demand can also cause more significant price fluctuations locally. Expanding subsidies for MOP, if necessary, may also be relatively more justified than doing so for urea or DAP, as current effective subsidy rates for MOP are lower. For example, Chakraborty et al. (2024) indicate that in 2022, the de facto subsidy rate for MOP was 29 percent, compared to 89 percent for urea and 61 percent for DAP.

7 Conclusions

The cumulative effects of inputs on production, including the effects of inputs in one year on production in subsequent years, have been widely recognized. Nonetheless, unlike sectors outside agriculture (e.g., cumulative effects of education on human capital formation) or public-good inputs within the agriculture sector, evidence on the cumulative effects of private-good inputs in agriculture has generally remained anecdotal or is confined to agronomic trial settings, and direct field-level evidence has remained scarce. Addressing this knowledge gap is important because, for example, cumulative input effects can affect optimal production advice provided by agricultural extension efforts or the demand for these inputs in a particular year and thus the effectiveness of policies such as subsidies. Addressing this knowledge gap has been challenging, particularly within the production function framework, due to the general scarcity of long-term panel data on agricultural production, significant multicollinearity among inputs, and variations in input quantities across years.

We attempted to narrow this knowledge gap by applying LASSO and GMM-LASSO methods to estimate the flexible translog production function using unique datasets of district-level panel data (DLD data) and farm household-level panel data (VDSA data) in India. Our results show that certain inputs, particularly K, remain among the most persistent key predictors of production levels, even when compared to other major inputs like land, labor, capital, irrigation, and other fertilizer nutrients. More importantly, the aggregate quantity of K from the current and previous years is consistently a more important determinant of production than the quantity of K from the current year alone, demonstrating the significant cumulative effects of K. These patterns do not hold for N and most other inputs. These results hold for both the district level (DLD data) and the farm household level (VDSA data). The dynamic panel LPM and the dynamic panel Tobit analyses further suggest that K used in the current year is significantly negatively affected by the quantity used in the previous year, while no such dynamics were observed for N, consistent with production function results.

References

Ahrens A, CB Hansen & ME Schaffer. 2020. lassopack: Model selection and prediction with regularized regression in Stata. *Stata Journal 20*(1):176-235.

Alem Y & E Demeke. 2020. The persistence of energy poverty: A dynamic probit analysis. *Energy Economics 90*, 104789.

Arellano M & S Bond. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economics Studies 58*:277-297.

Arshad S, JH Kazmi, FA Prodhan & S Mohammed. 2023. Exploring dynamic response of agrometeorological droughts towards winter wheat yield loss risk using machine learning approach at a regional scale in Pakistan. *Field Crops Research 302*:109057.

Askegaard M, J Eriksen & AE Johnston. 2004. *Sustainable management of potassium*. In *Managing soil quality: challenges in modern agriculture.* (pp. 85-102). Wallingford UK: CABI Publishing.

Baltagi B. 2013. *Econometric Analysis of Panel Data. 5th ed.* Chichester, UK: Wiley.

Beg S, M Islam & KW Rahman. 2024. Information and behavior: Evidence from fertilizer quantity recommendations in Bangladesh. *Journal of Development Economics 166*:103195.

Berger T. 2001. Agent-based spatial models applied to agriculture: a simulation tool for technology diffusion, resource use changes and policy analysis. *Agricultural Economics 25*(2-3):245-260.

Bernal R. 2008. The effect of maternal employment and child care on children's cognitive development. *International Economic Review 49*(4):1173-1209.

Blazsek S & A Escribano. 2016. Score-driven dynamic patent count panel data models. *Economics Letters 149*:116-119.

Blundell R, R Griffith & J Van Reenen. 1999. Market share, market value and innovation in a panel of British manufacturing firms. *Review of Econometric Studies 66*(3):529-554.

Blundell R, R Griffith & F Windmeijer. 2002. Individual Effect and Dynamics in Count Data Model. *Journal of Econometrics 108*:113-131.

Caner M. 2009. Lasso-type GMM estimator. *Econometric Theory 25*(01):270-290.

Chakraborty P, A Chopra & L Contractor. 2024. *The Equilibrium Impact of Agricultural Support Prices and Input Subsidies*. Economics Discussion Paper 123. Ashoka University, India.

Chand R. 1994. Economics of perennial crops: some methodological issues. *Indian Journal of Agricultural Economics 49*(2):246-249.

Choi I. 2001. Unit root tests for panel data. *Journal of International Money and Finance 20*:249-272.

Fan S & PG Pardey. 1997. Research, productivity, and output growth in Chinese agriculture. *Journal of Development Economics 53*(1):115-137.

FAO (Food and Agriculture Organization) /IIASA (International Institute for Applied Systems Analysis) /ISRIC (International Soil Reference and Information Centre) /ISSCAS (Institute of Soil Science – Chinese Academy of Sciences) /JRC (Joint Research Centre of the European Commission). 2012. *Harmonized World Soil Database (version 1.2)*.

Rome: FAO; Laxenburg, Austria: IIASA. http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/.
Food and Agriculture Organization (FAO). 2024. *FAOSTAT*. Rome, Italy.
Fuglie K, B Dhehibi, AAI El Shahat & A Aw-Hassan. 2021. Water, policy, and productivity in Egyptian agriculture. *American Journal of Agricultural Economics 103*(4):1378-1397.
Funk C, P Peterson, M Landsfeld, D Pedreros, J Verdin, S Shukla, G Husak, J Rowland, L Harrison, A Hoell & J Michaelsen. 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data 2*(1):1-21.
Gebru M, ST Holden & M Tilahun. 2019. Tenants' land access in the rental market: evidence from northern Ethiopia. *Agricultural Economics 50*(3):291-302.
Goncharov AA, TA Safonov, AM Malko, GA Bocharov & SV Goncharov. 2023. Climate change expected to increase yield of spring cereals and reduce yield of winter cereals in the Western Siberian grain belt. *Field Crops Research 302*:109038.
Gong B. 2018. Agricultural reforms and production in China: Changes in provincial production function and productivity in 1978–2015. *Journal of Development Economics 132*:18-31.
Haque T & A Goyal. 2021. Access to institutional credit by farmers in Eastern India. *Journal of Asian Development Research* 2633190X211040622.
Hendricks NP, A Smith & DA Sumner. 2014. Crop supply dynamics and the illusion of partial adjustment. *American Journal of Agricultural Economics 96*(5):1469-1491.
Hlouskova J & M Wagner. 2006. The performance of panel unit root and stationarity tests: Results from a large scale simulation study. *Econometric Reviews 25*:85-116.
ICRISAT (International Crops Research Institute for the Semi-Arid Tropics). 2024. *Village Dynamics in South Asia Dataset 2010–2014.* Hyderabad, India.
ICRISAT & TCI. 2024. *District-Level Database for Indian Agriculture and Allied Sectors*. Available at http://data.icrisat.org/dld/src/about-dld.html. Accessed on April 1, 2024.
Jacoby HG. 1993. Shadow wages and peasant Family Labour Supply: An Econometric Application to the Peruvian Sierra. *Review of Economic Studies 60*(4):903-21.
Johnston AE. 1986. *Potassium fertilization to maintain a K-balance under various farming systems*. In *Nutrient balances and the need for potassium*, p.199-226.
Kerr WR, WF Lincoln & P Mishra. 2014. The dynamics of firm lobbying. *American Economic Journal: Economic Policy 6*(4):343-379.
Kihara J et al. 2016. Maize response to macronutrients and potential for profitability in sub-Saharan Africa. *Nutrient Cycling in Agroecosystems 105*(3):171-181.
Kim HY. 1992. The translog production function and variable returns to scale. *Review of Economics and Statistics 74*(3):546-552.
Kishore A, M Alvi & TJ Krupnik. 2021. Development of balanced nutrient management innovations in South Asia: perspectives from Bangladesh, India, Nepal, and Sri Lanka. *Global Food Security 28*:100464.
López RA & N Yadav. 2010. Imports of intermediate inputs and spillover effects: Evidence from Chilean plants. *Journal of Development Studies 46*(8):1385-1403.
Máñez JA, ME Rochina-Barrachina & JA Sanchis. 2020. Foreign sourcing and exporting. *The World Economy 43*(5):1151-1187.
Meena BP, AK Biswas, M Singh, RS Chaudhary, AB Singh, H Das & AK Patra. 2019. Long-term sustaining crop productivity and soil health in maize–chickpea system through integrated nutrient management practices in Vertisols of central India. *Field Crops Research 232*:62-76.

Meng M, L Yu & X Yu. 2024. Machinery structure, machinery subsidies, and agricultural productivity: Evidence from China. *Agricultural Economics 55*(2):223-246.
Mondal S, S Dutta, L Crespo-Herrera, J Huerta-Espino, HJ Braun & RP Singh. 2020. Fifty years of semi-dwarf spring wheat breeding at CIMMYT: Grain yield progress in optimum, drought and heat stress environments. *Field Crops Research 250*, 107757.
Moser CM & CB Barrett. 2006. The complex dynamics of smallholder technology adoption: the case of SRI in Madagascar. *Agric. Econ. 35*(3):373-388.
Mourtzinis S, JIR Edreira, P Grassini, AC Roth, SN Casteel, IA Ciampitti & SP Conley. 2018. Sifting and winnowing: Analysis of farmer field data for soybean in the US North-Central region. *Field Crops Research 221*:130-141.
Nayak HS, JV Silva, CM Parihar, TJ Krupnik, DR Sena, SK Kakraliya & TB Sapkota. 2022a. Interpretable machine learning methods to explain on-farm yield variability of high productivity wheat in Northwest India. *Field Crops Research 287*:108640.
Nayak HS, JV Silva, CM Parihar, SK Kakraliya, TJ Krupnik, D Bijarniya & TB Sapkota. 2022b. Rice yield gaps and nitrogen-use efficiency in the Northwestern Indo-Gangetic Plains of India: Evidence based insights from heterogeneous farmers' practices. *Field Crops Research, 275*:108328.
Nickell SJ. 1981. Biases in dynamic models with fixed effects. *Econometrica 49*:1417-1426.
National Oceanic and Atmospheric Administration (NOAA). 2022. *Version 4 DMSP-OLS Nighttime Lights Time Series*. https://www.ngdc.noaa.gov/eog/dmsp/downloadV4composites.html. Accessed on May 1, 2024.
Qiao L, JV Silva, M Fan, I Mehmood, J Fan, R Li & MK van Ittersum. 2021. Assessing the contribution of nitrogen fertilizer and soil quality to yield gaps: A study for irrigated and rainfed maize in China. *Field Crops Research 273*:108304
Quilty JR, J McKinley, VO Pede, RJ Buresh, TQ Correa Jr & JM Sandro. 2014. Energy efficiency of rice production in farmers' fields and intensively cropped research fields in the Philippines. *Field Crops Research 168*:8-18.
Rashid S, PA Dorosh, M Malek & S Lenma. 2013. Modern input promotion in sub-Saharan Africa: insights from Asian green revolution. *Agricultural Economics 44*(6):705-721.
Sanchis Llopis JA, JA Mañez & AM Gómez-Sánchez. 2024. The dynamic links between product and process innovations and productivity for Colombian manufacturing. *Applied Economic Analysis 32*(94):62-82.
Shi Z. 2016. Estimation of sparse structural parameters with many endogenous variables. *Econometric Reviews 35*(8-10):1582-1608.
Takeshima H. 2017. Custom-hired tractor services and returns to scale in smallholder agriculture: A production function approach. *Agricultural Economics 48*(3):363-372.
Takeshima H, R Adhikari, BD Kaphle, S Shivakoti & A Kumar. 2017. Heterogeneous returns to chemical fertilizer at the intensive margins: Insights from Nepal. *Food Policy 69*:97-109.
Takeshima H, I Masias, MT Win & PP Zone. 2023. Effects of COVID-19 restrictions on mechanization service providers and mechanization equipment retailers: Insights from phone surveys in Myanmar. *Review of Development Economics 27*(1):323-351.
Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B58*:267-288.

Vicente-Serrano SM, S Beguería & JI López-Moreno. 2010. A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *Journal of Climate 23*(7):1696-1718.

Wakeel A & M Ishfaq. 2022. *Potash use and dynamics in agriculture.* Springer.

Wang N, Z Ai, Q Zhang, P Leng, Y Qiao, Z Li & F Li. 2024. Impacts of nitrogen (N), phosphorus (P), and potassium (K) fertilizers on maize yields, nutrient use efficiency, and soil nutrient balance: Insights from a long-term diverse NPK omission experiment in the North China Plain. *Field Crops Research 318*:109616.

Williams CB. 1937. The use of logarithms in the interpretation of certain entomological problems. *Annals of Applied Biology 24*:404-414.

Wooldridge JM. 2005. Simple solutions to the initial conditions problem in dynamic nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics 20*:39-54.

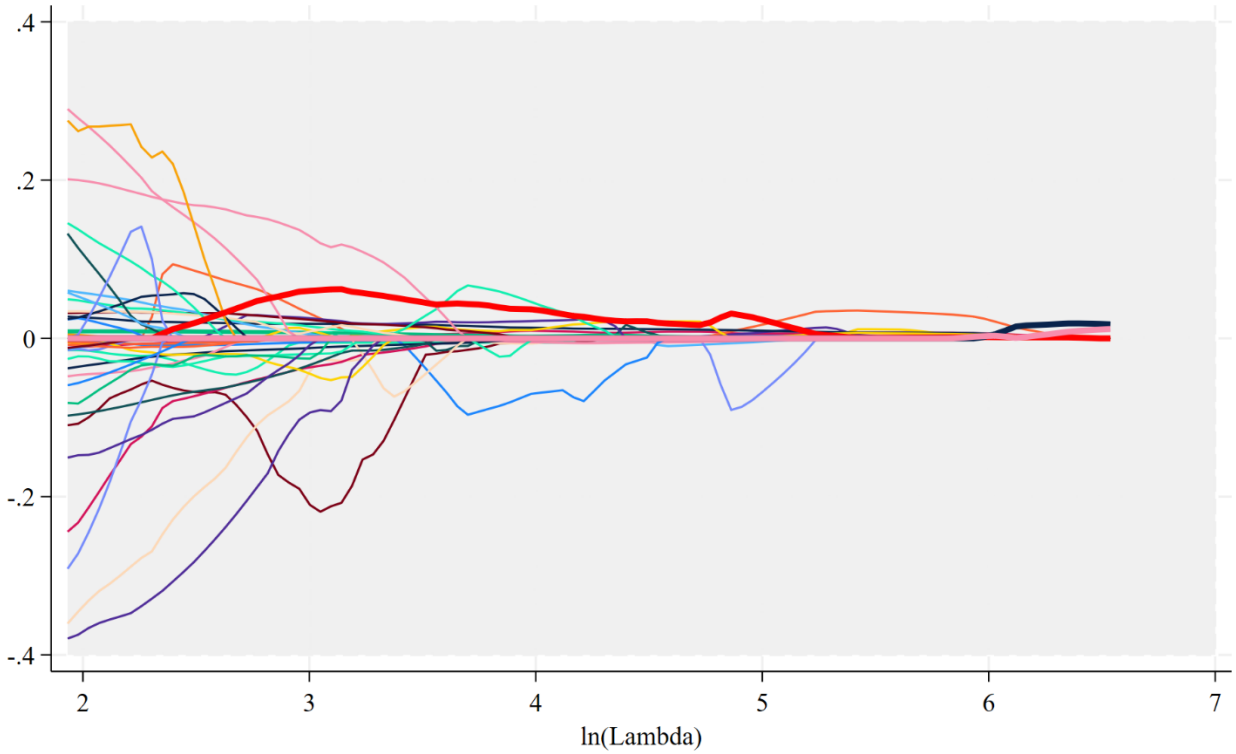

Figure 1. Illustration of coefficient estimates evolution in LASSO parameter selection process[a]

Source: Authors.

[a]Vertical axis indicates the standardized coefficients of each variable at given level of penalization parameter λ. Bold lines indicate the 3 RLASSO selected parameters in ***Table 5***

Table 1. Descriptive statistics of DLD data

Variables	Mean	Std.dev
Production revenue (current million Rupee)	7.268	8.804
Land (1,000 ha)	261.090	213.434
Agricultural population (1,000)	483.367	305.931
Agricultural credit (current million Rupee)	6.644	11.059
Share of land irrigated	0.489	0.317
Nitrogen used (1,000 ton)	27.555	26.467
Phosphorus used (1,000 ton)	11.165	11.646
Potassium used (1,000 ton)	4.538	7.047
Nighttime light (district average, nanoWatt per cm^2 per $radian^2$)	5.543	7.202
Absolute rainfall deviations (absolute value of z-score)	0.793	0.569
Absolute drought index deviations (absolute value of z-score)	0.806	0.609
Number of sample districts	496	
Total sample size	7,054	

Source: Authors' compilation from DLD data.

Table 2. Descriptive statistics of VDSA data (annual total)

Variables		Mean	Median	Std.dev
Production revenue (current 1,000 Rupee)		83.914	34.436	168.895
Land cultivated (ha)		2.104	1.113	3.540
Labor use (person-hours)		408.176	281.000	444.823
Use of own machines (horsepower-hour)		257.495	12.000	1041.677
Purchased inputs excl. fertilizer (1,000 Rupee)		37.106	18.181	62.716
Use irrigation (yes = 1)		0.299	0.000	0.458
Nitrogen used (kg)		98.604	39.000	196.314
Phosphorus used (kg)		67.943	25.760	135.792
Potassium used (kg)		12.013	0.000	38.154
Land owned (ha)		3.196	1.250	5.771
Household size		5.530	5.000	2.650
Agricultural equipment owned (1000 rupees)		48.674	7.800	176.073
Wage (current rupee per hour)		22.252	21.124	9.486
Mechanization service fees (rupees per hour)		451.566	450.000	127.625
Price – Nitrogen (current rupee per kg)		7.472	7.000	2.402
	2010	6.438	6.250	
	2011	8.055	7.000	
	2012	7.913	7.527	
	2013	7.467	7.280	
	2014	7.445	7.125	
Price – Phosphorus (current rupee per kg)		20.905	23.500	7.823
	2010	11.737	11.000	
	2011	16.502	16.350	
	2012	25.062	23.500	
	2013	25.281	25.000	
	2014	25.781	25.000	
Price – Potassium (current rupee per kg)		13.535	14.500	5.282
	2010	6.653	5.250	
	2011	8.796	8.500	
	2012	15.809	14.000	
	2013	18.394	18.000	
	2014	17.938	17.508	
Total sample size		3,104		

Source: Authors' compilation from VDSA data.

Table 3. Fisher panel data tests for the unit root of DLD data variables[a, b]

Variables	Inverse normal Z Statistics[a]	p-value (H_0: Unit root for all panels)
Ln(Production revenue)	–17.328	.000
Ln(Land)	–14.073	.000
Ln(Agricultural population)	–9.865	.000
Ln(Agricultural credit)	–13.559	.000
Share of land irrigated	–22.064	.000
Ln(Nitrogen used)	–17.417	.000
Ln(Phosphorus used)	–11.318	.000
Ln(Potassium used)	–11.526	.000

Source: Authors.

[a] Inverse normal Z statistics offers the best tradeoff between size and power, and recommended by Choi (2001).

Table 4. Set of LASSO-selected parameters (DLD data)

LASSO penalization parameter (λ)	400	500	600	690.65 (RLASSO-selected λ)
LASSO-selected parameters of translog function[a]	$N_{jt}A_{jt}$ $K_{jtc}A_{jt}$ $A_{jt}W_{jtc}$ $K_{jtc}K_{jtc}$ $A_{jt}A_{jt}$ $K_{jtc}N_{jt}$ $N_{jt}N_{jt}$	$N_{jt}A_{jt}$ $K_{jtc}A_{jt}$ $A_{jt}W_{jtc}$ $K_{jtc}K_{jtc}$ $A_{jt}A_{jt}$ $K_{jtc}N_{jt}$	$N_{jt}A_{jt}$ $K_{jtc}A_{jt}$ $A_{jt}W_{jtc}$ $K_{jtc}K_{jtc}$	$N_{jt}A_{jt}$ $K_{jtc}A_{jt}$ $A_{jt}W_{jtc}$

Source: Authors based on LASSO estimation.

A = land, W = capital, N = nitrogen, K = potassium.

Table 5. Post Double-Selection (PDS) LASSO-approximation of translog production function for district level data (DLD data) by different levels of penalization parameter λ[a, b]

LASSO selected variables	λ= 200	λ= 300	λ= 600	λ= 690.65 selected by RLASSO	λ= 690.65 selected by RLASSO Post Double-Selection LASSO
	Coef. (std.err)	Coef. (std.err)	Coef. (std.err)	Coef. (std.err)	Coef. (std.err)
$N_{jt}A_{jt}$	–0.098***	0.004	0.025***	0.025***	0.027
	(0.033)	(0.014)	(0.008)	(0.008)	(0.043)
$\boldsymbol{K_{jtc}A_{jt}}$	**0.042***	**–0.030***	**–0.007**	**0.024*****	**0.035*****
	(0.024)	**(0.017)**	**(0.013)**	**(0.005)**	**(0.008)**
$A_{jt}W_{jtc}$	0.014	0.011	0.022***	0.014**	0.008
	(0.010)	(0.010)	(0.008)	(0.007)	(0.026)
$\boldsymbol{K_{jtc}K_{jtc}}$	**0.022****	**0.012**	**0.015*****		
	(0.009)	**(0.008)**	**(0.006)**		
$A_{jt}A_{jt}$	0.098***	0.062**			
	(0.029)	(0.027)			
$\boldsymbol{K_{jtc}N_{jt}}$	**–0.037***	**0.017**			
	(0.019)	**(0.012)**			
$N_{jt}N_{jt}$	0.052***				
	(0.015)				
Nighttime light	Included	Included	Included	Included	Included
Weather shocks	Included	Included	Included	Included	Included
Year * State	Included	Included	Included	Included	Included
dummies	Included	Included	Included	Included	Included
District fixed effects	Included	Included	Included	Included	Included
Intercept	Included	Included	Included	Included	Included
Samples	7,054	7,054	7,054	7,054	7,054
Average returns-to-scale				0.960	1.038

Source: Authors. Asterisks indicate statistical significance: ***1% **5% *10%.

[a] A = land, W = capital, R = irrigation, N = nitrogen, P = phosphorus, K = potassium.

[b] Bold texts refer to coefficients that contain $\boldsymbol{K_{jtc}}$.

Table 6. GMM-LASSO-approximation of translog production function for farm household level data (VDSA data)

LASSO-selected inputs[a]	GMM-LASSO selection based on AIC	GMM-LASSO selection based on BIC
	Coef. (std.err)	Coef. (std.err)
$\boldsymbol{K_{itc}}$	**0.214*** **	**0.228*** **
	(0.078)	**(0.066)**
$\boldsymbol{K_{itc}A_{it}}$	**–0.177** **	**–0.243*** **
	(0.088)	**(0.062)**
$E_{it}E_{it}$	0.527***	0.492***
	(0.096)	(0.067)
A_{it}	0.162*	0.227***
	(0.094)	(0.078)
$W_{it}W_{it}$	0.066*	0.080**
	(0.039)	(0.038)
R_{it}	0.091*	0.109**
	(0.054)	(0.051)
$E_{it}L_{it}$	0.055	
	(0.091)	
$P_{itc}P_{itc}$	–0.046	
	(0.061)	
$A_{it}N_{it}$	–0.020	
	(0.087)	
Year * Village dummies	Included	Included
Farm household fixed effects	Included	Included
Intercept	Included	Included
Samples	3041	3041
p-value (H_0: under-identified)	.006	.000
p-value (H_0: not over-identified)	.296	.376
Information Criteria Statistics of GMM-LASSO selection as shown in Appendix A, equation (9)	AIC = 34.807	BIC = 139.124

Source: Authors. ***1% **5% *10%.

[a] A = land, W = capital, L = labor, E = other expenditures, R = irrigation, N = nitrogen, P = phosphorus, K = potassium.

AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.

Table 7. Dynamics of fertilizer nutrient use decision at extensive margins (Arellano & Bond 1991) dynamic panel linear probability model)

Variables	Nitrogen		Potassium	
	One-step GMM Coef. (std.err)	Two-step GMM Coef. (std.err)	One-step GMM Coef. (std.err)	Two-step GMM Coef. (std.err)
Lagged dependent variable (quantity at $t-1$)	–0.055 (0.230)	–0.241 (0.226)	–0.597** (0.249)	–0.670** (0.265)
Price at t	–0.006** (0.003)	–0.008*** (0.002)	–0.063** (0.031)	–0.069** (0.032)
Year dummies	Included	Included	Included	Included
Other variables Z_{it}	Included	Included	Included	Included
Intercept	Included	Included	Included	Included
Samples	2,944	2,944	2,944	2,944
p-value (H_0: no serial correlation at third-lag)	.740	.985	.297	.371
Most recent lag of dependent variable used as an IV	$t-3$	$t-3$	$t-3$	$t-3$
Number of IVs	18	18	18	18
p-value (H_0: no overidentification)	.356	.356	.397	.397

Source: Authors. ***1% **5% *10%.

Table 8. Dynamics of fertilizer nutrients use at the farm household level – Wooldridge (2005) random-effects dynamic panel Tobit regression (6)

Variables	Nitrogen (in 100kg)		Potassium (in 100kg)	
	Coef. (std.err)	Coef. (std.err)	Coef. (std.err)	Coef. (std.err)
Lagged dependent variable (quantity at $t-1$)	–0.065 (0.054)	–0.078 (0.050)	–0.556*** (0.051)	–0.558*** (0.050)
Price[a] at t	–0.445* (0.247)		–0.236 (0.646)	
Price (average between t and $t-3$)		–1.851*** (0.313)		–0.314 (0.250)
Initial period value of dependent variable	Included	Included	Included	Included
Year * Village dummies	Included	Included	Included	Included
Other variables Z_{it}	Included	Included	Included	Included
Time-invariant random effects	Included	Included	Included	Included
Intercept	Included	Included	Included	Included
Samples	1,172	1,172	1,172	1,172
p-value (H_0: coefficients are jointly insignificant)	.000	.000	.000	.000

Source: Authors. ***1% **5% *10%.

[a]Price of respective fertilizer nutrients.

Appendix A: LASSO and GMM-LASSO

The literature has discussed various approaches for selecting the penalization parameter λ (Ahrens et al. 2020). These include (a) Information Criteria (Zou et al. 2007; Zhang et al. 2010), (b) K-fold cross-validation and h-step-ahead rolling cross-validation approach (Geisser 1975; Arlot & Celisse 2010), and (c) theory-driven ("rigorous") LASSO (RLASSO) approach (Belloni et al. 2012, 2014, 2016).

Among these methods, we use RLASSO approach which is considered more suitable for causal inference rather than forecasting, and is shown to perform well in terms of type-I errors (mistakenly identifying parameters to be non-zero when they are actually zero ("false positive")) (Ahrens et al. 2020). Unlike other approaches, RLASSO selects λ which, based on theoretical grounding, guarantees that optimal rates of convergence for parameter estimation are achieved, resulting size of the models has the same order as the true model, and selected parameters are consistently estimated (Ahrens et al. 2020). Specifically, λ in RLASSO is selected as (Jing et al. 2003; Belloni et al. 2012),

$$\lambda = 2\theta\sigma\sqrt{n}\Phi^{-1}\{1 - \gamma/(2p)\} \tag{7}$$

in which γ is the probability that "false positive" does not occur, $\theta(> 1)$ is a constant slack parameter which is set at slightly greater than 1 (often 1.1 as in Ahrens (2020)), σ is the standard errors of residual ε_{jt} from (2).

GMM-LASSO

GMM-LASSO method proposed by Shi (2016) is essentially an extension of standard LASSO (4) to GMM framework. Specifically, GMM-LASSO can be expressed as

$$\widehat{\beta_X}(\lambda) = argmin\frac{1}{n^2}M + \lambda\sum_{X}^{p}|\beta_X|$$
$$M = E'ZWZ'E \tag{8}$$

in which M is the standard GMM criterion function, E is a $n \times 1$ vector with ε_{it} as elements, Z is $n \times Z_n$ matrix of IVs (Z_n is the number of IVs), and W is a suitable $Z_n \times Z_n$ weighting matrix estimated in GMM.

Based on Shi (2016), under the GMM-LASSO, practical criteria to selected λ are AIC and BIC. Specifically,

$$\widehat{\beta_X}^{AIC} = argmin\frac{1}{n^2}M + \frac{2}{n}B_X|\beta_X|_0$$
$$\widehat{\beta_X}^{BIC} = argmin\frac{1}{n^2}M + \frac{\log n}{n}B_X|\beta_X|_0 \tag{9}$$

in which $|\beta_X|_0$ is the number of nonzero coefficients, and B_X is a slowly diverging deterministic sequence (Shi 2016).

References for Appendix A

Arlot S & A Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys 4*:40-79.

Belloni A, D Chen, V Chernozhukov & C Hansen. 2012. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*:2369-2429.

Belloni A, V Chernozhukov & C Hansen. 2014. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies 81*:608-650.

Belloni A, V Chernozhukov, C Hansen & D Kozbur. 2016. Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics 34*:590-605.

Caner M. 2009. Lasso-type GMM estimator. *Econometric Theory 25*(01):270-290.

Geisser S. 1975. The predictive sample reuse method with applications. *Journal of the American Statistical Association 70*: 320-328.

Jing BY, QM Shao & Q Wang. 2003. Self-normalized Cramér-type large deviations for independent random variables. *Annals of Probability 31*: 2167-2215.

Shi Z. 2016. Estimation of sparse structural parameters with many endogenous variables. *Econometric Reviews 35*(8-10):1582-1608.

Zhang Y, R Li, and CL Tsai. 2010. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association 105*: 312-323.

Zou, H., T. Hastie, and R. Tibshirani. 2007. On the “degrees of freedom" of the lasso. *Annals of Statistics 35*: 2173-2192.

Appendix B: LASSO plots for each input variable in translog production function estimation (2) with DLD data

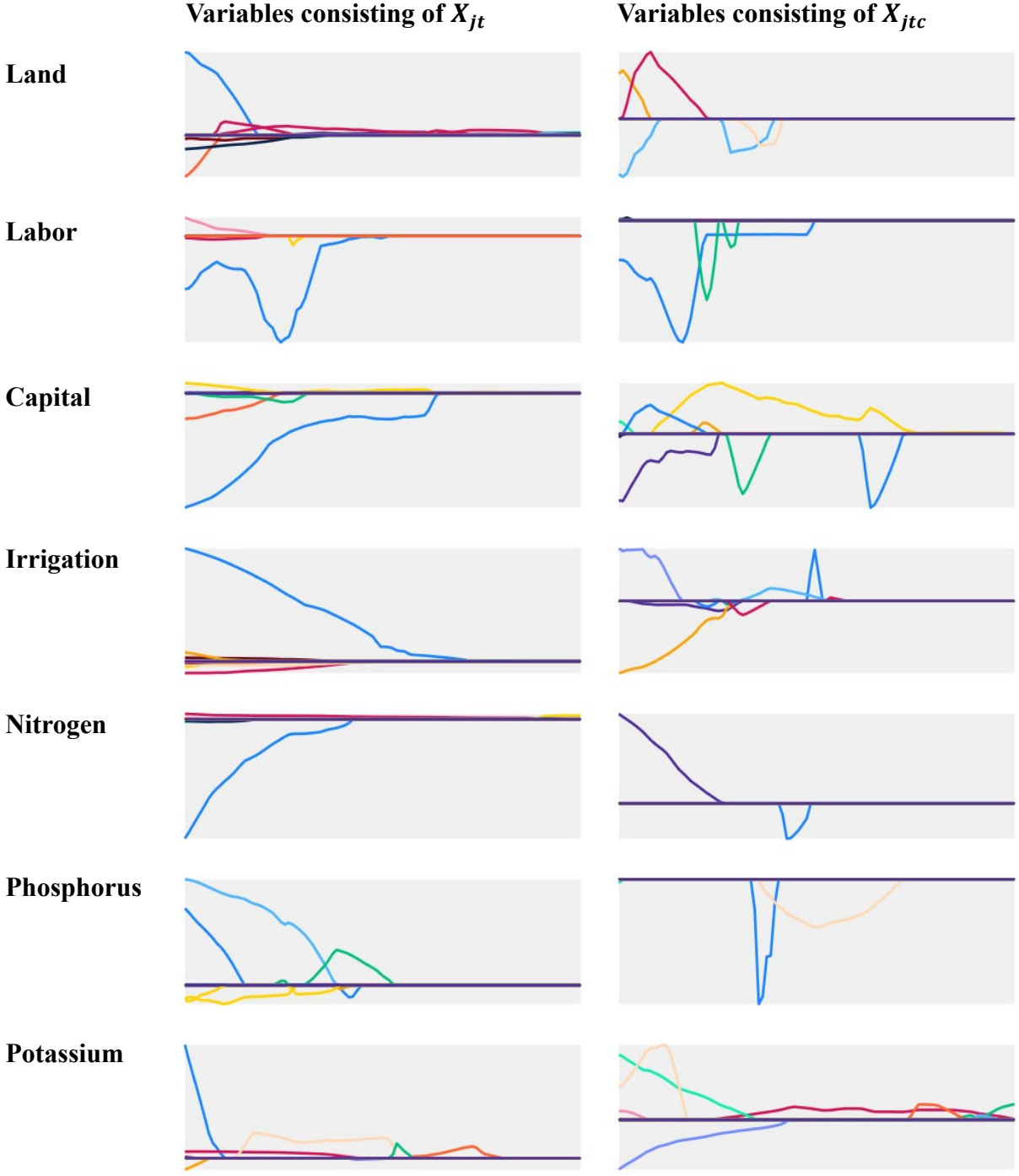

Figure 2. LASSO plots for each input variable[a]

Source: Authors.

[a]Horizontal axes indicate the level of penalization (greater penalization and more parsimonious models to the right), and vertical axes indicate the standard coefficients of each variable. Black lines indicate 0 for coefficient, i.e., variables are dropped due to insignificance conditional no other variables.

Appendix C: Additional results

Table C1. LASSO estimation with DLD data (Table *5*) – subsample districts based on soil types

Soil attributes / LASSO selected variables[a]	Sub-samples with higher soil attributes	Sub-samples with lower soil attributes
	Coef. (std.err)	Coef. (std.err)
Organic contents	Higher	Lower
$K_{jtc}A_{jt}$	0.045*** (0.015)	0.032*** (0.010)
$N_{jt}A_{jt}$		0.264** (0.081)
Other controls	Included	Included
λ	461.894	478.478
Samples	3,438	3,616
Alkalinity (pH)	Higher	Lower
$K_{jtc}A_{jt}$	0.044*** (0.013)	
$N_{jt}A_{jt}$	0.278*** (0.126)	
$P_{jt}A_{jt}$	0.029 (0.118)	
$N_{jt}N_{jt}$		0.025** (0.008)
Other controls	Included	Included
λ	459.299	484.746
Samples	3,517	3,537
Drainage	More excessive	Poorer
$K_{jtc}A_{jt}$	0.016* (0.010)	0.061** (0.015)
$N_{jt}A_{jt}$	0.025 (0.073)	0.097 (0.145)
$P_{jt}A_{jt}$		–0.288*** (0.103)
$W_{jtc}A_{jt}$	0.011 (0.015)	
Other controls	Included	Included
λ	465.290	468.771
Samples	3,628	3,426

Source: Authors. ***1% **5% *10%.

[a] A = land, W = capital, N = nitrogen, P = phosphorus, K = potassium.

Table C2. LASSO estimation with DLD data (Table *5*) – subsample districts based on area shares of grains, pulses, fruits/vegetables

Cropping systems / LASSO selected variables[a]	Sub-samples with higher area shares of the relevant crop-group	Sub-samples with lower area shares of the relevant crop-group
	Coef. (std.err)	Coef. (std.err)
Grain	Higher	Lower
$K_{jtc}N_{jt}$	0.170** (0.054)	
$K_{jtc}A_{jt}$		0.032***(0.012)
$N_{jt}A_{jt}$		0.223 (0.148)
Other controls	Included	Included
λ	543.729	381.970
Samples	3,529	3,525
Pulses	Higher	Lower
$K_{jtc}A_{jt}$	0.032***(0.011)	0.037***(0.013)
$N_{jt}A_{jt}$	0.018 (0.129)	–0.048 (0.074)
Other controls	Included	Included
λ	412.871	516.372
Samples	3,451	3,603
Fruits/Vegs	Higher	Lower
$K_{jtc}A_{jt}$	0.040***(0.012)	
$K_{jtc}N_{jt}$		0.016***(0.006)
$W_{jt}A_{jt}$	0.219***(0.036)	
$N_{jt}N_{jt}$		–0.103* (0.060)
Other controls	Included	Included
λ	356.365	549.493
Samples	3,802	3,252

Source: Authors. ***1% **5% *10%.

[a]A = land, W = capital, N = nitrogen, K = potassium.

Table C3. Same sets of results for Table 6 but using ln(X + 0.001) instead of ln(X + 1)

LASSO-selected inputs[a]	GMM-LASSO based on AIC	GMM-LASSO based on BIC
	Coef. (std.err)	Coef. (std.err)
K_{itc}	0.186** (0.073)	0.213*** (0.063)
$A_{it}K_{itc}$	–0.123 (0.082)	–0.206*** (0.053)
$E_{it}E_{it}$	0.519*** (0.095)	0.489*** (0.068)
A_{it}	0.086 (0.095)	0.111 (0.073)
$W_{it}W_{it}$	0.072* (0.039)	0.084** (0.038)
R_{it}	0.095* (0.055)	0.117** (0.054)
$E_{it}L_{it}$	0.059 (0.089)	
$P_{itc}P_{itc}$	–0.025 (0.058)	
$A_{it}N_{it}$	–0.043 (0.073)	
Other controls	Included	Included
Samples	3041	3041
p-value (H_0: under-identified)	.003	.000
p-value (H_0: not over-identified)	.247	.326

Source: Authors. ***1% **5% *10%.

[a] A = land, W = capital, L = labor, E = other expenditures, R = irrigation, N = nitrogen, P = phosphorus, K = potassium.

ALL IFPRI DISCUSSION PAPERS

All discussion papers are available **here**

They can be downloaded free of charge

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
www.ifpri.org

IFPRI HEADQUARTERS
1201 Eye Street, NW
Washington, DC 20005 USA
Tel.: +1-202-862-5600
Fax: +1-202-862-5606
Email: ifpri@cgiar.org