Panel Data Analysis: "Agriculture in China"

Gokul Prakash Kolakkattil (1312877)

12/01/2024

Contents

List of Tables

List of Figures

Abstract

This research focuses on estimating a Cobb-Douglas production function to examine the intricate relationships between agricultural inputs and outputs in China from 1986 to 1991. Leveraging a robust data set of 218 farm-level observations, panel data analysis is employed to capture multifaceted insights. The Pooled OLS, fixed effects, random effects, and time-fixed effects models are used for this study. The study evaluates the importance of key inputs, including labor, capital, land, and intermediate inputs, and examines trends in agricultural productivity over time. The "Trend" variable, which measures the impact of technical change, has positive and statistically significant coefficients across all three models (Pooled OLS, Fixed Effects, and Random Effects) that point to a significant and steady increase in agricultural output over time. This suggests that there has been continuous advancement or improvement in agricultural production technologies, methods, or efficiency between 1986 and 1991. The Fixed Effects model provides a more nuanced understanding of the interactions between inputs and outputs within each entity (farm) with its inclusion of individual-specific characteristics. The contrast between time fixed effects and fixed effects models leads to the conclusion that individual-specific characteristics play a more crucial role in explaining the observed variations within this data set. The analysis highlights the significance of panel data and how it influences future agricultural plans and policy decisions.

keywords: Cobb-Douglas production function, panel data, pooled OLS, fixed effects, random effects, time fixed effects, technical change

1 Introduction

In the field of agricultural economics, it is critical to comprehend the complex interactions that exist between inputs and outcomes. The estimation of a Cobb-Douglas production function stands as a fundamental tool in unravelling these connections (Hayami 1970). In this report, we embark on an analysis centred on estimating a Cobb-Douglas production function, employing agricultural output as the dependent variable. This investigation encompasses four pivotal inputs land, labour, capital, and intermediate inputs alongside a linear trend. Our data set encompasses a robust collection of 218 farm-level observations, meticulously compiled to capture the nuances of input utilization and agricultural output within the Chinese agricultural landscape. Spanning the years from 1986 to 1991, this data set encapsulates a critical period in China's agricultural evolution, providing a comprehensive window into the dynamics of agricultural production during a time of significant economic shifts and reforms.

When dealing with data sets that include cross-sectional and time-series characteristics, panel data analysis seems to be an effective tool (Baltagi 2008). This analytical method is widely used in many different fields, including the social sciences, economics, and finance. In contrast to conventional data sets that isolate singular dimensions, panel data also known as longitudinal or cross-sectional time-series data integrates multifaceted insights from multiple entities across different time intervals. Therefore, observations in panel data involve at least two dimensions; a cross-sectional dimension, indicated by subscript 'i', and a time series dimension, indicated by subscript 't'. Panel data, by blending the inter-individual differences and intra-individual dynamics, have several advantages over cross-sectional or time-series data.

Panel data analysis offers several key benefits. One advantage lies in its capacity to yield more precise model parameter inference. Panel data typically exhibits higher sample variability and degrees of freedom compared to pure cross-sectional or time-series data. For instance, cross-sectional data could be seen as a panel with a single time period (T = 1), while time-series data could be viewed as a panel with only one entity (N = 1). Hsiao (1995) highlights how econometric estimates are greatly enhanced by this abundance of data. Moreover, it facilitates the control of unobserved heterogeneity through fixed effects or random effects models, allowing researchers to isolate individual-specific characteristics from the analysis. Thirdly, it enables the investigation of dynamic relationships, such as causality and feedback effects, by examining changes within units over time. This richer data set structure enables the exploration of complex phenomena that cannot be adequately captured by cross-sectional or time-series data alone.

There are two primary categories of panel data structures: unbalanced panels and balanced panels. Balanced panels comprise data sets where every unit appears at each time point, ensuring consistency throughout the data set. However, panels that are not balanced have different numbers of observations for every unit. This disparity often occurs due to missing data, units entering or exiting the study over time, or irregular intervals between observations. In our study, we concentrate on a consistent data set of 218 observations, tracked across a span of 6 years.

Panel data analysis usually requires the use of specialized econometric methods designed for this type of data format. Pooled Ordinary Least Squares (OLS), fixed effects, random effects, and dynamic panel data models are examples of common approaches. When using pooled OLS, individual heterogeneity and potential correlation across time periods are ignored and all observations are treated similarly. While random effects models assume that individual effects are random and uncorrelated with the regressors, fixed effects models account for time-invariant individual effects by introducing dummy variables for each entity.

Each observation in the China agriculture data set captures a specific instance of the intricate relationships that exist between different inputs and agricultural production across a six-year period on multiple farms. Our objectives are twofold as we proceed with this analysis: first, to identify potential trends or patterns in agricultural productivity over time; second, to determine the relative significance and impact of labor, capital, land, and intermediate inputs on agricultural output. The Cobb-Douglas function model is used to determine the coefficients linked to each input. However, incorporating a linear trend variable enables us to capture the impact of time and to identify changes and patterns in agricultural productivity over the specified years.

Throughout this report, we navigate through the intricate fabric of agricultural production in China, leveraging the robustness of panel data comprising multiple farms observed over several years. We recognize the significance of this analysis as it offers insights into the factors influencing production and productivity within the Chinese agricultural sector. These insights can inform forthcoming policy decisions and shape future agricultural strategies.

2 Data and descriptive statistics

The data set captures 218 distinct records detailing agricultural practices within China over the span of 1986 to 1991. These entries delve into input factors and agricultural output, encompassing diverse socio-economic indicators like capital, land, labor, and intermediate inputs. Each entry is labeled with a unique index, allowing for individual identification.

Table 1: Summary statistics

Statistic	N	Mean	St. Dev.	Min	Max
Index	1,308	109.500	62.955	1	218
Capital	1,308	2,319.390	4,242.694	19.200	62,136.020
Land	1,308	3.196	2.198	0.300	53.500
Labour	1,308	493.217	218.612	25	1,436
IntermediateInputs	1,308	2,697.717	5,220.415	13.538	107,960.300
Output	1,308	6,409.341	6,738.235	161.680	110,274.500
Year	1,308	1,988.500	1.708	1,986	1,991
Trend	1,308	3.500	1.708	1	6

The index corresponds to a sequential numbering of observations from 1 to 218. Year denotes the data collection period, primarily concentrated between 1986 and 1991. The village column indicates specific village numbers where data collection took place. This range of values implies the data set covers multiple villages, each with multiple entries. Output reflects the quantity of output generated, showcasing a wide range from 161.7 to 110274.5, indicating substantial disparities in productivity across observations. Labor quantifies the level of employed labor, varying from 25.0 to 1436.0, showcasing differing labor intensities. Capital records investment levels, spanning from 19.2 to 62136.0, highlighting varying investment magnitudes across observations. Land represents the land area associated with the activities recorded, displaying differences in land utilization or availability across entries (ranging from 0.300 to 53.500). Intermediate inputs signify the use of intermediate resources, showcasing diverse usage or investment levels (ranging from 13.54 to 107960.31). The “Trend” column denotes qualitative trends in the observations, presenting categorical variations from 1 to 6 among entries.

Figure 1 shows a left-skewed normal distribution of the variables, which include labor, capital, output, land, and intermediate inputs. This pattern suggests that the data is concentrated in lower values, with fewer occurrences of very high values. For Output, this suggests that most observations tend to have lower output levels, with only a few instances of very high outputs. Similarly, Capital, Land, Labour, and Intermediate Inputs display a similar trend, where the majority of observations lean towards lower values.

Figure 2 shows the relationship between independent variables and output. Capital and intermediate inputs appear to be the main factors influencing output levels. Labour also plays a moderate role in influencing output but to a lesser extent compared to intermediate inputs and capital. Land have a weaker associations with Output, suggesting it might have less direct impact or might be indirectly related to output through other variables.

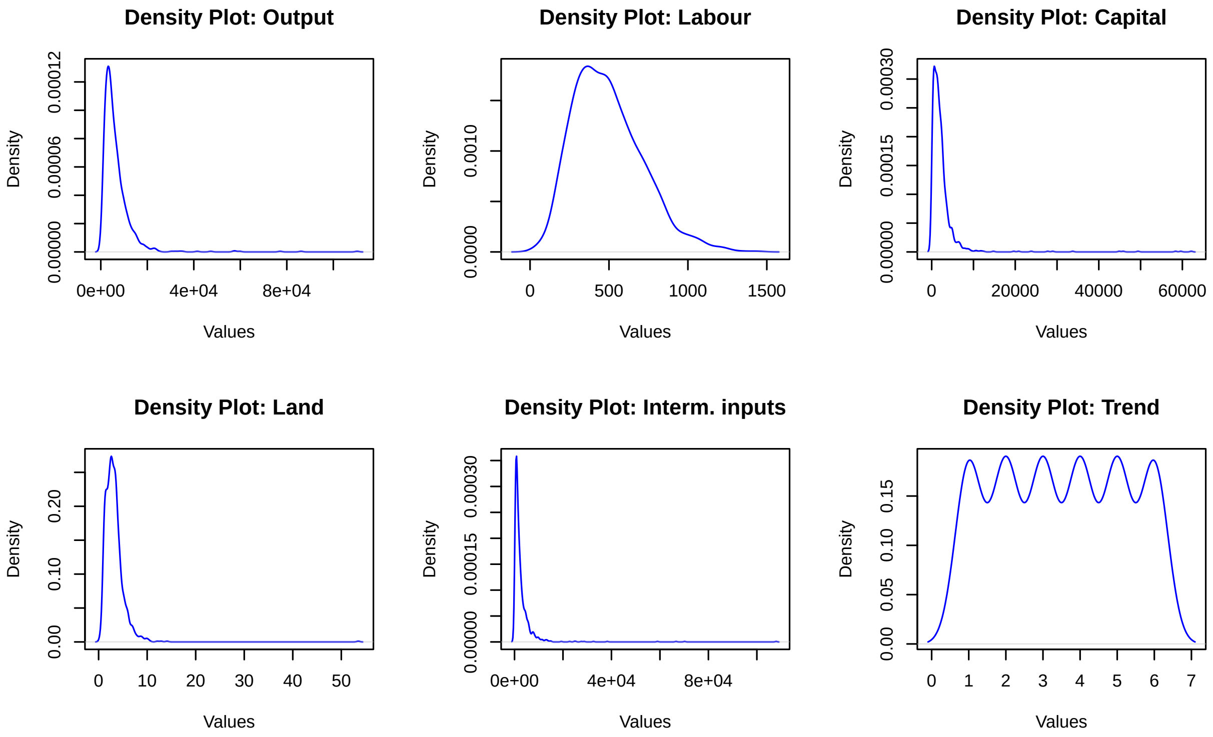

Figure 1: Overall density plots of features

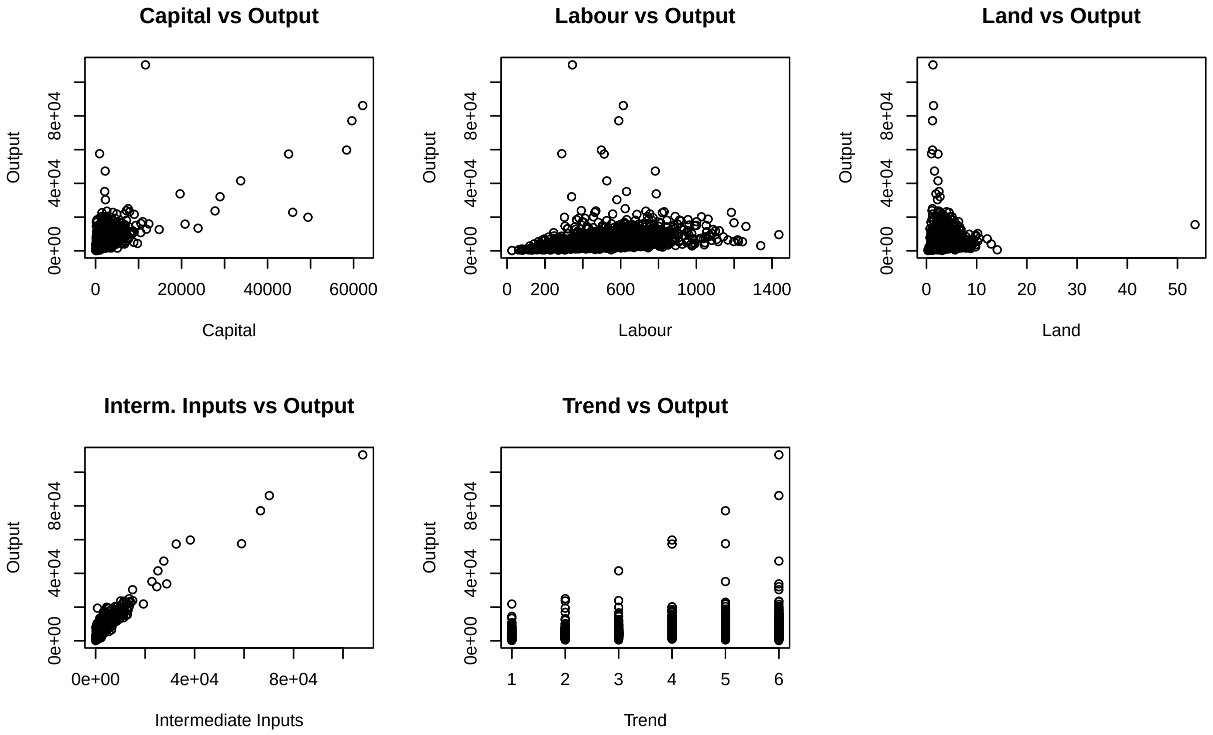

Figure 2: Overall scatter plots of features

3 Methodology

In this section, we delve into four pivotal models used in panel data analysis: Pooled OLS, Fixed Effects, Random Effects, and Time Fixed Effects models.

3.1 Pooled OLS (Ordinary Least Squares)

Pooled OLS model serves as an initial methodological approach in panel data analysis (Wooldridge 2010). This method considers all entities or individuals as a homogeneous group, estimating a shared intercept and slope for all entities across all time periods through a simple linear regression model:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}$$

Where: - Y_{it} represents the dependent variable for entity 'i' and time 't'. - X_{it} denotes the independent variables for entity 'i' and time 't'. - β_0 and β_1 are the intercept and coefficients to be estimated. - u_{it} is the error term.

However, Pooled OLS assumes no individual-specific effects, potentially leading to biased estimators if such effects exist and are unaccounted for. This can violate the assumption of independent error terms, potentially impacting the reliability of the results obtained.

3.2 Fixed Effects Model

Fixed Effects model (Baltagi 2008) aim to address the limitations of Pooled OLS by accounting for individual-specific effects. This model incorporate entity-specific dummy variables or fixed effects to capture unobserved heterogeneity across entities that remain constant over time:

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it}$$

Where: - α_i represents entity-specific intercept for entity 'i'.

By controlling for individual effects, Fixed Effects models can mitigate potential biases arising from omitted variables that are constant over time but vary across entities.

3.3 Random Effects Model

Random Effects model (Bell and Jones 2015) extend the analysis further by integrating both entity-specific effects and random errors. This model assume that the entity-specific effects are uncorrelated with the regressors:

$$Y_{it} = \alpha + \beta_1 X_{it} + \gamma_i + u_{it}$$

Where: - α denotes the overall intercept. - γ_i represents the entity-specific random effects.

Random Effects model estimate variance components of entity-specific effects and the error term, offering efficiency gains over Fixed Effects models when entity-specific effects are uncorrelated with the regressors.

3.4 Time Fixed Effects Model

Time fixed effects model involve incorporating time-specific variables into regression models to account for and control variations occurring uniformly across all entities during specific time intervals (Imai and Kim 2021):

$$Y_{it} = \alpha + \beta_1 X_{it} + \gamma_t + u_{it}$$

Where: - γ_t refers to the time fixed effects capturing time-specific influences.

4 Results and Discussions

In the following section Pooled OLS, Fixed Effects, Random Effects, and Time Fixed Effects models are estimated and analyzed. The optimal model for the data set is identified using Pooled test and Hausmann test. Subsequently, the chosen model undergoes tests for heteroskedasticity and auto-correlation. The identified heteroskedasticity issue is resolved by incorporating robust standard errors.

4.1 Regression Models

Table 2: Regession models

	Dependent variable:		
	Output		
	Pooled OLS	Fixed Effects	Random Effects
	(1)	(2)	(3)
Land	-123.643^{***}	-66.976^{**}	-90.442^{***}
	(27.288)	(28.357)	(26.497)
Labour	4.306^{***}	5.431^{***}	5.017^{***}
	(0.282)	(0.307)	(0.282)
Capital	0.233^{***}	0.126^{***}	0.181^{***}
	(0.017)	(0.022)	(0.018)
IntermediateInputs	1.009^{***}	0.992^{***}	0.998^{***}
	(0.014)	(0.012)	(0.012)
Trend	474.060^{***}	548.204^{***}	514.236^{***}
	(34.642)	(28.350)	(27.980)
Constant	-241.237		-687.569^{***}
	(196.110)		(201.245)
Observations	1,308	1,308	1,308
R^2	0.908	0.919	0.914
Adjusted R^2	0.907	0.902	0.914
F Statistic	$2,558.103^{***}$ (df = 5; 1302)	$2,459.983^{***}$ (df = 5; 1085)	$13,866.120^{***}$
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 2 shows the regression results for Pooled OLS, Fixed Effects, and Random Effects models. The Pooled OLS model reveals insights into the relationships between input factors and agricultural output. Land displays an inverse relationship with output, indicating that a unit increase in land leads to a decline in output by approximately 123 units. However, labor and capital show a positive impact, with each unit increase resulting in a rise of about 4.31 units and a 0.23-unit rise in output respectively. Moreover, intermediate inputs display a robust positive influence, where an increase of one unit corresponds to around a 1.01-unit rise in output. Finally, the trend variable illustrates a notably strong positive relationship with output, suggesting an increase of roughly 474 units in output per unit change in the trend. The model explains roughly 90.8% of the variability in output, reflecting its high explanatory power and reliability in capturing the relationships between these inputs and agricultural output.

The Fixed effects model explains 91.9% of the variability in output, highlighting the significant influences of these inputs on agricultural output within the data set. Based on the figures, there is a negative correlation between Land and output, which means that for every unit increase in Land, the output decreases by roughly 67 units. Conversely, labor and capital exhibit a positive relationship; a unit increase yields 5.43 units and 0.13 units of increased output, respectively. Additionally, there is a strong positive correlation found in Intermediate Inputs; an increase of one unit corresponds to an almost 0.99-unit rise in output. Moreover, a very significant positive correlation is shown by the Trend variable, indicating a rise in output of around 548 units for every unit change in the trend.

The Random Effects model reveals substantial relationships between the independent variables and agricultural output. Notably, land exhibits an inverse relationship, a unit increase is associated with a decrease of approximately 90 units in output. Labor and Capital exhibit positive effects, indicating that a one-unit rise in both labor and capital corresponds to increases of approximately 5.02 units and 0.18 units in output, respectively. Intermediate inputs demonstrate a robust positive influence, with a unit increase associated with roughly a one-unit rise in output. The trend variable showcases a robust positive relationship with output, implying an increase of around 514 units in output per unit change in the trend. The model explains roughly 91.4% of the variability in output, reflecting its high explanatory power and reliability in capturing the relationships between these inputs and agricultural output.

4.2 Pool test (Pooled OLS vs Fixed Effects)

The poolability test assesses whether there is a significant difference in the coefficients between the Pooled OLS and Fixed Effects model, helping to determine if including individual fixed effects significantly improves the model fit compared to the Pooled OLS approach. The pool test from Table 3 yields a p-value of less than 0.05, which suggests strong evidence against the poolability null hypothesis and suggests the presence of individual-specific effects that significantly affect the model.

Table 3: Pool test (Pooled OLS vs Fixed Effects)

	df1	df2	statistic	p.value	method	alternative
1	217	1085	4.985	0	F statistic	unstability

4.3 Hausmann Test (Fixed Effects vs Random Effects)

The “Hausman test” is a statistical test employed to assess the appropriateness of choosing between fixed effects and random effects models in panel data analysis (Park 2010). The null hypothesis is that the coefficients estimated by the fixed effects and random effects models do not significantly differ from one another. Since the p-value from Table 4 is less than 0.05, the null hypothesis is rejected. It suggests that the fixed effects model is the better choice and that the random effects model is inconsistent. The fixed effects model is considered appropriate when there are unobserved time-invariant individual effects that might be correlated with the independent variables. It accounts for individual-specific intercepts and controls for unobserved heterogeneity.

Table 4: Hausmann Test (Fixed Effects vs Random Effects)

	statistic	p.value	parameter	method	alternative
1	37.754	0	c(df = 5)	Hausman Test	one model is inconsistent

4.4 Test for Heteroskedasticity

The Breusch-Pagan test was conducted to assess the presence of heteroskedasticity in the residuals of our best model - the fixed effects model. The null hypothesis assumes homoskedasticity, implying that the variance of the residuals remains constant across observations. Conversely, the alternative hypothesis indicates different variances and implies the presence of heteroskedasticity (Baltagi 2008). Table 5 shows a high test statistic (BP = 110.43) with 5 degrees of freedom, yielding a remarkably low p-value (< 2.2e-16). The low p-value provides strong evidence to reject the null hypothesis of homoskedasticity. Hence, we conclude the presence of heteroskedasticity in the residuals.

Table 5: Test for Heteroskedasticity

	statistic	p.value	parameter	method
1	110.435	0	c(df = 5)	studentized Breusch-Pagan test

4.5 Test for Autocorrelation

The purpose of the Breusch-Godfrey/Wooldridge test for serial correlation in panel models is to determine whether the idiosyncratic errors exhibit any signs of serial correlation. The null hypothesis assumes no serial correlation, while the alternative hypothesis suggests the presence of serial correlation (Baltagi 2008). Table 6 shows a p-value less than 0.05, hence the data provide strong evidence to reject the null hypothesis. As a result, the panel model's idiosyncratic errors exhibit a strong serial association.

Table 6: Test for Auto-correlation

	statistic	p.value	parameter	method	alternative
1	233.84	0	c(df = 6)	Breusch-Godfrey/Wooldridge	serial correlation in idiosyncratic errors

4.6 Time Fixed Effects and Fixed effects (Robust std. errors) models

The heteroskedasticity-robust standard errors are used to adjust the fixed effects panel model for heteroskedasticity based on the Arellano method. The test results in Table 7 indicate that all coefficients are statistically significant (p-values < 0.05), suggesting that the corresponding variables have a significant impact on the dependent variable in the fixed effects model. While the positive coefficients for labor, capital, intermediate inputs, and trend indicate positive connections with the dependent variable, the negative coefficient for land implies an inverse relationship.

The results of two-way fixed effects within model in Table 7 reveal significant associations between the independent variables and agricultural output. In particular, Land shows a statistically significant negative association, with a unit increase associated with a decrease of approximately 65 units in output. Positive influences are seen in labor and capital, suggesting that a one-unit increase in each variable results in increases in production of roughly 5.30 units and 0.12 units, respectively. Moreover, Intermediate inputs showcase a robust positive impact, with a unit increase linked to approximately 1 unit rise in output. These findings underscore the importance of considering individual and time-specific effects in understanding the dynamics of agricultural production in China. The model's overall statistical importance is highlighted by its strong R-squared value and significant F-statistic, which both point to a high goodness-of-fit.

Table 7: Time Fixed Effects vs Fixed effects model (Robust std. errors)

	Dependent variable:	
	Output	
	Fixed effects (Robust std. errors)	Time Fixed Effects
	(1)	(2)
Land	-66.976^{***}	-65.306^{**}
	(14.692)	(27.598)
Labour	5.431^{***}	5.296^{***}
	(0.355)	(0.300)
Capital	0.126^{***}	0.122^{***}
	(0.034)	(0.021)
IntermediateInputs	0.992^{***}	0.998^{***}
	(0.039)	(0.012)
Trend	548.204^{***}	
	(38.087)	
Observations	1,308	1,308
R^2	0.919	0.903
Adjusted R^2	0.902	0.883
F Statistic	$2,459.983^{***}$ (df = 5; 1085)	$2,527.626^{***}$ (df = 4; 1081)
Note:		*p<0.1; **p<0.05; ***p<0.01

4.7 Hausmann Test (Time Fixed Effects vs Fixed Effects)

The test result obtained from Table 8 shows a p-value larger than 5% significance level, hence there is no strong evidence to reject the null hypothesis. This result suggests that there is no substantial difference between the coefficients estimated by the models. Therefore, it indicates that the inclusion of time fixed effects is unnecessary. In essence, the fixed effects model alone adequately captures and explains the variations within the data set. Hence, the Cobb Douglas production function with trend variable is more accurate in revealing the variations in the data set.

Table 8: Hausmann Test (Time Fixed Effects vs Fixed Effects)

	statistic	p.value	parameter	method	alternative
1	8.408	0.078	c(df = 4)	Hausman Test	one model is inconsistent

5 Conclusion

The regression models in this study were able to capture the intricate relationship between the various factors namely, capital, land, labour, intermediate inputs, and trends with the agricultural output, shedding light on the dynamics of the Chinese agricultural sector during the time frame 1986 to 1991. In conclusion, the adoption of a Fixed Effects model emerges as the optimal choice for our data set, outperforming alternative models such as Pooled OLS, Random Effects, and Time Fixed Effects. This preference emphasizes the importance of unobserved, time-invariant individual-specific characteristics in our data. The Fixed Effects model excels at capturing the nuanced effect of these specific factors on the dependent variable, resulting in a more accurate picture of the underlying dynamics. The positive and statistically significant coefficients of the "Trend" variable across all three models indicate a substantial and consistent increase in agricultural output over time. This pattern illustrates the influence of technological development, implying that agricultural production technologies, procedures, or efficiency have been continuously enhanced or advanced between 1986 and 1991. The negative coefficient in the Pooled OLS model suggests that, on average, an increase in land area is associated with a decrease in output. The positive and statistically significant coefficients for labor and capital in all models highlight the positive impact of labour intensity and investment in agricultural capital on output. Increased capital investment, such as in machinery or infrastructure, is associated with higher agricultural productivity. The positive and statistically significant coefficients for intermediate inputs across all models indicate that employing more resources improves agricultural output. Fertilizers, insecticides, and other inputs may fall within this category. Considering these results, our analysis emphasizes the critical role of unobserved individual-specific characteristics and technological progress in shaping the outcomes within our data set. These findings might have significant policy implications, emphasizing the importance of acknowledging entity-specific characteristics and cultivating an environment that supports ongoing technological innovation within the agricultural sector.

References

Baltagi, H Badi. 2008. *Econometric Analysis of Panel Data.* Vol. 4. Springer.

Bell, Andrew, and Kelvyn Jones. 2015. "Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data." *Political Science Research and Methods* 3 (1): 133–53.

Hayami, Yujiro. 1970. "On the Use of the Cobb-Douglas Production Function on the Cross-Country Analysis of Agricultural Production." *American Journal of Agricultural Economics* 52 (2): 327–29. http://www.jstor.org/stable/1237509.

Hsiao, Cheng. 1995. "Panel Analysis for Metric Data." In *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, 361–400. Springer.

Imai, Kosuke, and In Song Kim. 2021. "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis* 29 (3): 405–15.

Park, Hun M. 2010. "Practical Guides to Panel Data Analysis." *International University of Japan. Recuperado de Http://Www. Iuj. Ac. Jp/Faculty/Kucc625/Writing/Panel_guidelines. Pdf.*

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* MIT press.