Supplementary Materials for "Social interactions and spillover effects in Chinese family farming"

1. Method

The demeaning procedure

Plot-fixed effects and crop seasons fixed effects are controlled for by using either a demeaning approach or a bias correction procedure applied on the demeaning approach proposed by Lee and Yu (2010a, 2010b).

The demeaning procedure consists in using data transformation either to remove individual fixed effects with a deviation from the cross-sectional mean operator ($J_n = I_n - \frac{1}{N} l_n l_n'$) or to delete time fixed effects with a deviation from the time mean operator ($J_t = I_t - \frac{1}{T} l_t l_t'$) (where l_n and l_t are vectors of ones) and then using the quasi-maximum likelihood (QML).

However, in order not to create linear dependence on the resulting disturbances which may generate inconsistency, Lee and Yu (2010a) suggest a bias correction procedure which states that the transformations be based on the orthonormal eigenvector matrices of J_n and J_t. This implies reducing the number of observations from N*T to (N-1)*(T-1) observations. See Lee and Yu (2010a, pp.170) and Elhorst (2014, pp.48) for a detailed presentation.

It is worth noting that the coefficient estimates and their significances from the demeaning approach (col. 4 in Table 2) are close to the coefficient estimates using the transformation approach in the last column. This can be explained by the relatively small time periods observed compared to the number of plots in our sample. In fact, Lee and Yu (2010a) argue that in a two-way fixed effects spatial model (such as one used in col. 4 in Table 2), the ML estimator leads to inconsistency in the coefficient estimates only when both N and T are high. However, while they show that the ML estimator cannot yield a consistent estimator of the variance parameters if N is large and T is small (compared to N – so what we have), the significance of our results do not change.

Direct and indirect effects in the SDM model

The coefficients cannot be interpreted in the same way as parameter estimates in a non-spatial model in which they are the marginal effects of a change of explanatory variables on the dependent variable. Put differently, the coefficient estimates of X (β) cannot be interpreted as direct effects, and the coefficient estimates of WX (θ) cannot be interpreted as spillovers.

Firstly, the coefficient estimates of X (β) are different from direct effects due to feedback effects (related to the coefficients of Wy (δ) and WX (θ)) that happen as a result of impacts moving through neighboring plots and back to the plot itself. For instance, an increase of one particular input in plot i modifies the yield on i but this effect moves through neighboring plots of i since they are feedback effects between plots (i.e., a change of X modifies both Wy and WX which in turn modify y). More precisely, Elhorst (2010) mentions the direct effect of an explanatory variable of an individual i on the outcome of i through neighbours of i as higher-order direct effects. Thus the overall direct effect is the first-order direct effect (a change in X only) (β) plus the higher-order direct effects (a change in WX and WY). Thus, Elhorst (2010) points out that the direct effect in an SDM model is higher than a direct effect without spatial effects.

Secondly, in the same vein, the coefficient estimates of WX (θ) are different from indirect effects due to the same feedback effects. For instance, a change of a particular input on plot j has an effect on the yield of plot i but this effect passes through all other neighbours of j and i.

Interestingly, spillovers produced by the SDM are global by nature due to the presence of the spatial lag (Elhorst, 2010, LeSage and Pace, 2014). In others words, in the SDM, any change in an explanatory variable at any location will be transmitted to all other locations following W even if two locations, according to W, are unconnected. By contrast, local spillovers occur only between locations that, according to W, are connected to each other (this is the case of a spatial model without a spatial lag variable, e.g., the spatial error Durbin model (SEDM)).

However, given the block-diagonal structure of our study, the spillovers estimated by the SDM model are local (not global). In fact, the block-diagonal structure states that all individuals in a block are neighbors without any other connections with any other individuals in the other blocks. This means that spillovers only occur between individuals who are connected according to W or, put differently, that spillovers cannot be global because two individuals in two different blocks cannot be connected even by the intermediary of their neighbors (and the neighbors of their neighbors, and so on).

The self-selection model

The social interaction effect may be sensitive to self-selection problem in the adoption of organic farming that implies that conditions for a farmer to practice organic farming may be different from those of his conventional counterpart. The artificial division of the sample will thus create a biased estimation if self-selection exists. To rule out this potential problem, we implement a Heckman correction to the estimation of the two-regime model (model (4) in the main document).

To be valid, this model needs to have an exclusion restriction. Here, this is the distance for a farmer to reach the plot from his house. In fact, the distance should determine the yield (the dependent variable in the two-regime model) only through the farmer's choice to practice organic farming. We assume that the distance is a valid exclusion restriction because organic farming requires much more labor due to transport and application of organic compost and manure. Therefore long distance from house to plot should discourage organic farming.

The self-selection model is as follows:

$$D_{i,t} = \gamma_0 + \gamma_1 Distance_{i,t} + \gamma_2 X_{i,t} + \gamma_3 WX_{j,t} + \tau_t + S_{i,t} + \varepsilon_{i,t} , \tag{1}$$

where $D_{i,t}$ is the dummy variable indicating whether plot i at season t has employed organic or conventional farming (D is also the dummy regime in the two-regime model). The variable $Distance_{i,t}$ is the exogenous variable. X is a matrix of the same explanatory variables used in model (4) (except conventional pesticides that have been dropped because of collinearity (none organic farmers use them)). WX is a matrix of those X at the group level. t_t is a dummy fixing crop seasons to separate the sample for each of the five study periods. $S_{i,t}$ is a dummy variable that controls for the type of seed used on the plot (one type has been removed to avoid collinearity (only one organic farm uses this type)). $e_{i,t}$ is the error term. From model (1), we calculate the *inverse Mills ratio* and use it as a new control variable in model (4) in the main document.

Tests of the SDM model against alternative models (SAR, SEM, SLX and non-spatial model)

Before turning to the comparisons between the SDM model and the other models, it is worth noting that all these models are fixed spatial effects models. However, a random spatial effects model with different specifications (with or without time fixed effects, with or without *WX*) has also been estimated and tested against the spatial fixed effects model with Hausman tests. In each case, this model is rejected, i.e., the fixed effects model is more appropriate (cf. the Hausman tests hereafter).

Hausman tests (spatial random effects against spatial fixed effects)

Variables	Wy	X	WX	TE	FE	Hausman test
Model 1	x	x				220.18***
Model 2	x	x		x	x	149.94***
Model 3	x	x	x	x	x	59.95***

Note: *** Significant at 1 percent; probability < 0.025 implies rejection of random effects model in favor of fixed effects model. Y, X and WX are the same variables used in equation 3 in the main document. Fixed effects (FE) are used only in spatial fixed effects models. TE are time-fixed effects.

The SDM model shown in the fourth and fifth columns of Table 2 is tested against a spatial lag model (SAR - with *Wy* only) and a spatial error model (SEM - with spatial correlations between errors only) in two ways. Firstly, the coefficient estimates of the SDM model are used to test the two following hypotheses: $H_0: \delta = 0$ and $H_0: \delta = -\theta\delta$. The first hypothesis states whether the SDM model can be simplified to the SAR model, whereas the second hypothesis states whether the SDM model can be simplified to the SEM model. Both are Wald tests following an c^2 distribution with *K* degrees of freedom, where *K* is the number of explanatory variables. We also use the Likelihood Ratio (LR) test after estimating the SAR and SEM models alternatively (the results of these two models are not presented here but are available upon request). Both types of tests (the Wald and LR tests) lead us to the conclusion that the SDM model should be adopted. Moreover, we use an LR test after estimating a panel SLX model (with only *WX*) and a panel non-spatial model in which the yield is not explained by any spatial dependence (neither *Wy* or *WX* or spatial correlations between errors). These tests confirm that the SDM model should be preferred both to an SLX model and a non-spatial model.

Test of spatial interaction effects

We also employ tests for the presence of spatial interaction effects in our panel data setting. We use Lagrange Multiplier (LM) tests for a spatially lagged dependent variable and for spatial error correlation as well as their robust version counterpart developed by Anselin et al. (2006) for a panel regression model (Elhorst, 2009). We use Matlab routines written by Donald J. Lacombe and available at http://community.wvu.edu/~djlacombe/matlab.html. These tests are done with time fixed-effects and all together suggest both the existence of a spatially lagged dependent variable and spatial error correlation. While these tests cannot help to choose between the types of spatial dependence (i.e., between the SAR, SDM, SEDM and SEM models), they confirm that there are spatial interactions at stake in our case study (and so the non-spatial model or the SLX model (with only WX as spatial dependence) must be rejected). Results are available upon request.

Presentation of the SEDM model and the GNS model

More formally, the SEDM requires rewriting the structural model in matrix notation as follows:

$$y_{i,t} = \beta X_{i,t} + \theta W X_{j,t} + \alpha_i + \tau_t + \vartheta_{i,t} \,, \tag{2}$$

$$\vartheta_{i,t} = \lambda W \vartheta_{j,t} + \varsigma_{i,t} \,, \tag{3}$$

where λ denotes the coefficient estimates of spatial autocorrelation. Thus, interaction effects in the SEDM model occur through exogenous interactions among independent variables (WX) and interaction among the disturbance term ($W\vartheta$).

However, the GNS model, or the Manski model (Manski, 1993), is a full model with all types of group effects and takes the following form:

$$y_{i,t} = \delta W y_{j,t} + \beta X_{i,t} + \theta W X_{j,t} + \alpha_i + \tau_t + \vartheta_{i,t} \,, \tag{4}$$

$$\vartheta_{i,t} = \lambda W \vartheta_{j,t} + \varsigma_{i,t}. \tag{5}$$

Thus, group effects occur through the social interaction effect (δ), exogenous effects among independent variables (θ) and interaction effects among the disturbance term (λ). The GNS model thus helps to identify the social interaction effect by controlling for contextual effects and spurious correlation.

2. Results

Table S1. Robustness checks: disentangled group effects

	(1)	(2)	(3)	(4)
Determinants	SEDM model		GNS/Manski model	
W*Yield			-.584***	-.579***
			(0.131)	(0.133)
λ	-.246**	-.262**	0.312***	0.306***
	(0.106)	(0.107)	(0.110)	(0.113)
Log(Labor)	0.078	0.06	0.086	0.071
	(0.06)	(0.062)	(0.058)	(0.061)
Log(Cost)	0.174***	0.174***	0.181***	0.18***
	(0.02)	(0.02)	(0.019)	(0.02)
Log(Nitrogen)	0.501***	0.505***	0.486***	0.49***
	(0.038)	(0.038)	(0.037)	(0.037)
Log(Phosphate)	-.122***	-.124***	-.117***	-.119***
	(0.017)	(0.018)	(0.017)	(0.017)
Log(Water)	0.08***	0.081***	0.086***	0.087***
	(0.028)	(0.028)	(0.028)	(0.028)
Log(Pesticide Conv)	0.007	0.014	0.008	0.015
	(0.01)	(0.014)	(0.01)	(0.013)
Log(Pesticide Org)	0.013	0.011	0.015	0.014
	(0.011)	(0.011)	(0.011)	(0.011)
Lamp (0-1)	0.026	0.027	0.025	0.026
	(0.025)	(0.025)	(0.024)	(0.024)
Chemical influence (0-1)	0.059**	0.064**	0.059**	0.063**
	(0.028)	(0.028)	(0.028)	(0.028)
W*Log(Labor)	1.597***	1.653***	1.473***	1.494***
	(0.313)	(0.323)	(0.314)	(0.324)
W*Log(Cost)	0.214***	0.223***	0.229***	0.231***
	(0.079)	(0.08)	(0.081)	(0.081)
W*Log(Nitrogen)	-.379**	-.386**	-.368**	-.369**
	(0.151)	(0.151)	(0.154)	(0.154)
W*Log(Phosphate)	-.156*	-.145	-.151*	-.148
	(0.09)	(0.092)	(0.09)	(0.091)
W*Log(Pesticide Conv)	-0.176	-0.202	-0.206	-0.21
	(0.165)	(0.171)	(0.163)	(0.169)
W*Log(Pesticide Org)	0.19***	0.113	0.176***	0.152
	(0.052)	(0.12)	(0.052)	(0.119)
W*Log(Water)	0.157**	0.146**	0.185***	0.182**
	(0.07)	(0.073)	(0.071)	(0.073)
W*Lamp (0-1)	0.071	0.079	-.012	-.010
	(0.132)	(0.134)	(0.139)	(0.14)
W*Chemical influence (0-1)	0.01	0.008	0.005	0.002
	(0.124)	(0.124)	(0.124)	(0.124)
Organic		0.038		0.035
		(0.043)		(0.042)
W*Organic		-0.264		-0.086
		(0.364)		(0.362)
Seed dummies	x	x	x	x
FE	x	x	x	x
TE	x	x	x	x
N	990	990	990	990
N plots	198	198	198	198
Corrected R2	0.36	0.357	0.347	0.347
Log L	578.13	578.89	581.03	581.43
LR test			5.79**	5.08**

Notes: standard errors in parentheses. The dependent variable is the yield defined as the log of raw rice output per land area. *** statistical significance at 1%, ** statistical significance at 5%, * statistical significance at 10%.

Table S2. Direct and indirect effects based on the coefficients estimates of the GNS model reported in Table S1

	(1)	(2)	(3)	(4)
	From col. 3 Table S2		From col. 4 Table S2	
Determinants	Coef.	Std errors	Coef.	Std errors
Direct				
Log(Labor)	0.056	(0.06)	0.046	(0.07)
Log(Cost)	0.114***	(0.03)	0.114***	(0.02)
Log(Nitrogen)	0.309***	(0.05)	0.313***	(0.05)
Log(Phosphate)	-0.075***	(0.02)	-0.076***	(0.02)
Log(Water)	0.054*	(0.03)	0.056*	(0.03)
Log(Pesticide Conv)	0.005	(0.01)	0.009	(0.01)
Log(Pesticide Org)	0.009	(0.01)	0.009	(0.01)
Lamp (0-1)	0.015	(0.03)	0.015	(0.03)
Chemical influence (0-1)	0.038	(0.03)	0.04	(0.03)
Indirect				
Log(Labor)	0.938***	(0.35)	0.955***	(0.37)
Log(Cost)	0.145	(0.09)	0.147	(0.09)
Log(Nitrogen)	-0.233	(0.17)	-0.235	(0.17)
Log(Phosphate)	-0.095	(0.1)	-0.094	(0.1)
Log(Water)	0.115	(0.08)	0.113	(0.08)
Log(Pesticide Conv)	-0.134	(0.18)	-0.138	(0.2)
Log(Pesticide Org)	0.112**	(0.06)	0.096	(0.13)
Lamp (0-1)	-0.003	(0.17)	-0.002	(0.21)
Chemical influence (0-1)	0.001	(0.14)	-0.001	(0.1)
Total				
Log(Labor)	0.994***	(0.36)	1.001***	(0.37)
Log(Cost)	0.259***	(0.09)	0.261***	(0.09)
Log(Nitrogen)	0.076	(0.17)	0.078	(0.17)
Log(Phosphate)	-0.170*	(0.1)	-0.170*	(0.1)
Log(Water)	0.169**	(0.09)	0.169**	(0.09)
Log(Pesticide Conv)	-0.129**	(0.18)	-0.129	(0.17)
Log(Pesticide Org)	0.122	(0.06)	0.105	(0.14)
Lamp (0-1)	0.012	(0.15)	0.013	(0.17)
Chemical influence (0-1)	0.039	(0.14)	0.039	(0.14)

Note: standard errors in parentheses. *** statistical significance at 1%, ** statistical significance at 5%, * statistical significance at 10%.

Table S3. Robustness checks: other group-structures

Determinants	(1) SDM	(2) SEDM	(3) GNS	(4) SDM	(5) SEDM	(6) GNS
W*Log(Yield)	-.324*		-.018	-.555*		-.199
	0.186		0.48	0.328		0.629
Lambda		-.439*	-.410		-.652	-.379
		0.245	0.595		0.401	0.690
Log(Labor)	0.079	0.079	0.078	0.083	0.08	0.081
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Log(Cost)	0.179***	0.179***	0.178***	0.178***	0.178***	0.178***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Log(Nitrogen)	0.5***	0.495***	0.502***	0.499***	0.495***	0.497***
	(0.038)	(0.037)	(0.038)	(0.038)	(0.037)	(0.038)
Log(Phosphore)	-.122***	-.120***	-.122***	-.120***	-.119***	-.119***
	(0.018)	(0.017)	(0.018)	(0.018)	(0.017)	(0.018)
Log(Water)	0.082***	0.084***	0.08***	0.084***	0.084***	0.084***
	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)
Log(Pesticide Conv)	0.007	0.007	0.007	0.006	0.007	0.007
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Log(Pesticide Org)	0.015	0.015	0.015	0.014	0.016	0.015
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Lamp (0-1)	0.024	0.026	0.022	0.026	0.025	0.026
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
Chemical influence (0-1)	0.057**	0.058**	0.056**	0.061**	0.059**	0.06**
	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)
W*Log(Labor)	1.605***	1.546***	1.626***	1.538***	1.528***	1.531***
	(0.315)	(0.316)	(0.319)	(0.316)	(0.315)	(0.316)
W*Log(Cost)	0.214***	0.223***	0.208***	0.217***	0.216***	0.215***
	(0.079)	(0.08)	(0.081)	(0.081)	(0.08)	(0.081)
W*Log(Nitrogen)	-.373**	-.373**	-.373**	-.386**	-.372**	-.376**
	(0.152)	(0.154)	(0.151)	(0.154)	(0.153)	(0.154)
W*Log(Phosphore)	-.148	-.147	-.150*	-.149	-.155*	-.153*
	(0.091)	(0.091)	(0.091)	(0.091)	(0.091)	(0.091)
W*Log(Water)	-.178	-.190	-.172	-.184	-.190	-.187
	(0.167)	(0.166)	(0.169)	(0.167)	(0.166)	(0.166)
W*Log(Pesticide Conv)	0.19***	0.184***	0.192***	0.187***	0.182***	0.183***
	(0.053)	(0.052)	(0.053)	(0.052)	(0.052)	(0.053)
W*Log(Pesticide Org)	0.152**	0.166**	0.146**	0.166**	0.169**	0.168**
	(0.072)	(0.071)	(0.074)	(0.071)	(0.071)	(0.071)
W*Lamp (0-1)	0.089	0.043	0.106	0.031	0.025	0.027
	(0.136)	(0.136)	(0.142)	(0.136)	(0.135)	(0.136)
W*Chemical influence (0-1)	0.018	0.008	0.02	-.012	0.001	-.003
	(0.126)	(0.125)	(0.126)	(0.125)	(0.125)	(0.126)
Seed dummies	x	x	x	x	x	x
FE	x	x	x	x	x	x
TE	x	x	x	x	x	x
N	990	990	990	990	990	990

N plots	198	198	198	198	198	198
Corrected R2	0.391	0.378	0.379	0.341	0.372	0.359
Log L	561.93	562.21	562.22	558.44	558.51	558.61

Notes: standard errors in parentheses. The dependent variable is the yield defined as the log of raw rice output per land area. The group-structures are defined by the farmer's family (four reference groups) from columns 1 to 3 and by the geographical location of the plot (two reference groups) from columns 4 to 6. *** statistical significance at 1%, ** statistical significance at 5%, * statistical significance at 10%.

References

Anselin L, Le Gallo J. & Jayet H. (2006). Spatial panel econometrics. In: Matyas L, Sevestre P. (Eds). *The econometrics of panel data, fundamentals and recent developments in theory and practice*, 3rd edition. Kluwer, Dordrecht, 901-969.

Elhorst, J.P. (2014). *Spatial Econometric: From Cross-Sectional Data to Spatial Panels*, New-York: Springer.

Elhorst, J.P. (2010). Applied spatial econometrics: Raising the bar, *Spatial Economic Analysis* 5 (1), 9–28.

Elhorst, J.P. (2009). Spatial Panel Data Models. In MM. Fischer & A. Getis (Eds.), in *Handbook of Applied Spatial Analysis*, 377-407.

Lee, L. F., & Yu, J. (2010a). Estimation of spatial autoregressive panel data models with fixed effects, *Journal of Econometrics* 154 (2), 165–185.

Lee, L. F., & Yu, J. (2010b). Some recent developments in spatial panel data models, *Regional Science and Urban Economics*, 40(5), 255-271.

LeSage, J.P., & Pace, R.K. (2014). Interpreting Spatial Econometric Models, in *Handbook of Regional Science*, M. M. Fischer & P. Nijkamp (Eds.) Springer, Berlin, 1535-1552.