SELECTION AND COMPARATIVE ADVANTAGE IN TECHNOLOGY ADOPTION: A RECONSIDERATION

BY JOHN M. ANTLE

May 2018

Abstract

This paper reconsiders Suri's (2011) analysis of hybrid maize seed and fertilizer adoption in Kenya. Using a correlated random coefficient (CRC) production function model, Suri estimated a counterfactual gross *relative* return to hybrid adoption for permanent *non-adopters* averaging over 100 percent, at least 70 percent higher than the estimated average return to hybrid *adopters*. Suri argued this result was explained by high fixed costs of market access, proxied by distance to market, and concluded that policies reducing those costs could increase adoption rates and "increase yields dramatically." This finding contradicts research showing relatively poor performance of modern maize varieties in marginal areas where most permanent non-adopters are located, due to a lack of investment in breeding targeted to such areas. In this paper I reconsider Suri's data and methods, and find that Suri's results are the consequence of two errors. First, a mathematical error in the derivation of Suri's adoption model led to a mis-specified empirical relationship between hybrid returns and distance to market; second, estimates of the CRC model's parameters were biased by the use of data that violate the common support condition required for identification. A corrected CRC analysis based on data stratified by agro-ecological zones, satisfying common support, reverses Suri's results. I also present results from an additive error switching regression model that is computationally simpler and more flexible than the CRC model. This model produces results similar to the CRC model estimated by zone, and also shows that observed adoption behavior may be explained, in part, by the risk attributes of the hybrid seed and fertilizer technology.

1. INTRODUCTION

THE CENTRAL ROLE of agricultural growth in economic growth (World Bank 2007) has led to a rich literature of econometric studies of agricultural technology adoption (Feder, Just and Zilberman 1985; Sunding and Zilberman 2001; Foster and Rosenzweig 2010). In both more and less developed areas of the world with suitable agronomic conditions, technologies such as hybrid maize seed varieties and mineral fertilizers are much more productive and profitable than non-hybrids, and adoption rates are high. Yet, in some areas of the developing world where conditions are less favorable, relatively low rates of adoption of agricultural technologies such as improved seed and mineral fertilizers persist. These less favorable agricultural areas are also populated by some of the poorest, most food insecure people in the world.

This technology under-adoption "puzzle" has led to many studies attempting to explain it. In addition to widely varying agro-ecological conditions including soils and climate, researchers have identified various farm-specific factors affecting technology adoption (e.g., risk attitudes, human capital, financial constraints) as well as external factors such as market access and policies. The World Development Report (World Bank 2007) identifies adverse agro-ecological conditions and limited market access as two key factors constraining technology adoption in regions with low rates of technology adoption.

Underlying the adoption question is the challenge of quantifying the productivity effect of a technology while accounting for the heterogeneous agro-ecological, economic and social conditions typical of small-scale farm households in many parts of the developing world. Suri's analysis (2011) contributed to this literature by showing how a production function could be specified and estimated as a structural correlated random coefficient (CRC) model with panel data to account for unobserved heterogeneity in the presence of technology self-selection (i.e.,

adoption). Consistent with many agronomic and economic studies, Suri's analysis provided evidence of heterogeneous productivity of hybrid and non-hybrid maize varieties. However, Suri's CRC model estimates produced a result that runs counter to conventional agronomic and economic understanding. Suri found that *non-adopters* were estimated to have high counterfactual returns to hybrid maize seed technology *relative* to non-hybrid, averaging over 100 percent, far higher than the returns to farmers that are permanent *adopters*. Suri argued that this finding can be explained by high costs of market access for non-adopters, and has important policy implications (2011, pp. 162-163):

> The estimated mean gross return from my approach is 60%, but some farmers have returns as high as 150%, while there are many who have returns either close to zero or (in some cases) negative. These estimated returns control for input use, but do not account for other costs (such as the costs of accessing the technology) and are therefore gross, rather than fully net returns. The joint distribution of estimated returns and adoption decisions displays some remarkable features….A small group of farmers has extremely high counterfactual returns to hybrid (about 150%), yet they choose not to adopt. This is rather striking and seems to deepen the initial puzzle, but is well explained by supply and infrastructure constraints, such as long distances to seed and fertilizer distributors….The heterogeneity in returns to this technology has important implications for policy…. For example, for farmers who would have high returns but are constrained on the supply side, alleviating their constraints by targeted distribution of inputs and infrastructure improvements could improve yields dramatically.

Thus, if valid and generalizable, Suri's findings would have important implications for development policy, implying an allocation of resources away from agronomic research, such as breeding crop varieties targeted to regions with low adoption rates of existing varieties, towards investments in market infrastructure. However, Suri's results appear to contradict studies of agricultural research investments which show that the performance of modern crop varieties such as hybrid maize depends on their adaptation to local agro-ecological conditions, and that most breeding efforts have been directed toward development of varieties suited to high potential

areas (Evenson and Gollin 2003). Suri's results also seem to contradict analysis by Mathenge, Smale and Olwande (2014) with the same Kenyan data, who find larger impacts of hybrid maize use on indicators of economic welfare in the more productive maize growing regions, and smaller impacts of hybrid use in other regions.

In sections 2-4 of this paper, I reconsider Suri's analysis and its implications, using the same data as Suri from 1997 and 2004, supplemented with data from 2000, 2007 and 2010. The first striking feature of these data is the spatial pattern of hybrid and fertilizer use: in agro-ecological zones favorable to maize production, adoption rates exceed 80 percent, and exceed 90 percent in the most suitable areas. Moreover, consistent with agronomic understanding, the data show that low adoption rates are confined largely to unfavorable areas where productivity of both non-hybrid and hybrid maize varieties is low. The data also show that the majority of permanent non-adopting farms are located in these low-productivity areas, and they are farther from input markets, by about 3 kilometers on average. Could the high returns to hybrid seed and fertilizer estimated by Suri among permanent non-adopters – who are mostly located in these low productivity areas – be explained by this difference in distance to markets?

My analysis provides a negative answer to this question. I show that Suri's findings are the consequence of two errors, and are reversed in a correct analysis. The first error is mathematical and occurs in Suri's theoretical analysis of the adoption decision by farmers. This error leads to an incorrect decision rule for hybrid adoption by a profit maximizer that compares the gross returns to hybrid measured in relative (unit free) terms to the cost of the technology measured in yield units (i.e. kilograms of maize per acre). This error led Suri to correlate the *relative* returns to hybrid with the cost of market access (proxied by distance to market) to explain the spatial variation in hybrid returns. I show that the correct adoption rule compares

gross hybrid returns to the cost of the technology, with both gross returns and costs measured in the same units.

The confounding of units of measurement in Suri's adoption analysis has important consequences for the empirical analysis Suri presented to explain the spatial distributions of hybrid returns. The Kenyan data show that most permanent non-adopters of hybrid are located in low-productivity zones where maize yields for both hybrid and non-hybrid varieties are 50-70 percent lower than in the areas where hybrid is widely adopted. Thus, a high *relative* counterfactual return for non-adopters does not necessarily imply a high return when translated into yield units, and there is no reason for a systematic relationship between relative hybrid returns and the cost of market access across low and high productivity areas.[1]

Consistent with these facts, I find that a correctly formulated analysis does not show a positive or statistically significant relationship between gross hybrid returns and distance to market infrastructure. My replication of Suri's analysis shows that distance variables explain a very small share of the variation in hybrid returns, whether in relative or absolute units, and that agro-ecological zones provide a better explanation for spatial productivity patterns than distance to market. Thus, my replication contradicts Suri's claim that the spatial pattern of hybrid

[1] For example, farm *A* that is 8 km from the market in a low productivity zone could have a *relative* return of 100 percent, whereas farm *B* that is 4 km from the market in a higher productivity zone could have a relative return of 50 percent, implying a positive correlation between gross relative returns and distance. But because farm *A's* non-hybrid yield is very low, say 300 kg/ac, its gross return to hybrid adoption in yield units would be 300 kg/ac; whereas if farm *B* in a high productivity zone has a non-hybrid yield of 1000 kg/ac, its gross return to hybrid adoption would be 500 kg/ac, implying a negative correlation between gross hybrid returns (in yield units) and distance to market.

productivity is "…well explained by supply and infrastructure constraints, such as long distances to seed and fertilizer distributors."[2]

My analysis of the Kenyan data and the CRC model shows that Suri's estimate of a high counterfactual return to non-adopters is explained by biases caused by lack of identification in the data, combined with properties of the structural CRC model. I show that the Kenyan data fail to satisfy the common support or "overlap" condition required for identification of average treatment effects from data (Wooldridge 2010). The lack of common support is due to Suri's use of data pooled across extremely different agro-ecological zones. The data show that virtually all farms that are permanent adopters of hybrid seed and fertilizer are located in agro-ecological zones favorable to maize production, and in these zones there are virtually no permanent non-adopters. Thus, in these zones, a counterfactual return for non-adopters cannot be identified or reliably estimated. Virtually all farms that are permanent non-adopters are located in areas unfavorable to maize production, and in these areas there are virtually no permanent adopters, thus, the returns to permanent adopters cannot be identified or reliably estimated in these areas. Under these conditions, "apparent" structural identification may be possible when the data provide a small number of observations satisfying the common support condition, but too few observations to provide an unbiased or reliable estimate of a causal effect. My analysis shows that the use of an over-identified nonlinear structural form under these conditions can result in an unstable model that is sensitive to the data, specification and estimation method, and can result in large parameter biases and erroneous inferences.[3]

[2] Suri (2011) did not report goodness-of-fit statistics for the regressions on which this claim was based. See section 4 and Table IV for further discussion. The R^2 in the first column of Table IV was obtained in a personal communication.

[3] Similarly, Heckman (2010, p. 357) notes, "There have been many demonstrations of the sensitivity of estimates of structural models to assumptions about functional forms and distributions of unobservables." He cites various

Although the CRC model cannot be identified with the data pooled across agro-ecozones, I show that it can be identified and used to estimate agro-ecozone-specific effects of hybrid adoption. I do this by using transitory non-adopters as controls for permanent adopters in the high and medium productivity zones, and using transitory adopters as controls for permanent non-adopters in the low productivity zone. The CRC models estimated by agro-ecozone show very different patterns of relative hybrid returns from Suri's analysis, with returns to permanent non-adopters lower on average than returns to permanent adopters. Moreover, when gross relative returns are translated into returns in maize yield units (kg/ac), returns to permanent adopters in the medium and high productivity zones are much higher than the returns to permanent non-adopters in the low-productivity zone. The CRC models estimated by zones also show that most permanent adopters earn a gross return to hybrid that exceeds the additional cost of seed and fertilizer including transportation costs, whereas the gross returns to most non-adopters are less than the cost of the technology, as implied by profit maximization.

Interestingly, the results from the zone-specific CRC models also show that some of the farms that are adopting hybrid technology may be earning gross returns that are *less* than the cost of the hybrid seed and fertilizer used. Thus, if there is an adoption "puzzle" among Kenyan maize producers, it is why some farmers with *low* returns to hybrid *are* adopting, rather than why farms with potentially *high* returns *are not* adopting. Using the same data to estimate a yield function, Sheahan, Ariga, and Jayne (2016) similarly found that the return to fertilizer may actually be less than its cost for some Kenyan farmers in the higher-productivity zones.

authors who "…gave early warnings about the fragility of standard econometric estimates of explicit economic models." However, he does not relate this fragility to the application of structural models to data that violate the common support condition.

In sections 4 and 5, I discuss several limitations of the CRC model used by Suri, related to its log-linear functional form and multiplicative error specification, and then present an additive-error switching regression (AES) model that does not impose these restrictions. I show that this model can be identified using observables if two conditions are met: first, that permanent adoption behavior is determined by observables (a condition supported by the data); and second, that transitory adoption (or switching) behavior is determined by factors randomly distributed across farms and independent of technology. Estimating this model for the two years used by Suri (1997 and 2004) as well as for the 5-year panel data for 1997, 2000, 2004, 2007 and 2010, I obtain results consistent with the CRC models estimated by agro-ecozone for 1997 and 2004. I use the additive-error switching regression model to estimate the risk characteristics of the hybrid technology, and find that risk may help explain the observed spatial pattern of adoption across low and high productivity zones.

2. TEGEMEO HOUSEHOLD SURVEY: IMPLICATIONS FOR RESEARCH DESIGN

In this section I describe some key features of the data derived from the Tegemeo Rural Household Surveys, conducted in 1997, 2000, 2004, 2007 and 2010, referred to henceforth as the full panel. The original survey contained over 1500 maize-producing households, but due to attrition and missing values, the full panel contains 1045 households. The balanced 1997 and 2004 sample used by Suri to estimate the CRC model contains 1202 households, although as discussed below, Suri dropped observations from two districts with high HIV rates for the analysis of hybrid returns distributions and distance to market. Although Suri did not describe this "low HIV" sample in detail or identify the number of observations, in my analysis dropping

these two districts resulted in a sample of 1061 households.[4] Due to data sharing restrictions of the Tegemeo Institute, Suri could not provide the data used in her analysis, and did not provide the computer code used to construct the data used for model estimation and analysis, but did provide a description of how the variables were constructed. I was able to reasonably approximate the summary statistics presented by Suri for the 1997 and 2004 samples with 1202 observations. Further details about the data are in Suri (2011), Sheahan, Ariga and Jayne (2016), and the Supplemental Material for this study.

2.1 *Agro-ecological Zones, Yields and Hybrid and Fertilizer Use*

Table 1 presents summary statistics for some key variables, for the full panel and stratified by three agro-ecozones that I refer to as high-productivity, medium productivity and low productivity. The high zone is identified by the Kenyan government as "high maize potential," the medium zone is comprised of somewhat less productive highland areas, and the low zone is comprised of lowland areas with soils and climate not well suited to maize production (see the Supplemental Material for further details). The data show three types of hybrid adoption behavior: farms that always use hybrid seed in the data (permanent adopters, referred to Suri as "hybrid stayers"); farms that never use hybrid (permanent non-adopters, referred to by Suri as "non-hybrid stayers"); and farms that are transitory users of hybrid seed, meaning that they use it in some growing seasons and not in others. As noted in the Introduction, the data show large differences in maize yields between the agro-ecological zones (also see Figure 1). These differences reflect the well-known causal relationship between soils, climate and maize productivity.

[4] This is the same number reported by Suri in a personal communication.

A key feature of the data is the distribution of adopter types across the three zones (Table 1 and Figure 1). Hybrid adoption rates are high in the medium and high zones (80 and 93 percent), and low in the low productivity zone (30 percent). 2 percent of farms in the low zone are permanent hybrid users, compared to 58 percent in the medium zone and about 75 percent in the high zone. 25 percent of the low-zone farms are permanent non-adopters, and this represents 77 percent of all non-adopters. In the medium zone, only 4.5 percent of farms are permanent non-adopters (about 22 percent of all non-adopters), and less than 0.1 percent of farms in the high zone never used hybrid in the five years of data. In contrast, transitory users are observed in all zones, with the largest proportion in the low zone.

In Suri's CRC analysis, transitory users were divided into "joiners" who did not use hybrid in 1997 but did in 2004, and "leavers" who did the opposite. The full panel shows that about 9 percent of the observations are joiners over the 1997-2010 period, and 2 percent of farms were "leavers." Some of the farms classified as "permanent" may be switchers in intervening years, but it seems reasonable to assume that a farm that shows the same behavior over 13 years is exhibiting stable behavior. The joiner and leaver data suggest a modest trend towards increased hybrid use over time, although there is no evidence of a trend in the overall use of hybrid, equal to 66, 68, 60, 68 and 72 percent of farms in the five years.

2.2 *The Importance of Agro-ecological Zones*

In contrast to hybrid use, there is a clear trend in fertilizer use in the data, with 50 percent of farms using fertilizer in 1997, increasing each year to about 70 percent in 2010. Figure 1 illustrates the trend in fertilizer use and yields, by zone and by type of maize variety. This figure also shows the large differences between the low productivity agro-ecozones and the medium and high zones. Figure 1 also shows that in the low zone, hybrid outperforms non-hybrid by a

small absolute amount, although this difference increases in 2010 when fertilizer use also increased. In contrast, in the medium and high zones, hybrid varieties performed much better than non-hybrid, with much higher fertilizer applications that increased over time.

Table II demonstrates the importance of agro-ecological zones using linear probability models for hybrid seed and fertilizer use for the full panel (model estimates are presented in the Supplementary Material), and contrasts the behavior of farmers exhibiting permanent adoption behavior with those exhibiting transitory behavior. These models demonstrate that agro-ecological zone dummy variables explain a larger share of the variation in hybrid and fertilizer use than any other observables. This is particularly true for the sub-sample of farms that show permanent behavior, where the zone dummies alone explain 67 percent of the variation, whereas province dummy variables alone explain only 25 percent, and all covariates together explain 71 percent. It is notable that Suri used province dummies to represent spatial fixed effects in describing the spatial features of the data and in some econometric analysis, but did not use agro-ecological zones.

Production functions for maize yield estimated using the same covariates as the linear probability models confirm the importance of agro-ecological zones as the most important covariate explaining yields (see the Supplementary Material). For example, across all farms, the agro-ecological zone dummies explain the largest share of the variation, 24 percent, compared to 11 percent explained by province dummies and 38 percent explained by all observables. Statistical tests also confirm that the parameter differences between zones are highly significant.

A key issue that will be discussed below is the relationship between productivity and distance to market infrastructure. A regression of yield on distance variables alone confirms a negative correlation, but the production functions in the Supplementary Material show that this

Table I. Selected Summary Statistics for Tegemeo Household Survey Data, 1997, 2000, 2004, 2007 and 2010 (means with standard deviations in parentheses)

		Agro-eco Zone		
	All Farms	Low	Medium	High
Maize yield (kg/ac)	797	394	868	1119
	(655)	(368)	(667)	(661)
Hybrid Use (0/1)	0.693	0.303	0.804	0.927
	(0.461)	(0.460)	(0.397)	(0.260)
Permanent Hybrid (0/1)	0.466	0.020	0.582	0.752
	(0.499)	(0.142)	(0.493)	(0.432)
Perm Non-Hybrid (0/1)	0.093	0.253	0.045	0.004
	(0.291)	(0.435)	(0.207)	(0.062)
Seed (kg/ac)	8.897	7.479	9.069	10.180
	(5.724)	(6.654)	(5.486)	(4.561)
Fertilizer (kg/ac)	40.182	5.553	50.122	60.597
	(63.861)	(16)	(79.121)	(47.876)
Fertilizer Use (0/1)	0.670	0.245	0.822	0.862
	(0.470)	(0.430)	(0.383)	(0.345)
Permanent Fertilizer (0/1)	0.462	0.072	0.572	0.698
	(0.499)	(0.258)	(0.495)	(0.459)
Perm Non-Fertilizer (0/1)	0.191	0.576	0.037	0.0465
	(0.393)	(0.494)	(0.188)	(0.211)
Maize Area (ac)	2.113	2.115	1.168	3.919
	(4.249)	(2.472)	(1.335)	(7.602)
Farm Size (ac)	3.784	3.644	2.853	5.721
	(5.153)	(3.644)	(2.845)	(8.458)
Distance to Fertilizer Market (km)	4.445	7.725	2.692	4.061
	(6.183)	(9.027)	(3.294)	(4.808)
High HIV District (0/1)	0.115	0.409	0.000	0.000
	(0.319)	(0.492)	(0)	(0)
Sasonal Rainfall (mm)	667	491	749	712
	(279)	(259)	(270)	(208)
No. Observations	5225	1470	2465	1290

Note: (0/1) indicates variable equal to 1 if true and 0 otherwise. These data represent the full panel. Results presented in the paper utilize various combinations of the data according to the model and sample specification.

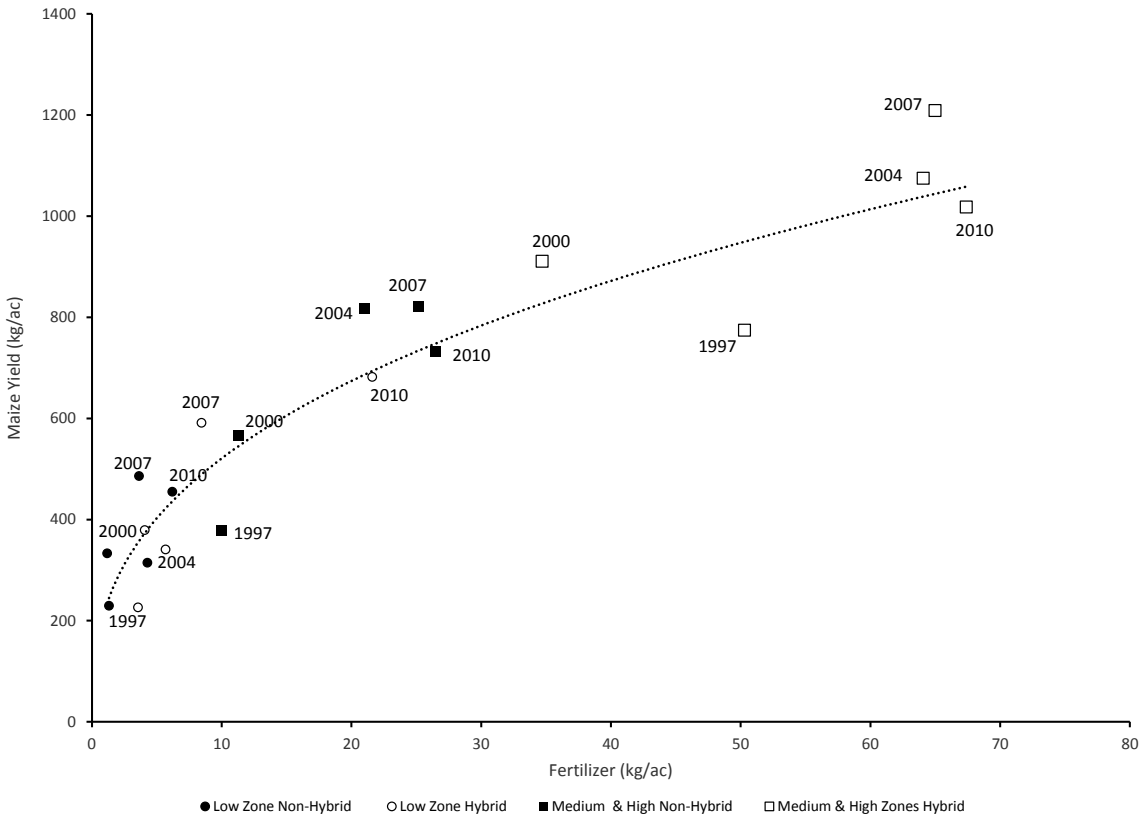

Figure 1. Kenya maize yield and fertilizer use by agro-ecological zone and hybrid use, 1997-2010 full panel. Dashed line is a constant elasticity curve fitted to the data with elasticity 0.37.

Table II. Percent of Variation Explained by Agro-ecozone Dummies, Province Dummies and Other Covariates in Linear Probability Models for Hybrid and Fertilizer Use. All Farms in Full Panel, and Farms with Permanent and Transitory Hybrid Seed and Fertilizer Use

	Hybrid			Fertilizer		
	All	Permanent	Transitory	All	Permanent	Transitory
Zone Dummies	29	67	6	32	37	22
Province Dummies	14	25	4	18	15	20
Other Covariates	25	44	13	31	31	27
All Covariates	36	71	16	42	45	37

Note: Linear probability models were estimated with an intercept and the variables indicated. Values in the table are R^2 statistics from each model.

correlation is weak and largely disappears when other covariates are in the model, suggesting that the causal relationship between distance to market and productivity is weak. In contrast, the agro-ecozone variables remain highly significant in explaining hybrid use, fertilizer use, and yield in the presence of all other covariates including distance to market.

2.3 *Common Support and Covariate Balance*

Table 1 and the other data presented above suggest that selection behavior has a large effect on the pattern of hybrid use and non-use behavior across the agro-ecozones in Kenya. These patterns in turn induce the lack of overlap in the data between treated and control observations that is needed to identify treatment effects of hybrid adoption. Figure 2 illustrates the problem using propensity scores estimated with logistic regressions using the same covariates as in the linear probability models presented in the Supplementary Material. The upper two panels show the lack of overlap for the data pooled across zones for permanent users and non-users (upper left), but with better overlap for transitory users (upper right) due to the fact that a substantial share of farms in each zone exhibit transitory behavior. The lower panels suggest an identification strategy based on stratification of the data that I utilize in the analysis presented below. The lower left panel suggests that the counterfactual hybrid productivity of permanent non-users in the low zone can be identified with transitory hybrid users as controls. The lower right-hand panel indicates that the productivity of permanent hybrid users in the medium and high zones can be identified with transitory non-adopters as controls.

Analysis of covariate balance for the full panel using standardized differences confirms that balance is poor for a number of variables, particularly for climatic variables (mean temperature and rainfall) that are associated with agro-ecological conditions, for agronomic factors such as the use of inter-cropping, and for other activities on the farm such as dairy

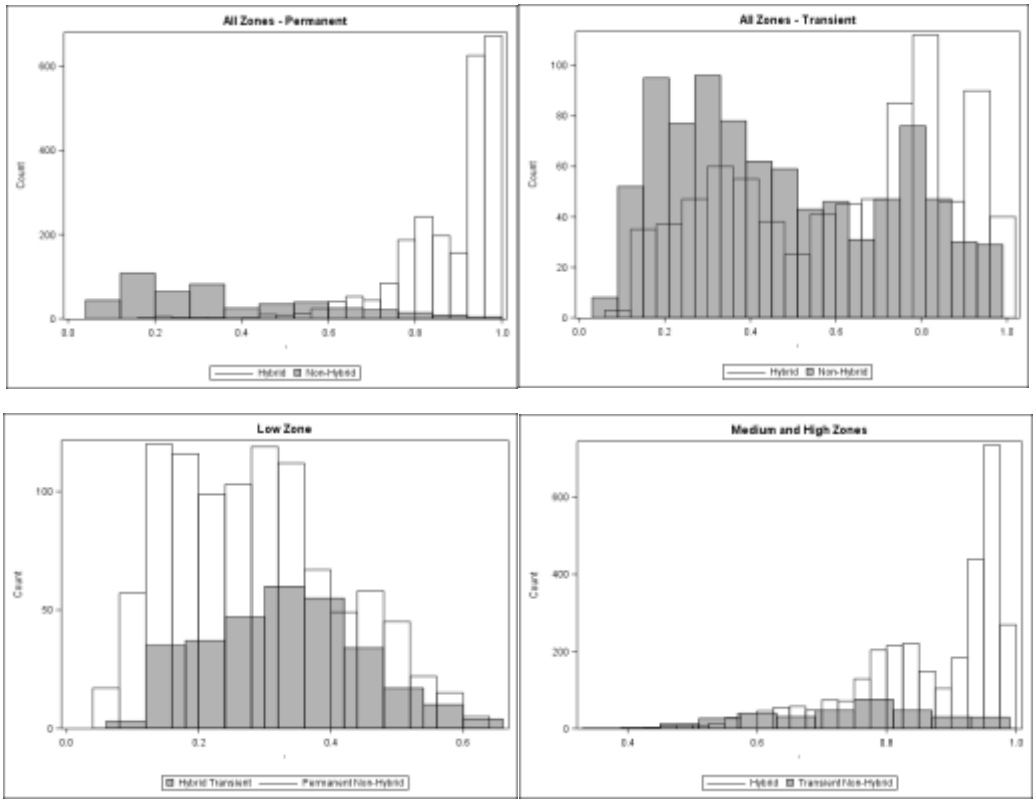

Figure 2. Histograms of Propensity Scores for Hybrid Use, 1997-2010 Full Panel

production that are likely to interact with maize production (e.g., by making manure available for use as an organic fertilizer). Stratification clearly improves balance, and propensity score matching further reduces standardized differences for most covariates to less than 0.25.

3. THE CRC PRODUCTION MODEL

In this section I briefly summarize the derivation of the CRC production model developed by Suri (2011) and then use it to demonstrate the issues that arise in Suri's interpretation of the model for technology adoption, the identification issues created by the use of data violating common support, and how the model can be identified and estimated with data stratified by agro-ecozone. I use (S*x*) to denote equation *x* of Suri (2011). The model is:

(S8) $$y_{it}^{H} = \beta_{t}^{H} + x_{it}'\gamma^{H} + u_{it}^{H}$$

(S9) $y_{it}^{N} = \beta_t^{N} + x_{it}'\gamma^{N} + u_{it}^{N}$

where y_{it}^{s} is the log of maize yield for production system *s* = *H* (hybrid), *N* (non-hybrid), x_{it} is a vector of inputs and other covariates in logs, and u_{it}^{s} are random errors.[5]

The error terms are further decomposed as $u_{it}^{H} = \theta_i^{H} + \xi_{it}^{H}$ and $u_{it}^{N} = \theta_i^{N} + \xi_{it}^{N}$. The individual productivity components are assumed to follow $\theta_i^{k} = b_k\left(\theta_i^{H} - \theta_i^{N}\right) + \tau_i$, *k* = *H*,*N*, where τ_i is an individual-specific productivity component that does not vary by system. The parameter $\theta_i \equiv b_N(\theta_i^{H} - \theta_i^{N})$ is interpreted as the farmer's comparative advantage in using hybrid. Defining $\phi \equiv \frac{b_H}{b_N} - 1$, it follows that $\theta_i^{H} = (1+\phi)\theta_i + \tau_i$ and $\theta_i^{N} = \theta_i + \tau_i$. The sign of ϕ indicates selection on gain, with $\phi < 0$ indicating negative selection, i.e., farms with low initial productivity have highest gain from adoption. Combining with (S8) and (S9) gives[6]:

(S17) $y_{it}^{H} = \beta_t^{H} + \tau_i + (\phi + 1)\theta_i + x_{it}'\gamma^{H} + \xi_{it}^{H}$

(S18) $y_{it}^{N} = \beta_t^{N} + \tau_i + \theta_i + x_{it}'\gamma^{N} + \xi_{it}^{N}$

Defining h_{it} as an indicator variable equal to 1 for hybrid and zero for non-hybrid, the log of yield can be expressed as $y_{it} = h_{it}\, y_{it}^{H} + (1 - h_{it})y_{it}^{N}$. Combining this equation with the production functions for hybrid and non-hybrid gives Suri's estimation model

(S20) $y_{it} = \beta_t^{N} + \theta_i + (\beta_t + \phi\theta_i)h_{it} + x_{it}'\gamma^{N} + h_{it}x_{it}'\gamma + u_{it}$,

where $\beta_t \equiv \beta_t^{H} - \beta_t^{N}$, $\gamma \equiv \gamma^{H} - \gamma^{N}$, $u_{it} \equiv \tau_i + h_{it}\,\xi_{it}^{H} + (1 - h_{it})\xi_{it}^{H}$, $\theta_i \equiv \theta_i^{H} - \theta_i^{N}$.

[5] As discussed below, Suri actually estimated the CRC model with inputs in non-log form, apparently to accommodate the estimation method used and the presence of zeros in the data. See section 4.4 below.

[6] Note Suri's equations (S17) and (S18) as printed contain a typographical error, as they show the inputs in non-log form X_{it}, whereas they should be in log form x_{it}' (the same error appears in equation S20).

3.1. *Quantifying the Returns to Hybrid Adoption*

Suri refers to $\beta_t + \phi\theta_i$ as an individual-specific estimate of the "gross return to hybrid" under the assumption that the production function parameters for hybrid and non-hybrid are the same, implying $\gamma = 0$. From (S17) and (S18),

(1) $\quad E[y_{it}^H - y_{it}^N] = \beta_t + \phi\theta_i + x_{it}'\gamma.$

Thus $\beta_t + \phi\theta_i$ is equal to the difference in *log yield* between hybrid and non-hybrid when $\gamma = 0$. In this section I show that $\beta_t + \phi\theta_i$ can be interpreted as an approximate *relative* or unit-free measure of hybrid productivity, under the additional assumption that the technology shocks to the hybrid and non-hybrid systems follow the same distribution.

Using (S17) and (S18), expected yields are

(2) $\quad E[Y_{it}^H] = E[\exp(y_{it}^H)] = \exp(\beta_t^H + \tau_i + (\phi + 1)\theta_i + x_{it}'\gamma^H)E[\exp(\xi_{it}^H)]$

(3) $\quad E[Y_{it}^N] = E[\exp(y_{it}^N)] = \exp(\beta_t^N + \tau_i + \theta_i + x_{it}'\gamma^N)E[\exp(\xi_{it}^N)]$

Noting that h_{it} is a discrete variable, the percentage change in expected yield with respect to a change in h_{it} from zero to 1 is thus:

(4) $\quad r_{it}^H \equiv \frac{E[Y_{it}^H] - E[Y_{it}^N]}{E[Y_{it}^N]} = \exp(\beta_t + \phi\theta_i + x_{it}'\gamma)\, E[\exp(\xi_{it}^H)]/E[\exp(\xi_{it}^N)] - 1.$

For $\gamma = 0$, rearranging (4) gives

(5) $\quad \beta_t + \phi\theta_i = \ln(r_{it}^H + 1) + \ln(E[\exp(\xi_{it}^N)]) - \ln(E[\exp(\xi_{it}^H)]).$

Thus, under these assumption that the ξ_{it}^s follow the same distributions, we have $\beta_t + \phi\theta_i = \ln(r_{it}^H + 1)$ which is approximately equal to r_{it}^H for small values of r_{it}^H, and we can interpret

$\beta_t + \phi\theta_i$ as an approximate unit-free measure of the *relative* productivity effect of hybrid on yield. Using (4), an appropriate measure of the hybrid returns *in yield units* is $r_{it}^H \; E[Y_{it}^N]$.

3.2. *Analysis of the Adoption Decision*

Define p_{it} as the expected maize output price; the price of hybrid and non-hybrid seed is b_t and c_{it}; a_{it} is the fixed cost of acquiring hybrid seed and fertilizer (e.g., transportation cost); s_{it} is the quantity of seed (assumed by Suri to be the same for hybrid and non-hybrid); and w_{jit} is the price of other inputs X_{jit}. Also define $A_{it} \equiv a_{it}/p_{it}$ and $\Delta_{it}^s \equiv (b_t - c_{it})s_{it}^*/p_{it}$, which are measures of fixed transportation cost and cost of seed at the profit maximum, *both measured in units of maize yield*. Denoting profit-maximizing values with an asterisk, the hybrid technology is adopted if:

$$\text{(S4)} \quad \left(Y_{it}^{*H} - \sum_{j=1}^{J} \frac{w_{jit}}{p_{it}} X_{jit}^{*H}\right) - \left(Y_{it}^{*N} - \sum_{j=1}^{J} \frac{w_{jit}}{p_{it}} X_{jit}^{*N}\right) > A_{it} + \Delta_{it}^s$$

Suri additionally assumed: $\gamma^H \approx \gamma^N$ and $X_{it}^{*H} \approx X_{it}^{*N}$ (except for fertilizer used with hybrid seed); and that fertilizer is used in fixed proportions to hybrid seed and can be incorporated into Δ_{it}^s. Under these assumptions, it follows that hybrid is used if:

$$\text{(S5)} \quad Y_{it}^{*H} - Y_{it}^{*N} > A_{it} + \Delta_{it}^s.$$

Importantly, (S5) is expressed in *yield units* on both sides of the inequality.

In section 4.3, Suri (2011) attempts to manipulate (S5) to derive a relationship between the individual-specific measure of productivity, θ_i, and the cost of acquiring the technology given on the right-hand side of (S5). Suri (2011, page 179) asserts the following:

"Rewriting (S4) in log output and using (S8) and (S9), a farmer uses hybrid if

(S21) $E(u_{it}^{H} - u_{it}^{N}) > A_{it} + \Delta_{it}^{s} + \beta_{t}^{H} - \beta_{t}^{N} + \sum_{j=1}^{J}(\gamma_{j}^{N} X_{jit}^{*N} - \gamma_{j}^{H} X_{jit}^{*H})$"

Given the assumptions that $\gamma^{H} \approx \gamma^{N}$ and $X_{it}^{*H} \approx X_{it}^{*N}$, the last term on the right-hand side of (S21) is zero, leading to:

(S24) $\phi\theta_{i} > (A_{it} + \Delta_{it}^{s}) - (\beta_{t}^{H} - \beta_{t}^{N})$,

However, as demonstrated by equation (1), the left-hand side of (S24) and $(\beta_{t}^{H} - \beta_{t}^{N})$ are defined in *log yield units*, whereas the two terms A_{it} and Δ_{it}^{s} on the right-hand side of (S24) are defined in *yield units*, as is apparent from (S5). Thus equation (S24) relates variables in incommensurable units and thus is erroneous.

The error in (S24) is due to the fact that in equation (S21) Suri equated the terms Y_{it}^{*H} and Y_{it}^{*N} in (S5) with expectations of the logs of yield, y_{it}^{H} and y_{it}^{N}, whereas they should be equated with $Y_{it}^{*H} = E[Y_{it}^{H}] = \mathrm{E}[\exp(y_{it}^{H})]$ and $Y_{it}^{*N} = E[Y_{it}^{N}] = \mathrm{E}[\exp(y_{it}^{N})]$. Using (S5), the correct condition for hybrid adoption, under the assumptions noted above, is:

(6) $Y_{it}^{*H} - Y_{it}^{*N} = E[Y_{it}^{H}] - E[Y_{it}^{N}] = \left(\frac{E[Y_{it}^{H}]}{E[Y_{it}^{N}]} - 1\right) E[Y_{it}^{N}] = r_{it}^{H} E[Y_{it}^{N}] > (A_{it} + \Delta_{it}^{s})$.

Alternatively, using the assumptions discussed below equation (5), it follows that $\beta_{t} + \phi\theta_{i} \approx r_{it}^{H}$ and the adoption condition can be approximated in relative terms by:

(7) $\beta_{t} + \phi\theta_{i} > (A_{it} + \Delta_{it}^{s}) / E[Y_{it}^{N}]$,

or in yield units by:

(8) $(\beta_{t} + \phi\theta_{i}) E[Y_{it}^{N}] > (A_{it} + \Delta_{it}^{s})$.

Thus, (7) and (8) show that the site-specific productivity of the non-hybrid technology $E[Y_{it}^N]$ must be taken into account to relate $\beta_t + \phi\theta_i$ to the cost of the technology.

Suri (2011) wrote the fixed cost term in (S24) as $A_{it} = \alpha_i + \vartheta_{it}$, to obtain

(S25) $\phi\theta_i + \alpha_i > \Delta_{it}^s - (\beta_t^H - \beta_t^N) + \vartheta_{it}$

Equation (S25) is the basis for Suri's regressions of θ_i on distance to market and other variables discussed in section 4 below. Observe that the two terms on the left-hand side of (S25) are incommensurate, because $\phi\theta_i$ is unit-free whereas α_i is in maize yield units. Equation (8) shows that the correct relationship has $\phi\theta_i E[Y_{it}^N] + \alpha_i$ on the left-hand side and $E[Y_{it}^N]$ multiplying $(\beta_t^H - \beta_t^N)$ on the right-hand side. It follows that the correlation between θ_i and a proxy for the fixed cost of accessing markets, such as the distance to market, would confound cost of access with the effects of site-specific productivity represented by $E[Y_{it}^N]$.

3.3. *Identification and Estimation of the CRC Model*

Equation (S20) is a CRC model because the effect of hybrid on yield, $(\beta_t + \phi\theta_i)$, is random across farms indexed by *i*. Estimation must account for the expected correlation between θ_i and u_{it}. Suri used the Chamberlin (1984) method of expressing θ_i as a linear-in-parameters function of the histories of hybrid use h_{it}, fertilizer f_{it} and their interactions. For a two-period model this leads to the specification of θ_i as:

(S35) $\theta_i = \lambda_0 + \lambda_1 h_{i1} + \lambda_2 h_{i2} + \lambda_3 h_{i1} h_{i2} + \lambda_4 h_{i1} f_{i1} + \lambda_5 h_{i2} f_{i1} + \lambda_6 h_{i1} h_{i2} f_{i1} + \lambda_7 h_{i1} f_{i2} +$
$\lambda_8 h_{i2} f_{i2} + \lambda_9 h_{i1} h_{i2} f_{i2} + \lambda_{10} f_{i1} + \lambda_{11} f_{i2}$

Suri also imposed the normalization $\sum_i \theta_i = 0$ and assumed that β_t did not vary with time. Note also that in order to satisfy the orthogonality of the projection of θ_i with respect to the model's

error term, the projection must be linear in the parameters. Further, since fertilizer takes on values of zero, this requires that the production function be specified with fertilizer quantities rather than the logs of fertilizer, so the production function is not concave in inputs. I return to this functional form issue in section 4.4.

In this two-period formulation, permanent non-hybrid farms (PN) are identified by h_{i1}= $h_{i2} = 0$, permanent hybrid adopters (PH) have h_{i1}= $h_{i2} = 1$, leavers (L) have h_{i1}= 1, $h_{i2} = 0$, joiners (J) have h_{i1}= 0, $h_{i2} = 1$; switchers are the combination of joiners and leavers.[7] Using these definitions,

$$\theta_i^{PN} = \lambda_0 + \lambda_{10} f_{i1} + \lambda_{11} f_{i2}$$
$$\theta_i^{PH} = \lambda_0 + \lambda_1 + \lambda_2 + \lambda_3 + (\lambda_4 + \lambda_5 + \lambda_6 + \lambda_{10}) f_{i1} + (\lambda_7 + \lambda_8 + \lambda_9 + \lambda_{11}) f_{i2}\gamma$$
(9)
$$\theta_i^{L} = \lambda_0 + \lambda_1 + (\lambda_4 + \lambda_{10}) f_{i1} + (\lambda_7 + \lambda_{11}) f_{i2}$$
$$\theta_i^{J} = \lambda_0 + \lambda_2 + (\lambda_5 + \lambda_{10}) f_{i1} + (\lambda_8 + \lambda_{11}) f_{i2}$$

As noted above, total effect of hybrid on yield is $\beta + \phi\theta_i$, thus relative returns to the four types of adoption behavior are

(10) $r_{it}^j \equiv \beta + \phi\theta_i^j$ for j = PN, PH, J, L.

Two aspects of identification must be addressed, the common support condition for identification from the data, and the identification of the structural parameters β, ϕ, and the λ_j, j =0, 1, …, 11. As discussed in section 1, the data show that there are few observations of permanent adopters in the low zone, thus θ_i^{PH} and r_i^{PH} cannot be identified from the data in the

[7] The data also show that farmers are stayers, joiners and leavers in terms of fertilizer use. Accounting for this fact would further complicate the CRC model specification.

low zone. Likewise, the data show that there are few observations of permanent non-adopters in the medium and high zones, thus θ_i^{PN} and r_i^{PN} cannot be identified in the medium and high zones. We can also see that the common support condition is related to the conditions for structural identification. From (9), to identify r_i^{PH} it must be possible to estimate λ_3 which requires a sufficient number of observations of permanent adopters. Thus, in the low zone where there are almost no observations of farms using hybrid in both periods, λ_3 cannot be reliably estimated and r_i^{PH}cannot be identified. Similar logic implies that λ_0 cannot be identified in the medium or high zones because there are virtually no observations where $h_{i1}= h_{i2} = 0$. Thus the intercept λ_0 of the projection (S35) cannot be identified or reliably estimated separately from $\lambda_1$and λ_2 in the medium and high zones.

We can conclude that the consequence of attempting to estimate the model with data pooled across zones will be to produce biased parameter estimates that are likely to be sensitive to the data sample and model specification. As shown in Section 2, pooling across zones means that observations of transitory and permanent adopters in the medium and high zones are used to represent the counterfactual yield potential of hybrid for observations in the low zone, thus imparting an upward bias to the estimate of r_i^{PN}. Since fertilizer use is very low or zero among permanent non-adopters (Table 1), it follows from (9) that $r_i^{PN} \approx \beta + \phi\lambda_0$, and for the case of negative selection with $\phi < 0$, an upward bias in r_i^{PN} implies a downward bias in λ_0 and an upward bias in β. Similarly, pooling the data across zones suggests that r_i^{PH}would be over-estimated because yields of non-adopters in the low zone are used to estimate the counterfactual for adopters in the medium and high zones. However, this bias effect is diluted by the relatively large number of transitory non-adopters in the medium and high zones. Due to the normalization $\sum_i \theta_i = 0$, the overall effect of pooling across agro-ecozones is likely to be an upward bias in

r_i^{PN} and a downward bias in r_i^{PH}. Similar logic indicates a possible upward bias in r_i^{L} and a downward bias in r_i^{J}.

One strategy to overcome the bias problem caused by pooling the data across heterogeneous regions is to incorporate a set of spatial dummy variables. However, in the CRC model such fixed effects cannot be distinguished from the θ_i, and would not resolve the identification problems due to lack of common support. Instead, the research design strategy suggested by the data presented in Section 2 is to stratify the data by zone. As suggested by Suri (2011, section 4.4.3), if transitory hybrid use is due to factors uncorrelated with productivity, such as lack of availability in the local market at the time farmers need to purchase seed and fertilizer, observations of transitory hybrid use or non-use can be used as control observations. Data from the low zone, where most permanent non-hybrid farms and many transitory adopters are located, can be used to estimate r_i^{PN}, but r_i^{PH} cannot be estimated in the low zone because it depends on λ_3 which cannot be identified in the low zone (see equation 9). Likewise, data from the medium and high zones, where most permanent hybrid farms as well as transitory non-adopters are located, can be used to estimate λ_3 and r_i^{PH}; however r_i^{PN} cannot be estimated for the medium and high zones because λ_0 cannot be identified.

4. CRC MODEL RESULTS

This section presents a replication of the results for the CRC model highlighted in Suri (2011) sections 6 and 7. Then I present estimates of the CRC model estimated by agro-ecozone, and compare the parameters and hybrid returns distributions produced by Suri's specification and the zone-specific model.

4.1. *Replication of Suri's Results*

Suri argued that high HIV rates in some areas could affect the identification of the model, representing "shocks" realized before the hybrid and fertilizer choices were made, and argued that the results were "stronger" when observations from two high-HIV districts were dropped from the data. Below I discuss the legitimacy of this procedure and its effect on the results.

Table III presents Suri's parameter estimates for β and ϕ as well as my replicated estimates for these parameters using the low-HIV sample. I also include λ_0 in the table because of its importance to the estimation of returns to hybrid non-adopters as explained in section 3 (Suri did not present the estimates of λ_0 or any other structural parameters in equation S35). Table III also shows summary statistics for the distributions of the returns to hybrid in log yield units and in yield units. Like Suri, I estimated the reduced form using seemingly unrelated regression. For estimation of the structural parameters, I used a two-step procedure implemented with the GMM procedure in SAS 9.4, similar to Suri's Optimal Minimum Distance estimation procedure.

Since my data for yields and hybrid use are virtually identical to Suri's, as a first step in the replication I closely reproduced Suri's estimates of the model without covariates and with only hybrid endogenous (Suri 2011, Table VIIIA). Next, I replicated the model used by Suri to analyze the hybrid returns distributions, i.e, the model with both hybrid and fertilizer endogenous, with covariates but without covariate interactions with hybrid (Suri 2011, Table VIIIC). As my Table III shows, my estimates of β and ϕ are very close to Suri's, the average values for the relative hybrid returns are similar, and the distributions of θ_i by adopter type are similar to the ones presented by Suri (2011, Figure 5C). Thus, I conclude that I have successfully

replicated Suri's CRC estimates based on the low-HIV sample and the model without covariate interactions.

As discussed in section 3.2, Suri argued that the spatial distribution of hybrid returns across farms could be explained by permanent fixed costs of market access, proxied by distance to market and related infrastructure (see S25). The first column of Table IV presents one of Suri's regressions of the estimated θ_i on variables for distance to markets and transport infrastructure, credit use and province dummy variables (Suri 2011, Table IX, column 4). The second column presents my replication of this regression carried out with θ_i derived from the replicated model presented in the second column of Table 3. Note that the parameter ϕ in Table III is negative, and relative hybrid returns are calculated as $\beta + \phi\theta_i$ (see equation 12). Thus, the sign of the regression coefficients in the first two columns of Table III have signs opposite of the implied effect of each variable on relative hybrid returns. The replicated regression shows a similar negative parameter for the one statistically significant distance variable, distance to fertilizer market, implying a *positive* relationship between this variable and relative hybrid productivity. Other parameters are somewhat different between Suri's and the replication but all are statistically insignificant. As discussed in section 3.2, this regression does not follow logically from the correct analysis of adoption under profit maximization, so its interpretation is a moot issue. My replication also shows that the distance variables explain a very small amount of the variation in relative hybrid productivity, and thus do not provide a satisfactory explanation of the spatial pattern of relative hybrid productivity.[8]

[8] Suri (2011) did not report R^2 statistics for the θ_i regressions. In a personal communication Suri indicated they were all in the 0.13 range.

Table IV (column 3) shows a regression of hybrid returns in yield units, regressed on the same variables. As argued in section 3.2, this is the regression implied by the correct adoption rule. The R^2 statistic falls to 0.025, and only the distance to extension variable is significant, and its sign implies a *negative* effect of distance on hybrid returns. Thus, these results confirm the bias caused by the mis-specification of the regressions in the first two columns which attempt to relate relative returns to distance.

A more direct test of Suri's explanation for the high counterfactual returns to hybrid for permanent non-adopters is to estimate the cost of hybrid seed and fertilizer to farmers, inclusive of transport cost. The data show that the cost of hybrid seed and additional fertilizer associated with hybrid use averages about 100 kg/ac in terms of maize yield (see Figure 3 which shows the observed distribution of hybrid seed and fertilizer cost). Sheahan (2011, p. 76) estimated that the transport cost ranges from 25 to 50 percent of the fertilizer cost at point of sale. I obtained a similar range of cost by utilizing the transportation cost for fertilizer reported by farmers in the 2004 and 2010 surveys. Taken together, these data imply a purchase cost plus transport cost in maize yield units of less than 200 kg/ac. We can conclude from Table III that Suri's estimated relative returns to hybrid for non-adopters, which imply an average gross return in maize yield over 500 kg/ac, are far higher than the purchase cost plus transportation cost of seed and fertilizer for most farmers. Thus, after accounting for the cost of accessing the technology, Suri's estimate of the counterfactual return to non-adopters is inconsistent with profit maximization after accounting for the cost of market access.

4.2. *Estimation by Agro-ecozone*

Table III also presents parameter estimates and hybrid return distribution statistics for the CRC model estimated by agro-ecozone; also see the kernel density estimates in Figure 3.Table III

Table III. CRC Parameter Estimates and Gross Hybrid Returns Distributions, 1997 and 2004

	Low HIV Sample, All Zones		Estimation by Zone		
	Suri	Replication	Low Zone	Medium Zone	High Zone
λ_0	NA	-0.318	-0.104	0.085	0.316
	(NA)	(0.074)	(0.077)	(0.099)	(0.125)
β	0.603	0.611	0.203	0.28	0.312
	(0.060)	(0.054)	(0.155)	(0.092)	(0.114)
ϕ	-1.788	-1.757	0.117	-1.634	-1.658
	(0.277)	(0.264)	(0.603)	(0.249)	(0.193)
No. Observations	1061	1061	301	451	343
	Gross Relative Hybrid Returns Distributions (100 x log kg/ac) (means with standard deviations in parentheses)				
Permanent Non-Hybrid	106	99	19	NA	NA
	(24)	(33)	(1)	NA	NA
Permanent Hybrid	62	62	NA	34	40
	(5)	(20)	NA	(20)	(19)
Leavers	55	50	23	9	39
	(16)	(14)	(2)	(20)	(48)
Joiners	-25	-27	22	4	-53
	(24)	(16)	(2)	(46)	(8)
	Gross Hybrid Returns Distributions in Yield Units (kg/ac) (means with standard deviations in parentheses)				
Permanent Non-Hybrid	NA	542	49	NA	NA
	(NA)	(288)	(29)	NA	NA
Permanent Hybrid	NA	738	NA	362	548
	(NA)	(1008)	NA	(410)	(333)
Leavers	NA	515	59	50	600
	(NA)	(1243)	(19)	(134)	(945)
Joiners	NA	-234	79	44	-502
	(NA)	(240)	(70)	(240)	(100)

Note: Data for Relative Hybrid Returns in the first column were calculated from data provided by Suri (personal communication). Data show small numbers of observations and outliers in High Zone for Leavers and Joiners.

Table IV. Regressions of Theta (θ_i) and Gross Hybrid Returns (kg/ac) on Distance to Market and Other Variables

	Theta (x 100)		Hybrid Returns (normalized kg/ac x 100)				
	Suri CRC	CRC Replication	CRC Replication	CRC by Zone 1997 & 2004	CRC by Zone 1997 & 2004	AES by Zone 1997-2010	AES by Zone 1997-2010
Distance to Fertilizer Market	-0.285	-0.363	0.259	-0.792	-0.014	-1.055	0.493
	(0.121)	(0.142)	(0.667)	(0.785)	(0.759)	(0.183)	(0.156)
Distance to Motorable Road	-0.898	-0.615	-0.996	-2.338	0.337	0.391	0.911
	(0.501)	(0.568)	(1.922)	(2.210)	(2.171)	(0.863)	(0.714)
Distance to Matatu	-0.028	-0.101	0.813	-0.477	-0.423	NA	NA
	(0.299)	(0.336)	(1.360)	(1.632)	(1.622)	NA	NA
Distance to Extension	-0.061	0.120	-1.659	-0.133	0.205	-1.853	-0.873
	(0.155)	(0.170)	(0.571)	(0.836)	(0.806)	(0.252)	(0.229)
Credit Used	-0.470	-1.093	4.959	13.813	5.168	9.431	0.436
	(1.540)	(1.726)	(6.845)	(8.176)	(8.584)	(3.090)	(2.803)
Low Zone					-84.422		-139.024
					(14.724)		(4.070)
High Zone					25.242		35.586
					(15.651)		(6.406)
R2	0.130	0.095	0.025	0.125	0.154	0.142	0.274
No. Observations	1057	1058	1058	941	941	5014	5014

Note: Standard errors in parentheses. Suri's estimates in the first column were based on ordinary least squares regression; R2 statistics and number of observations were not reported in Suri (2011) and were obtained in personal communication. All other results based on heteroskedastic-consistent regression. Suri scaled distance variables by 100 and credit by 10, and dropped observations with distance greater than 70 km. Here, Theta was scaled by 100. CRC regressions include dummy variables for five provinces and for head of household education as defined by Suri. Hybrid Returns regressions have dependent variable defined as returns in kg/ac minus its mean divided by its standard deviation and multiplied times 100. AES regressions do not include education dummies and distance to matatu (public transport) due to data availability.

shows that the parameter estimates and hybrid returns distributions differ substantially from those estimated with the data pooled across zones. The parameters estimated by zone confirm the analysis of section 3.2 which showed that the lack of common support in the data pooled across zones would bias the estimate of λ_0 downwards and the estimate of β upwards, and result in an upward bias in the estimated counterfactual returns to hybrid for permanent non-adopters. For example, with the data pooled across zones, λ_0 is estimated to be –0.318 and significant, whereas the estimate of λ_0 for the high zone is +0.316 and significant. The estimate of β from the pooled model is 0.611 and highly significant; the estimates for the medium and high zones are 0.28 and 0.31 and significant.

Table III shows that the mean relative hybrid returns are in the range of 20 to 40 percent for the zone-based CRC models. These mean returns are similar to the parameters on the hybrid use dummies in the zone-specific production function estimates presented in the Supplemental Material, showing that with appropriate stratification, conventional production functions provide a reasonable estimate of mean hybrid productivity. Table 3 and Figure 3 show the estimated hybrid returns (in yield units) are much higher for permanent adopters than for permanent non-adopters and for transitory users. Figure 3 also shows an estimate of the distribution of the cost of hybrid seed and fertilizer, which averages about 108 kg/ac (in maize yield units) (see the Supplementary Data for details). As noted above, combining these data with Sheahan's (2011) estimate of transport costs, the cost of the hybrid-seed technology to farmers is estimated to be less than 200 kg/ac in maize units. We can conclude that the CRC model estimated by zone implies high positive net returns to most permanent adopters in the medium and high zones, but negative returns to most non-adopters in the low zone. Thus, these estimates are consistent with adoption based on profit maximization.

Table III also shows relative returns are low for Leavers and Joiners in the low and medium zones. In the high zone, average returns are high for Leavers and large and negative for Joiners, seemingly inconsistent with the other zones. These results are similar to the data for the model estimated with the data pooled across zones, and appear to be unreliable due to small numbers of observations and some large outliers.

Table IV (columns 4 and 5) presents regressions of hybrid returns in yield units from the zone-based CRC models, using the same covariates as the first two columns, as well as dummies for the low and high zones. Recall from section 3 that this regression is consistent with the corrected adoption model, because it relates hybrid returns in yield units to the cost of access

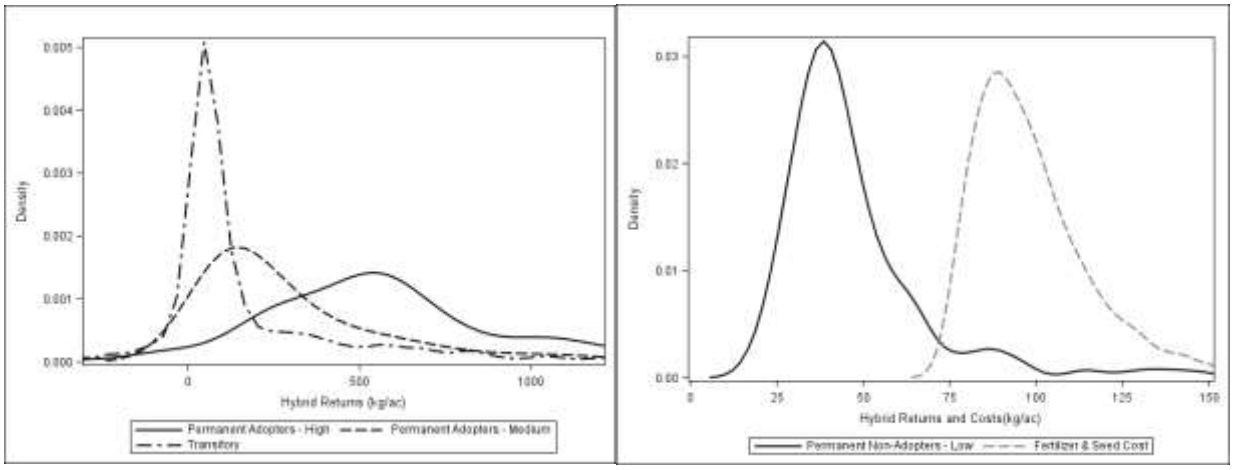

Figure 3. Hybrid returns distributions in maize units by adopter type, CRC model without covariate interactions, estimated by zone, 1997 and 2004 sample. Note: Transitory is both Joiners and Leavers. Right panel also shows the distribution of additional fertilizer and seed cost associated with hybrid technology, as described in the Supplemental Material.

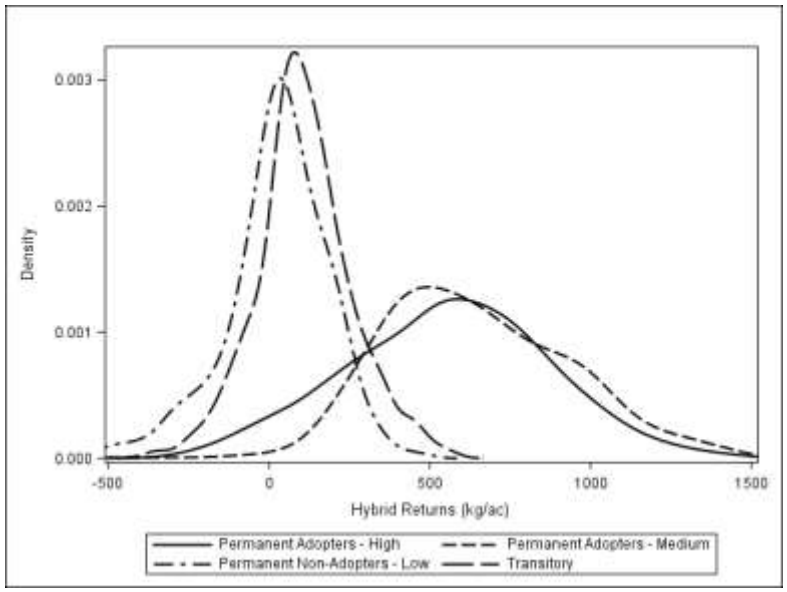

Figure 4. Hybrid returns distributions in maize units, AES model, estimated by zone, 1997-2010 sample. Means with standard deviations in parentheses: High 555 (313); Medium 515 (301); Low 29 (156); Transitory 109 (149).

proxied by distance to market. Without the zone dummies, all of the distance variables are *negatively* related to hybrid returns in yield units; inclusion of the zone dummies causes the distance to fertilizer market to become insignificant and others to switch sign. The low zone dummy is highly significant, again suggesting that the correlations between distance and hybrid returns are not causal, and that hybrid returns are better explained by agro-ecology.

4.3. *Sensitivity of the CRC Model to Data and Specification*

As discussed in Section 3.2, the identification problems associated with the data pooled across agro-ecozones is likely to give rise to estimation problems, particularly for an over-identified non-linear model. These identification problems explain the extreme sensitivity of the CRC model to data and specification evident in Suri's results.

As noted above, Suri's main results were based on a sample that excluded two districts with high HIV rates. However, as Table 1 shows, *all* of the observations Suri classified as "high HIV" were located in areas classified in the low productivity agro-ecological zone. These two districts represent about 40 percent of the observations in the low zone. Thus an alternative explanation for the "stronger" results associated with the "low-HIV" sample pooled across zones is that reducing the number of observations in the low zone amplifies the biases caused by lack of common support in the data. As Suri's (2011) Table VIIIC shows, there is a large difference between the parameter estimates based on the full sample and the low-HIV sample. With the full sample including the high HIV districts, Suri's estimate of the average hybrid productivity without covariate interactions, the parameter β, is 0.088 and not statistically significant, implying an implausible zero average return to the hybrid technology; excluding the high HIV districts the estimate increases to 0.603.

Suri's estimate of the parameter ϕ increased from a value of -0.449 in the full sample to a value of -1.788 with the high HIV districts deleted, implying a much larger degree of heterogeneity in the hybrid return distribution, even though eliminating the low HIV districts *reduces* heterogeneity. Other parameter estimates presented by Suri also show extreme instability, e.g., Suri's Table IIIA show estimates of ϕ that range from -0.794 to -17.82, and not surprisingly are most implausible when the model includes interactions with hybrid which induce high multicollinearity. In my attempts to estimate the model with various samples and specifications, I found that the GMM estimator often would not converge, or would converge to implausible values for β, ϕ and λ_0. This instability is consistent with the identification problem caused by the lack of observations needed to identify β, ϕ and λ_0 in the medium and high zones, which represent about 82 percent of the sample without the high HIV districts.

4.4 Other Limitations of the Log-linear CRC Model

The linear-in-parameters form required for the Chamberlin (1984) estimation procedure imposes strong restrictions on the form of the production function. The conventional log-transformed constant elasticity (or Cobb-Douglas) production function is linear in the parameters, but not linear in the variables. Suri estimated the CRC model using log yield as the dependent variable and inputs and other covariates specified as linear in the parameters and variables. Thus, the production function is *convex* in fertilizer and other inputs, violating the standard concavity assumption of production theory. This specification is necessary for fertilizer because it is a non-essential input with many zero values; the data also exhibit zero values for other variables such as hired labor and land preparation cost.

Another restriction of models estimated in log form is the assumption that the error terms u_{it}^{H} and u_{it}^{N} (see equations S8 and S9) satisfy restrictions implied by a multiplicative error structure. Just and Pope (1978) and Antle (1983a) showed that models with multiplicative errors impose restrictions between input use and the higher moments of the yield distribution, and thus restrict the production risk attributes of inputs, including how inputs affect downside and upside risk (Antle 2010). As I discuss in the Supplemental Material, these functional restrictions are likely to affect the estimation of the unobserved heterogeneity component θ_i in the CRC model, and in combination with the identification issues discussed above, are another reason why the estimates may be sensitive to specification.

Another issue with the CRC model specified with the factor structure (9) is its non-linear structural form and number of parameters. With more than two time periods, or more than two endogenous inputs, the model has a large number of parameters and is difficult to formulate, estimate and interpret. As noted above, even with the two-period, two-endogenous variable model presented in the previous section, the estimation procedure often fails to converge and is highly sensitive to small changes in the data. Additionally, with multiple time periods, the identification issues discussed in section 3.2 are more difficult to assess.

5. AN ADDITIVE-ERROR SWITCHING REGRESSION MODEL

In this section I present an alternative production function model, the additive-error switching (AES) regression model. As elaborated in the Supplementary Material, this alternative approach has several advantages over the CRC model: it is simpler and imposes fewer restrictions on the technology; it can be identified with appropriate assumptions similar to the ones required to identify the CRC model; and it can be specified and estimated for any number of time periods. I compare estimates of the hybrid returns distributions from the AES model to

those presented above for the CRC model, and also use it to investigate the possible risk effects of the hybrid seed and fertilizer technology.

The general AES model takes the form $E[Y_{it}^s|Z_{it}^s] = f(Z_{it}^s, \alpha_t)$ with parameter vector α_t, exogenous covariate vector Z_{it}^s, and f any suitable functional form. The yield for system *s* can be expressed as $Y_{it}^s = E[Y_{it}^s|Z_{it}^s] + u_{it}^s$, $E[u_{it}^s|Z_{it}^s] = 0$, giving the generalized yield function:

$$(11) \quad Y_{it} = h_{it}E[Y_{it}^H|Z_{it}^H] + (1 - h_{it})E[Y_{it}^N|Z_{it}^N] + u_{it}, u_{it} = h_{it}u_{it}^H + (1 - h_{it})u_{it}^N.$$

As discussed in section 2, the Kenyan data show that permanent use of hybrid seed and fertilizer is substantially explained by observables. I use this fact to justify the assumption that $E[u_{it}|Z_{it}^H] = 0$ for permanent hybrid users in the medium and high zones, and to justify $E[u_{it}|Z_{it}^N] = 0$ for permanent non-hybrid farms in the low zone. Under the assumption (also used by Suri) that transitory use is determined by random factors not observed at the time farmers make input decisions, it follows that $E[u_{it}|Z_{it}^H] = 0$ for transitory hybrid users in the low zone, and $E[u_{it}|Z_{it}^N] = 0$ for transitory non-hybrid farms in the medium and high zones. Given an appropriate functional form, the mean functions in equation (11) can then be estimated consistently by non-linear least squares regression with a heteroscedastic error. Unlike the CRC model, this additive-error switching regression model (AES) can be estimated for any number of panel data observations using standard non-linear regression methods. The expected gross return to hybrid can be estimated for farms in each zone as $E[Y_{it}^H|Z_{it}^H] - E[Y_{it}^N|Z_{it}^N]$.

In the results presented here I assumed parameter differences over time are captured by time dummies and other time-varying covariates, and I assume $f(Z_{it}^s, \alpha) = \exp[g(Z_{it}^s, \alpha)]$ where g is linear in logs of continuous positive covariates and dummy variables. To account for the

occurrence of zeros in fertilizer use, I utilize the method proposed by Battese (1997) to allow zero values in a log-transformed model.[9] Parameter estimates are presented in the Supplementary Material. As noted in section 2.3, covariate balance is improved by propensity score matching. Estimation of models with matched and stratified data showed similar results, indicating that stratification by zone is sufficient to address covariate balance.

Figure 4 presents the distributions of hybrid returns in yield units estimated with the AES model for the full five-year panel (1997, 2000, 2004, 2007 and 2010; similar results were obtained for the 1997 and 2004 years). This figure shows a pattern of hybrid returns similar to the results presented in Table 3 and in Figure 3 for the CRC models estimated by zone. The average gross returns to permanent hybrid use in the high and medium productivity zones are 515 and 555 kg/ac. The hybrid returns distribution for permanent non-adopters in the low zone and for transitory users in all zones have means near zero. Recalling that the additional cost of hybrid seed and increased fertilizer average over 100 kg/ac including transport cost, we can see that the AES model implies that most permanent hybrid adopters are found to have a positive net return to hybrid, and that most permanent non-adopters and switchers are found to have low or negative net returns.

The relatively wide spread of the returns distributions shown in Figures 3 and 4, however, also implies negative returns for some adopters and positive returns for some non-adopters. One explanation for this result is the sampling error in the estimates. Another explanation could be behavioral, such as the risk effects of hybrid seed and fertilizer. Evidence suggests that many farm decision makers are downside risk averse while seeking upside gains (Antle 2010; Kim et

[9] Define a fertilizer use dummy as d_f equal to 0 if fertilizer is used and equal to 1 if not used, and define fertilizer quantity as x_f. Battese's method is to include d_f and $\log(d_f + x_f)$ in a log-linear model. The coefficient on d_f is interpreted as a distinct intercept for fertilizer non-users, and the coefficient on $\log(d_f + x_f)$ is interpreted as the production elasticity for fertilizer users.

al. 2014). Following the method presented in Antle (2010), the AES model was used to estimate the impacts of hybrid and fertilizer use on the partial second moments of yield and other covariates (i.e., the negative and positive semi-variances), with the square root of the negative and positive partial second moments interpreted as a measures of downside and upside risk (see the Supplementary Material for further details). The results show a positive relationship between expected returns to hybrid and downside risk in the low zone, indicating that farmers with relatively high hybrid returns also experience high downside risk, thus, high downside risk could inhibit adoption. For the high zone, the results show a positive relationship between expected returns to hybrid and downside risk, and high upside risk for farms with low expected returns. This indicates that farmers with relatively low gross hybrid returns experience low downside risk and high upside risk, thus encouraging adoption by farmers who are downside risk averse but upside risk seeking, even when expected returns are low or negative.

Table IV (columns 6 and 7) present the regressions of the AES model's estimated hybrid returns in yield units on distance to market and other covariates. The table confirms the same pattern of correlations as the zone-based CRC models. Inclusion of the zone dummies substantially affects the parameters of the distance variables, again casting doubt on the causal relationship between distance and productivity.

In the Supplemental Material I develop a specification test for restrictions on the error distribution implied by the log-linear fixed-effect specification used for the CRC model (equation S20). These restrictions are due to the multiplicative error specification implied by the log transformation of the dependent variable (Just and Pope, 1978; Antle 1983a). For both hybrid and non-hybrid observations, this specification is rejected with p-values less than 0.001. Thus,

we can conclude that the data strongly reject the restrictions implied by the multiplicative error fixed-effects model.

6. CONCLUSIONS

In the Introduction I noted that agro-ecological factors and market access are considered to be important factors influencing technology adoption in developing countries. The results presented in this study show that, in the case of hybrid maize adoption in Kenya, the agro-ecological factors appear to be more important that market access, in as much as market access can be represented by distance to markets and related infrastructure. The results from both the CRC model estimated by agro-ecological zone and the AES model suggest that there does not appear to be a hybrid maize or fertilizer adoption "puzzle" in Kenya. Most farmers in areas favorable to maize production are using hybrid maize and fertilizer, and fertilizer use was increasing over the 1997-2010 period as improved hybrid varieties became available that were better adapted to the differing agro-ecological conditions across Kenya. The explanation for the low adoption rates of hybrid seed in the unfavorable areas for maize (the low zone in this paper) is primarily that the expected return is low, combined with the downside-risk increasing effects of the technology in the low productivity areas that appears to be associated with lower and more variable rainfall. This finding is consistent with the fact that the most widely used hybrid varieties in the high-productivity zones are not well-suited to the low productivity areas. As studies of crop breeding investments show (Evenson and Gollin 2003), new varieties can be developed that are more suited to the low productivity areas, with the result that hybrid and fertilizer use are both increasing but still lag behind the higher productivity areas (Figure 1). The policy implication of these results is that while infrastructure investments surely can improve the

well-being of farmers in the low-productivity areas, they cannot substitute for much needed variety improvements.

One of the lessons of the past several decades of micro-econometric research is the importance of research design to valid statistical inference (Angrist and Pischke 2010; Heckman 2010). The analysis presented here shows the relevance of this lesson in *ex post* technology impact assessment in cases such as Kenya, where agro-ecological conditions vary greatly and have a correspondingly large effect on farmers' technology choices. This example also should serve as a cautionary tale about the use of complex structural models when identification from the data is weak and "apparent" structural identification may lead to biased estimates that are extremely sensitive to data and model specification.

Finally, in this paper I have demonstrated that relatively simple models based on observables, such as the AES model, together with careful attention to identification in the data, can produce results that are similar to more complex structural models such as the CRC model. The AES approach is attractive because it can accommodate virtually any functional form, can be used with multiple years of panel data, and does not impose restrictions on the risk properties of technologies.

References

Angrist and Pishke (2010): "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24(2):3-30.

Antle, J.M. (1983a). "Testing the Stochastic Structure of Production: A Flexible Moment-Based Approach." *Journal of Business and Economic Statistics* 1(3):192-201.

Antle, J.M. (2010): "Asymmetry, Partial Moments and Production Risk." *American Journal of Agricultural Economics* 92:1294-1309.

Battese, G.E. (1997): "A Note on the Estimation of Cobb-Douglas Production Functions When Some Explanatory Variables Have Zero Values." *Journal of Agricultural Economics* 48:250-252.

Chamberlain, G. (1984): "Panel Data," in *Handbook of Econometrics*, ed. by Z. Griliches andM. Intriligator. Amsterdam: North-Holland.

Evenson, R.E., and D. Gollin, eds. (2003): *Crop Variety Improvement and Its Effects on Productivity: The Impact of International Agricultural Research.* CAB International, Wallingford, UK.

Feder, G., R.E. Just and D. Zilberman (1985): "Adoption of Agricultural Innovations in Developing Countries: A Survey." *Economic Development and Cultural Change* 30:59-76.

Foster, A.D. and M.R. Rosenzweig (2010): "Microeconomics of Technology Adoption." *Annual Review of Economics* 2:395-424. doi.org/10.1146/annurev.economics.102308.124433

Kim, K., J-P Chavas, B. Barham and J. Foltz (2014): "Risk, Irrigation and Downside Risk: A Quantile Analysis of Risk Exposure and Mitigation on Korean Farms." *European Review of Agricultural Economics* 41(5):775–815. doi:10.1093/erae/jbt041

Mathenge, M.K., M. Smale and J. Olwande (2014): "The Impacts of Hybrid Maize Seed on the Welfare of Farming Households in Kenya." *Food Policy* 44: 262-271.

Sunding, D. and D. Zilberman (2001): "The Agricultural Innovation Process: Research and Technology Adoption in a Changing Agricultural Sector." B.L. Gardner and G.C. Rausser, eds. *Handbook of Agricultural Economics, Volume 1A, Agricultural Production.* North-Holland.

Sheahan, M.B. (2011). *Analysis of Fertilizer Profitability and Use in Kenya*. Ms. Thesis, Michigan State University. ttp://fsg.afre.msu.edu/Megan_Sheahan_MS_Thesis_Final.pdf

Sheahan, M., Ariga, J., & Jayne, T. S. (2016): "Modeling the Effects of Input Market Reforms on Fertiliser Demand and Maize Production: A Case Study from Kenya." *Journal of Agricultural Economics*. DOI: 10.1111/1477-9552.12150

Suri, T.(2011): "Selection and Comparative Advantage in Technology Adoption." *Econometrica* 79: 159–209.

Wooldridge, J.M. (2010): *Econometric Analysis of Cross Section and Panel Data.* Cambridge, Mass., MIT Press.

World Bank (2007): *World Development Report 2008: Agriculture for Development.* Washington, D.C., World Bank.