Semiparametric Cost Allocation Estimation

Daniel Wikström, Ludo Peeters and Yves Surry
Swedish University of Agricultural Sciences, Department of Economics, Box 7013, 750 07 Uppsala, Sweden.
Daniel.Wikstrom@slu.se
Hasselt University, Hasselt, Belgium.
Ludo.Peeters@uhasselt.be
Swedish University of Agricultural Sciences.
Yves.Surry@ekon.slu.se

Paper prepared for presentation at the EAAE 2011 Congress
Change and Uncertainty
Challenges for Agriculture,
Food and Natural Resources

August 30 to September 2, 2011
ETH Zurich, Zurich, Switzerland

1. Introduction

The goal of cost-allocation estimation is to assign whole-farm (total) variable costs to each of the farms' enterprises (production activities). Despite the fact that parametric estimation of cost-allocation coefficients is fairly common in applied economics research, we are not aware of any application of semi-parametric modeling. The previous estimation methods can broadly be sorted into three branches: (i) classical (parametric) econometric modeling based on Least Squares (LS) (e.g. Dixon, Batte & Sonka 1984, Hornbaker, Dixon & Sonka 1989, Just, Zilberman, Hochman & Barshira 1990, Hallam, Bailey, Jones & Errington 1999); (ii) Generalized Maximum Entropy (GME)/Generalized Cross-Entropy (GCE) (e.g. Lence & Miller 1998, Léon, Peeters, Quinqu & Surry 1999, Zhang & Fan 2001, Peeters & Surry 2003); and Positive Mathematical Programming (PMP) (e.g. Howitt 1995). A problem, though, with the two latter branches of estimation methods is that they are not solidly grounded on statistical asymptotic theory, which, therefore, precludes proper statistical inference and we will not consider them any further in this paper.

In contrast to previous studies, this paper proposes semi-parametric estimation of a non-stationary Random-Coefficients Model (NS-RCM).[1] Specifically, our proposed NS-RCM is based on the semi-parametric varying-coefficients model developed by Li, Li, Huang, Li & Fu (2002) and further extended by Li & Racine (2010). That is, we estimate a linear cost-allocation model that allows the coefficients to be an unknown (unspecified) function of some selected covariates, here region and economic size, and an additive random error. That is, the cost-allocation coefficients are estimated non-parametrically, albeit in a semi-parametric setting. This modeling approach is more general than the parametric NS-RCM in that it dispenses with the need to impose a functional form on the cost-allocation coefficients.

The purpose of the paper is to introduce a semi-parametric methodology to the field of classical econometric modeling, branch (i), to enable more flexible specifications of the cost-allocation coefficients.

The proposed methodology does not only include the estimation procedure of a semi-parametric random-coefficient model, but also constrained estimation and model testing.[2]

The semi-parametric methodology is applied to farm-level data extracted from the French Farm Accountancy Data Network (FADN), for the year 2006. The data set contains information on a total of 920 specialized multi-crop farms located in nine French FADN regions.

The remainder of the paper is organized as follows. Section 2 discusses the conceptual modeling framework and the basic estimation approach. Section 3 focuses on the basic properties of the constrained kernel-estimation method. Section 4 describes the model-selection process. Section 5 gives an overview of the data and Section 6 presents the empirical results. Section 7 concludes the paper.

[1] We have adopted the terminology 'non-stationary' versus 'stationary' RCM from Hsiao (2002, p. 142). The latter is a model where the coefficients have constant means and variance-covariances while the counterparts of the coefficients of the former model are not constant either due to a stochastic trend or due to some exogenous variables. In our case the latter is true. Thus our modeling should not be confused with non-stationary time-series modeling.

[2] The model testing framework is designed to test parametric models, classical econometric modeling; branch (i), against semi-parametric specifications.

2. Conceptual Model and Estimation Approach

In choosing the conceptual modeling framework, we follow Dixon et al. (1984), Hornbaker et al. (1989) and Just et al. (1990) among others, who suggested a linear model with varying (random) cost-allocation coefficients.[3]

2.1. Basic Setup of the Model.

The basic cost-allocation model can be stated as follows:

$$X_i = Y_i'\beta_i = Y_i'\left[\bar{\beta}(Z_i) + \nu_i\right] \tag{2.1}$$

where X_i is total specific (variable) cost for farm i and Y_i is a $K \times 1$ vector of output values for farm i (see definitions in Table 1). The mean coefficient vector $\bar{\beta}(\cdot)$[4] is a $K \times 1$ vector of unknown functions of Z_i which in turn is a $Q \times 1$ vector of exogenous variables (here economic size and region; i.e. $Q = 2$), and ν_i is a $K \times 1$ vector of random errors. Let us denote the compound error as $\zeta_i \equiv Y_i'\nu_i$ for which we make the following assumptions:

$$E(\zeta_i|Y_i, Z_i) = 0, \tag{2.2}$$

$$V(\zeta_i|Y_i, Z_i) = E(\zeta_i^2|Y_i, Z_i) = \sum_{k=1}^{K} \gamma_k(Z_i)Y_{ki}^2, \tag{2.3}$$

$$E(\zeta_i\zeta_j|Y_i, Z_i) = 0, \quad \text{if} \quad i \neq j. \tag{2.4}$$

Note that assumption (2.2) implies $E(\zeta_i) = 0$ through the "law of iterated expectations"; that is, it is assumed that Y_i and Z_i are independent of the compound error term.[5] The heteroskedasticity assumption in (2.3) is broadly similar to Hornbaker et al., yet we have added unspecified heteroskedasticity conditioning on Z_i. That is, economic size and region are assumed to cause heteroskedasticity. By incorporating the assumption in (2.4), which we also adopted from Hornbaker et al., ensures that any cross-sectional correlation in the error term is ruled out.

In this respect, it should be noted that the crop farms have been sampled by stratification on the basis of economic size and region. Given this stratification procedure, it is natural to include a regional indicator and economic size as elements of Z_i. Accordingly, the cost-allocation mean coefficients are allowed to vary by region and economic size. This, however, remains a rather strong assumption. It means, for instance, that land allocations are homogeneous across crop farms from the same region and economic size. On the other hand, the assumption is less restrictive than (i) the assumption underlying the stationary RCM, which implies that the variation in input costs, X_i, is independent of the exogenous variables; and (ii) the assumption underlying the parametric non-stationary RCM, in that no explicit choice is needed concerning the functional form of the mean coefficient vector.

To obtain a feasible estimator of the cost-allocation mean coefficients, pre-multiply equation (2.1) by Y_i. Moreover, by taking the expectation conditional on $Z_i = z$, we get

$$E(Y_iX_i|z) = \left[E(Y_iY_i'|z)\right]\bar{\beta}(z).$$

Then, multiplying both sides by $[E(Y_iY_i'|z)]^{-1}$ yields

$$\bar{\beta}(z) = \left[E(Y_iY_i'|z)\right]^{-1} E(Y_iX_i|z). \tag{2.5}$$

[3] This model choice, which is standard in applied work, is motivated by the fact that it considerably simplifies the estimation procedure. However, a disadvantage of this choice is that it strongly restricts the underlying technology (for furter details see Chambers & Just 1989).

[4] In the stationary RCM the coefficient vector is $\beta_i = \bar{\beta} + \nu_i$ and , $\bar{\beta}$, is denoted the mean coefficient vector (of constants) because $E(\beta_i|Y_i) = \bar{\beta}$. In the case of the non-stationary RCM there is instead a mean coefficient vector of functions, $E(\beta_i|Y_i, Z_i) = \bar{\beta}(Z_i)$.

[5] A less restrictive assumption is that $E(Y_i\zeta_i|Z_i) = 0$. However, such an assumption would render the variance in Eq. (2.3) much more complicated.

2.2. **Kernel Estimation.** The mean coefficient vector in (2.5) is estimable by the local constant estimator

$$\hat{\hat{\beta}}(z) = \left[(I)^{-1}\sum_{i=1}^{I} Y_i Y_i' K(Z_i, z, H)\right]^{-1} \times \left[(I)^{-1}\sum_{i=1}^{I} Y_i X_i K(Z_i, z, H)\right] \tag{2.6}$$

where

$$K(Z_i, z, H) = \ell \times \kappa \tag{2.7}$$

is a product kernel. In the current setting, the product kernel consists of one univariate kernel to deal with unordered categorical data, ℓ, and a univariate kernel to deal with continuous data, κ. The bandwidth vector, H, contains one bandwidth, λ, for the unordered kernel and second one, h, for the continuous kernel.

The unordered kernel, for the regional indicator, is defined as

$$\ell(Z_{ig}, z_g, \lambda) = \begin{cases} 1, & Z_{ig} = z_g \\ \lambda \in [0,1], & \text{otherwise} \end{cases} \tag{2.8}$$

where the index g is used to denote the region component of z. Note further that if $\lambda = 0$, the kernel reduces to a simple indicator function, which acts as a collection of regional dummy variables. Conversely, if $\lambda = 1$, the kernel turns into a uniform weighting function, where cost-allocation coefficients become identical across all regions. In finite samples, there will always be a $\lambda > 0$ for which kernel smoothing is MSE efficient compared to estimators based on sample splitting/dummy variables (Brown & Rundell 1985, Wikström 2011).

For the economic size variable, we use a second-order normal kernel, designed for continuous data, which is defined as

$$\kappa(Z_{ie}, z_e, h) = \phi\left(\frac{Z_{ie} - z_e}{h}\right) \tag{2.9}$$

where $\phi(\cdot)$ is the standard normal density function, and e indexes the economic size component of z.

A crucial step in non-parametric estimation is to select an appropriate bandwidth vector. We follow the Least-Squares Cross-Validation (LSCV) procedure proposed by (Li & Racine 2010), for which they have outlined the theoretical underpinnings necessary to ensure consistency and asymptotic normality of the estimator in (2.6).

2.3. **Treatment of Heteroskedasticity.** The model defined in (2.1) through (2.4) incorporates heteroscedastic due to both Y_i and Z_i. The estimation procedure described above is robust to the presence of heteroskedasticity. Yet, in order to gain some efficiency, we decided to employ a heuristic two-step procedure akin to the FGLS estimator used for parametric linear models.

In Step 1 we estimate $\hat{\hat{\beta}}(z)$ by (2.6) and compute the residuals $\hat{\zeta}_i = X_i - Y_i'\hat{\hat{\beta}}(z)$ to obtain the variance estimates $\hat{\sigma}_i^2$ by the linear regression $\hat{E}(\hat{\zeta}_i^2|Y_i, Z_i) = \sum_{k=1}^{K} \hat{\gamma}_k(Z_i) Y_{ki}^2$. In Step 2 we re-estimate (2.6) based on the transformed variables $\tilde{Y}_i \equiv \frac{Y_i}{\sqrt{\hat{\sigma}_i^2}}$ and $\tilde{X}_i \equiv \frac{X_i}{\sqrt{\hat{\sigma}_i^2}}$ are used. Finally we iterate this procedure until the changes in $\hat{\hat{\beta}}(z)$ are negligible.[6]

[6]It should be emphasized that the theoretical properties of the proposed GLS-type transformation are not yet fully established for the semi-parametric estimation. Nevertheless, we conducted some Monte Carlo simulations which provide support for this approach. It was found that the GLS-type transformations improved the MSE of $\hat{\hat{\beta}}(Z_i)$.

3. CONSTRAINT ESTIMATION

Economic theory puts obvious (accounting) restrictions on the cost-allocation coefficients. First, the cost-allocation mean coefficients should all be positive:

$$\bar{\beta}(Z_i) \geq \mathbf{0}. \tag{3.1}$$

Second, there are also upper limit constraints on the cost-allocation mean coefficients. In the present case, total specific cost (total variable cost) is used as the dependent variable. Total revenue $Y_i'\mathbf{1}$ has to cover total specific cost $X_i = Y_i'\left[\bar{\beta}(Z_i) + v_i\right]$. Since our focus is on estimating the cost-allocation mean coefficients, conditional on region and economic size, we should get

$$\begin{gathered} E(Y_i|Z_i)'\mathbf{1} - E(Y_i|Z_i)'\bar{\beta}(Z_i) \geq 0 \\ \Leftrightarrow \\ E(Y_i|Z_i)'\bar{\beta}(Z_i) \leq E(Y_i|Z_i)'\mathbf{1} \\ \Leftrightarrow \\ \bar{\beta}(Z_i) \leq \mathbf{1} \end{gathered} \tag{3.2}$$

where $\mathbf{1}$ is a $K \times 1$ vector of ones.

In this study we impose the restrictions in (3.1) and (3.2) by means of recently developed constrained kernel estimation (Racine & Parmeter 2010). Consider the following generalization of the estimator (2.6) in Section 2:

$$\hat{\bar{\beta}}(z|p) = \sum_{i=1}^{I} A_i X_i p_i \tag{3.3}$$

where

$$A_i = \left[\sum_{j=1}^{I} Y_j Y_j' K\left(Z_j, z, H\right)\right]^{-1} I Y_i K\left(Z_i, z, H\right) \tag{3.4}$$

Note that if $p_i = p = 1/I$ for all $i = 1, \ldots, I$, (3.3) collapses to the unconstrained estimator (2.6) in Section 2. Let p_u be the I-vector of uniform weights and p the vector of weights to be selected. To impose constraints, Racine & Parmeter (2010) in line with Hall & Huang (2001) propose to minimize the distance from p to the uniform weights p_u. Racine & Parmeter (2010) make use of the L_2-metric and the minimization problem becomes

$$\min_{p} \quad D(p) = (p_u - p)'(p_u - p) \tag{3.5}$$

$$\text{subject to} \quad \mathbf{0} \leq \hat{\bar{\beta}}(z|p) \leq \mathbf{1} \quad \text{and} \quad \sum_i p_i = 1. \tag{3.6}$$

The idea behind (3.5) is to select the estimator which is "closest" the unconstrained estimator given the constraints in (3.6). An important advantage of the L_2-metric is that the minimization problem represented by (3.5) and (3.6) can be solved as an ordinary quadratic-programming problem which most statistical software packages have as a standard procedure (see Racine & Parmeter 2010, for detailed examples).

However, establishing the theoretical properties in terms of possible efficiency gains is beyond the scope of the present paper.

4. Model Selection

Despite the fact that the model with fixed (constant) mean coefficients is overly restrictive, researchers continue to use it in applied work. If the fixed mean coefficient assumption is correct, obvious efficiency gains can be achieved compared to the more general non-stationary RCM. Therefore, it is of primary importance to test less flexible parametric specifications against the more general semi-parametric model proposed in Section 2.

4.1. **Testing Parametric Specifications.** Let us state the null hypothesis as

$$H_0: \quad Pr\left[\bar{\beta}(Z) = \bar{\beta}_0(Z;\alpha_0)\right] = 1 \tag{4.1}$$

where $\bar{\beta}_0(Z;\alpha_0)$ is a vector of parameterized functions under the null (and thus the semi-parametric mean coefficient vector is the alternative hypothesis). For example if we assume a parametric stationary RCM (the fixed mean coefficient model), $\bar{\beta}_0(Z;\alpha_0)$ is reduced to α_0, which is a $K \times 1$ vector of constants. If $\bar{\beta}_0(Z;\alpha_0) = \alpha_0$ is, correctly, rejected it means that the parametric model is misspecified and will produce inconsistent estimates. On the other hand, if $\bar{\beta}_0(Z;\alpha_0) = \alpha_0$ is (correctly) not rejected the semi-parametric mean coefficient vector is still consistent but inefficient compared to fixed mean coefficient model.

Li et al. (2002) propose a test based on the integrated squared difference

$$\int \left[\bar{\beta}(z) - \bar{\beta}_0(z;\alpha_0)\right]' \left[\bar{\beta}(z) - \bar{\beta}_0(z;\alpha_0)\right] dz$$

and Li & Racine (2010) extend this test to the case of mixed data; that is, where the model includes both categorical and continuous data. Based on the integrated squared difference Li & Racine derive a test statistic, which is given by

$$\hat{T}_n \equiv n\sqrt{\hat{h}}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} Y_i'Y_j\hat{u}_i\hat{u}_j K\left(Z_i,Z_j,H\right)/\hat{\sigma}_0 \tag{4.2}$$

where $\hat{\sigma}_0 = 2n^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(Y_i'Y_i)^2\hat{u}_i^2\hat{u}_j^2 K\left(Z_i,Z_j,H\right)^2$ and $\hat{u}_i$ is the i^{th} estimated residual under H_0, $X_i - Y_i'\bar{\beta}_0(Z;\hat{\alpha}_0)$, and the other terms are explained before in Section 2.

Li & Racine (2010) show that asymptotic null distribution of $\hat{T}_n$ is the standard normal distribution. However, given that kernel-based specification tests are plagued with size distortion in finite samples, they propose to generate an empirical null distribution using the wild bootstrap. We follow their suggestion, although we are using the Rademacher distribution instead of the two-point distribution they proposed. Davidson & Flachaire (2008) compare these two distributions for the wild bootstrap and recommend the use of the former.[7]

4.2. **Three Different Model Specifications.** In the empirical analysis which follows below, our aim is to perform the T-test to check the adequacy of two distinct parametric RCMs, henceforth called Model M2 and Model M3, respectively, against the alternative model, represented by the semi-parametric non-stationary RCM, henceforth called Model M1 (defined in Section 2):

Model M1	Semi-parametric non-stationary RCM
Model M2	Parametric non-stationary RCM
Model M3	Parametric stationary RCM (or fixed mean coefficient model)

The two parametric RCMs, Model M2 and Model M3, are different from the semi-parametric RCM, Model M1, as well as from one another through their mean coefficient vector $\bar{\beta}(Z;\alpha)$.

[7]Since the GLS transformation of the data changes the X and Y variables differently for the model under the null and the alternative model, we perform the test without any transformations. However, since the wild bootstrap is robust to heteroskedasticity, the test is also appropriate in our setting.

Specifically, Model M2 is a parametric non-stationary RCM, in which the mean coefficient vector is defined as

$$\bar{\beta}(Z;\alpha_{M2}) = \begin{pmatrix} \alpha_{11}d_1 + \alpha_{21}d_2 + \cdots + \alpha_{G+1,1}Z_e \\ \alpha_{12}d_1 + \alpha_{22}d_2 + \cdots + \alpha_{G+1,2}Z_e \\ \vdots \\ \alpha_{1K}d_1 + \alpha_{2K}d_2 + \cdots + \alpha_{G+1,K}Z_e \end{pmatrix} \tag{4.3}$$

where $d_1,\ldots,d_G$ are regional dummies, and Z_e is the economic size covariate. On the other hand, Model M3 is a parametric stationary RCM, in which the mean coefficient vector is defined as

$$\bar{\beta}(Z;\alpha_{M3}) = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_K \end{pmatrix}. \tag{4.4}$$

In addition, we modify the heteroskedasticity assumption (2.3) for the two parametric models. Specifically, for Model M2, we make the following assumption:

$$E(\zeta_i^2|Y_i,Z_i) = \sum_{k=1}^{K}\left[\gamma_{1k}d_1 + \gamma_{2k}d_2 + \cdots + \gamma_{G+1,k}Z_e\right]Y_{ki}^2, \tag{4.5}$$

whereas for Model M3, we assume

$$E(\zeta_i^2|Y_i) = \sum_{k=1}^{K}\gamma_k Y_{ki}^2. \tag{4.6}$$

All models are estimated using a GLS transformation, where $E(\zeta_i^2|Y_i,Z_i)$ is estimated according to the assumptions specific to each model.

5. Data

5.1. **Data and Variables.** The empirical analysis presented in this section uses data extracted from the French Farm Accountancy Data Network (FADN), for 2006 (for a detailed description of the data, see European Commission (2009)).

The data is sampled according to a three-way stratification. Specifically, farms in the population are stratified (probability weighted) on the basis of three different criteria: (i) *type of farm* (TF); (ii) *economic size* (ES), measured in European Size Units (ESU); and (iii) *geographical unit* (GU). The analysis is limited to the so-called "COP farms"; that is, farms in the TF14(13) FADN grouping, which comprises farms specialized in cereals, oilseeds, and protein crops. The regions (Départements) that have been selected are those with a total of at least 50 COP farms. To make the estimation consistent with the stratified sample as well as to account for heterogeneity, we condition on region and economic size. Table 1 provides definition of the variables, including: total variable costs (X); total output values for wheat (Y_1), barley (Y_2), oilseeds (Y_3), other crops (Y_4), and other agricultural products (Y_5); economic size (Z_e), and the regional indicator (Z_g). Summary statistics are provided in Table 2.

6. Empirical Results

In this section, we first conduct model selection with help of the T-test presented in Section 4. When an appropriate model is selected, according to the T-test, estimated cost-allocation mean coefficienxts are presented. Finally, to further assess the validity of the empirical results, it is instructive to use real, farm-level accounting data on cost allocations –which are typically rarely available– for the purpose of comparison. Specifically, we use accounting data made available by the Centre d'Économie agricole de la Meuse, for Wheat, Barley, and Oilseeds.

Variable	Description	FADN Definition
X	Total specific (variable) cost	SE281
	Total output values for	
Y_1	Wheat	K120TP+K121TP
Y_2	Barley	K123TP
Y_3	Oilseeds	K132TP
Y_4	Other crop output	SE135-(K120TP+K121TP)-K123TP-K132TP
Y_5	Other agricultural outputs	SE131-SE135
Z_e	Economic size units (ESU)	SE005
Z_g	FADN region	A1

TABLE 1. Variable Definitions

Variable	Mean	Std. dev.	Min.	1st Qu.	Median	3rd Qu.	Max.
Total specific cost (1000 €)							
X	51.3	34.1	0.4	28.9	43.4	65.8	334.1
Total output value (1000 €)							
Y_1	47.7	37.2	0.0	23.0	39.2	64.4	349.9
Y_2	14.1	16.7	0.0	0.0	9.9	21.6	129.6
Y_3	20.4	18.6	0.0	7.1	16.0	28.9	151.7
Y_4	28.8	42.4	0.0	2.5	14.9	38.2	465.6
Y_5	20.7	36.1	0.0	2.7	8.9	24.8	391.4
Economic size units							
Z_e	107.0	62.8	10.0	64.7	94.7	136.1	510.6

TABLE 2. Summary statistics

6.1. **Specification-Test Results.** Table 3 presents the results of the hypothesis tests (recall that M1 is the alternative model). Model M3, the fixed mean coefficient model, is the most restrictive model, nevertheless, if it is a correct specification it will be efficient compared to both M1 and M2. However model M3 is rejected on the five percent level (P-value $= 0.042$). This suggest that Model M3 is a misspecified model, and that regional heterogeneity and/or economic size matter. The parametric non-stationary RCM, model M2, is also rejected in favor of model M1 (P-value < 0.001). Unlike model M3, model M2 includes both regional and economic size variables but still it is rejected. Normally, this is an indication of a misspecification of the functional form of the mean coefficients. However the bandwidth vector of model M1 includes an abnormally large value for economic size ($h \approx 375$). As shown by Hall, Li & Racine (2007) the cross-validation bandwidth approaches infinity for irrelevant continuous variables. Thus, a large bandwidth for economic size indicates that the variable is irrelevant for the estimation of cost-allocation mean coefficients of the crop farmers in our data set.

If economic size is irrelevant, the rejection of model M2 does not steam from misspecification of the coefficients but from inclusion of a superfluous variable. Inclusion of an irrelevant variable causes estimation error that makes the estimator inefficient. Hall et al. (2007) do not only show that a large bandwidth indicates the irrelevance of the attached variable but also that the kernel estimator has the ability to automatically smooth out the extra variability an irrelevant variable normally causes in parametric estimation.[8] The point we want to make is that the rejection, of model M2 in favor of model M1, rather originates from the inclusion of an irrelevant variable (economic size) than from misspecification of the functional form.

[8]The intuition behind this ability of the kernel estimator is quite simple. If the bandwidth, h, of the continuous kernel in (2.9) is huge the kernel weights will be uniform, irrespectively of the difference in economic size, $Z_{ie} - z_e$. And consequently the estimator in (2.5) will be independent of economic size.

H_0	T-test (p-values)
Parametric nsRCM (**M2**)	0.000
Parametric sRCM (**M3**)	0.042
Semi-parametric nsRCM (**M1'**)	0.876

TABLE 3. Results of specification test

To formally test if economic size is irrelevant, we exclude the economic size variable (Z_e) from model M1, we denote this model M1'. Model M1' is a semi-parametric non-stationary RCM, only including the regional indicator variable Z_g. We test this coefficient specification against model M1.[9] The P-value reported in Table 3 is 0.876 and model M1' is not rejected. Hence we find no significant support for economies of scale.

On basis of the T-test we have selected Model M1' which is a semi-parametric non-stationary RCM model, where the mean coefficients are varying by region but not by economic size. In the next subsection the results, in form of the cost-allocation mean coefficients, of Model M1' is presented.

6.2. **Estimated Cost-Allocation Mean Coefficients.** Table 4 presents the estimated cost-allocation mean coefficients of model M1'.The estimates are all within the [0,1] interval.[10]

The estimates of Model M1, M2 and M3 are presented in Table 7 in the appendix. Although these specifications are rejected, the estimates bear some interesting information. A first remark is that the estimates of model M1 are very similar to those given by M1'. The explanation of this is, as discussed in the previous subsection, that the large bandwidth on economic size effectively smooths out the variation of this irrelevant variable. A related remark is the larger variation shown in the estimates of the parametric non-stationary RCM (M2). This is an example of the extra variability an irrelevant variable causes in parametric estimation versus kernel estimation. The estimated coefficients of the parametric Model M3, on the other hand, show large resemblance to the average mean coefficient estimates of the M1' model. The M3 estimates are not affected by any extra variability caused by economic size and seem to produce quite good overall cost-allocation estimates, but at the same time completely misses out on the heterogeneity across regions.

	Wheat	Barley	Oilseeds	Other Crops	Other Ag. Prods	R^2
M1'–Semi-parametric nsRCM (regions only)						
Mean	0.39	0.43	0.51	0.27	0.42	
Std. dev.	0.03	0.05	0.06	0.01	0.05	
Min.	0.33	0.36	0.42	0.25	0.35	
1st Qu.	0.38	0.39	0.49	0.26	0.39	
Median	0.39	0.42	0.49	0.26	0.43	
3rd Qu.	0.42	0.49	0.55	0.28	0.45	
Max.	0.42	0.53	0.62	0.29	0.50	0.84

TABLE 4. Estimation results

[9]Model M1' is actually asymptotically equivalent to a parametric non-stationary RCM, only including regional dummies, if the bandwidth, λ, equals zero. However if $\lambda > 0$ some bias is induced to reduce variance. It has been shown theoretical that this type of kernel estimation is MSE ('Mean square error') efficient to parametric dummy variable estimation (Brown & Rundell 1985, Wikström 2011). This efficiency advantage has also been manifested in Monte Carlo simulations of several shapes (two recent examples are Li, Racine & Wooldridge 2009, Wikström 2011).

[10]The values in Table 4 are the unconstrained estimates; since these estimates relax the constrain on the estimator in (3.6) when $p = p_u$ the constrained and unconstrained estimators (and estimates) are identical.

The regional estimates for M1' are presented in Table 5, along with wild bootstrapped standard errors. Relatively large heterogeneity exists across regions (Other Crops is the only exception). For example, it can be seen that the Île de France region has the smallest cost-allocation coefficient for Wheat among the nine regions, and the largest one for Oilseeds. Conversely, the Lorraine region almost shows the opposite pattern, with the smallest coefficient for Oilseeds and second largest for Wheat.

Region	Wheat	Barley	Oilseeds	Other Crops	Other Ag. Prods
Île de France	0.325	0.492	0.620	0.288	0.350
	(0.029)[a]	(0.072)	(0.058)	(0.019)	(0.019)
Champagne-Ardenne	0.416	0.433	0.517	0.262	0.349
	(0.031)	(0.048)	(0.042)	(0.036)	(0.036)
Picardie	0.380	0.525	0.486	0.264	0.426
	(0.024)	(0.051)	(0.048)	(0.018)	(0.024)
Centre	0.389	0.387	0.495	0.247	0.501
	(0.030)	(0.052)	(0.06)	(0.027)	(0.020)
Bourgogne	0.388	0.417	0.560	0.263	0.410
	(0.025)	(0.067)	(0.047)	(0.017)	(0.017)
Lorraine	0.421	0.355	0.419	0.281	0.385
	(0.024)	(0.045)	(0.053)	(0.016)	(0.023)
Poitou-Charentes	0.381	0.395	0.553	0.260	0.437
	(0.026)	(0.058)	(0.078)	(0.023)	(0.028)
Aquitaine	0.368	0.400	0.535	0.293	0.446
	(0.023)	(0.039)	(0.042)	(0.018)	(0.022)
Midi-Pyrénées	0.421	0.437	0.429	0.271	0.436
	(0.033)	(0.043)	(0.059)	(0.014)	(0.020)

[a] *Standard errors in parenthesis*

TABLE 5. Estimated regional cost-allocation coefficients (M1')

6.3. **Validity of Empirical Results.** We have applied statistical tests to find the most appropriate model for our estimations. However, it remains difficult to say how good these estimates really are. To assess the validity of the empirical results (from M1'), it is instructive to use real, farm-level accounting data on cost allocations –which are typically rarely available– for the purpose of comparison. Specifically, we use accounting data made available by the Centre d'Économie agricole de la Meuse, for Wheat, Barley, and Oilseeds. The data set contains accounting information for 565 multi-crop farms in the Département de Meuse, which is located in the Lorraine region, along the border to Champagne-Ardenne.

The averages of the cost-accounting allocations for this set of farms are presented in Table 6. Comparison with the estimated cost-allocation coefficients returned by model M1'. There is a noticeable match between the M1 estimates and the accounting averages for the neighboring Champagne-Ardenne region. However, some moderate differences are found between the M1 estimates and the accounting averages for the Lorraine region itself.

Wheat	Barley	Oilseeds
0.405	0.431	0.517
(0.106)[a]	(0.108)	(0.120)

[a] *Standard deviations in parenthesis*

TABLE 6. Average cost-allocation coefficients in Department of Meuse

7. Conclusion

We have proposed a new way to estimate cost-allocation mean coefficients. The linear functional form with varying-/random-coefficients traditionally used in cost-allocation estimation fits very well within the semi-parametric framework. Semi-parametric modeling is more general than parametric varying-/random-coefficients models. The proposed methodology incorporates both constrained estimation as well as model testing.

In the empirical study of multi-crop farmers in nine French regions, we reject the linear model with fixed mean coefficients as well as the parametric non-stationary RCM in favor of the semi-parametric non-stationary RCM. The regional estimates obtained from the selected semi-parametric specification are also compared to actual cost-allocation coefficients, obtained from multi-crop farms in the Département Meuse, with the help of very detailed farm-accounting data. The estimates from the neighboring region Champagne-Ardenne show strong resemblance to the computed cost-allocation coefficients of the Department of Meuse. Hence, the obtained estimates do not only make sense on statistical ground.

Extension of the proposed methodology could be to incorporate panel data into the semi-parametric framework as well as develop system estimation with appropriate constraints. The latter should be relatively straightforward while the simplicity for the panel data modeling depends on the assumptions put on the farm specific effects. In either way we leave this to future research.

References

Brown, P. & Rundell, P. (1985), 'Kernel estimates for categorical data', *Technometrics* **27**, 293–299.

Chambers, R. G. & Just, R. E. (1989), 'Estimating multioutput technologies', *American Journal of Agricultural Economics* **71**(4), 980–995.

Davidson, R. & Flachaire, E. (2008), 'The Wild Bootstrap, Tamed at Last', *Journal of Econometrics* **146**(1), 162 – 169.

Dixon, B., Batte, B. & Sonka, S. (1984), 'Random Coefficients Estimation of Average Total Product Costs for Multiproduct Firms', *Journal of Business & Economic Statistics* .

European Commission (2009), 'Farm Accountancy Data Network'.
URL: *http://ec.europa.eu/agriculture/rica*

Hall, P. & Huang, L. S. (2001), 'Nonparametric kernel regression subject to monotonicity constraints', *Annals Of Statistics* **29**(3), 624–647.

Hall, P., Li, Q. & Racine, J. S. (2007), 'Nonparametric estimation of regression functions in the presence of irrelevant regressors', *Review Of Economics And Statistics* **89**(4), 784–789.

Hallam, D., Bailey, A., Jones, P. & Errington, A. (1999), 'Estimating input use and production costs from farm survey panel data', *Journal Of Agricultural Economics* **50**(3), 440–449.

Hornbaker, R., Dixon, B. & Sonka, S. (1989), 'Estimating production activity costs for multi-output firms with a random coefficient regression model', *American Journal of Agricultural Economics* **71**, 167–177.

Howitt, R. E. (1995), 'Positive Mathematical-Programming', *American Journal of Agricultural Economics* **77**(2), 329–342.

Hsiao, C. (2002), *Analysis of Panel Data*, Cambridge University Press.

Just, R. E., Zilberman, D., Hochman, E. & Barshira, Z. (1990), 'Input allocation in multicrop systems', *American Journal Of Agricultural Economics* **72**(1), 200–209.

Lence, S. H. & Miller, D. J. (1998), 'Recovering output-specific inputs from aggregate input data: A generalized cross-entropy approach', *American Journal Of Agricultural Economics* **80**(4), 852–867.

Li, Q., Huang, C. J., Li, D. & Fu, T. T. (2002), 'Semiparametric smooth coefficient models', *Journal Of Business & Economic Statistics* **20**(3), 412–422.

Li, Q. & Racine, J. S. (2010), 'Smooth Varying-Coefficient Estimation and Inference for Qualitative and Quantitative Data', *Econometric Theory* **26**, 1607–1637.

Li, Q., Racine, J. S. & Wooldridge, J. M. (2009), 'Efficient estimation of average treatment effects with mixed categorical and continuous data', *Journal Of Business & Economic Statistics* **27**(2), 206–223.

Léon, Y., Peeters, L., Quinqu, M. & Surry, Y. (1999), 'The use of the maximum entropy to estimate input-output coefficients from regional farm accounting data', *Journal of Agricultural Economics* **50 (3)**, 425–439.

Peeters, L. & Surry, Y. (2003), 'Entropy Estimation of a restricted Hildreth-Houck Random-Coefficients Model with an Application to Cost Allocation in Multi-Product Farming in France', *ITEO Research Paper 03/01. Limburgs Universitair Centrum. Diepenbeek, Belgium* .

Racine, J. S. & Parmeter, C. F. (2010), 'Constrained nonparametric kernel regression: Estimation and inference', *Unpublished manuscript* .

Wikström, D. (2011), 'A Finite Sample Improvement of the Fixed Effects Estimator of Technical Efficiency', *Unpublished manuscript* .

Zhang, X. B. & Fan, S. G. (2001), 'Estimating crop-specific production technologies in chinese agriculture: A generalized maximum entropy approach', *American Journal Of Agricultural Economics* **83**(2), 378–388.

8. Appendix

	Wheat	Barley	Oilseeds	Other Crops	Other Ag. Prods	R^2
M1–Semi-parametric nsRCM (alternative model)						
Mean	0.39	0.43	0.51	0.27	0.42	
Std. dev.	0.03	0.05	0.06	0.01	0.05	
Min.	0.32	0.35	0.42	0.24	0.35	
1st Qu.	0.38	0.39	0.49	0.26	0.38	
Median	0.39	0.42	0.50	0.26	0.43	
3rd Qu.	0.42	0.49	0.56	0.28	0.45	
Max.	0.43	0.53	0.63	0.29	0.51	0.84
M2 Parametric nsRCM						
Mean	0.43	0.57	0.55	0.27	0.40	
Std. dev.	0.18	0.38	0.22	0.07	0.14	
Min.	0.22	-0.17	-0.00	0.13	0.09	
1st Qu.	0.37	0.37	0.40	0.21	0.34	
Median	0.38	0.47	0.57	0.26	0.43	
3rd Qu.	0.46	0.65	0.66	0.29	0.48	
Max.	1.05	1.95	1.01	0.41	0.77	0.86
M3 Parametric sRCM						
$\hat{\alpha}_{M3}$	0.38	0.42	0.52	0.27	0.41	0.82

TABLE 7. Estimation results for model M1, M2 and M3