Empirical Economics manuscript No.
(will be inserted by the editor)

Transaction Costs and Social Networks in Productivity Measurement

Geraldine Henningsen · Arne Henningsen · Christian H.C.A. Henning

Forthcoming in **Empirical Economics**. The final publication is available at Springer via http://dx.doi.org/10.1007/s00181-014-0882-y.

Abstract We argue that in the presence of transaction costs, observed productivity measures may in many cases understate the true productivity, as production data seldom distinguish between resources entering the production process and resources of a similar type that are sacrificed for transaction costs. Hence, both the absolute productivity measures and, more importantly, the productivity ranking will be distorted. A major driver of transaction costs is poor access to information and contract enforcement assistance. Social networks often catalyse information exchange as well as generate trust and support. Hence, we use measures of a firm's access to social networks as a proxy for the transaction costs the firm faces. We develop a microeconomic production model that takes into account transaction costs and networks. Using a data set of 384 Polish farms, we empirically estimate this model and compare different parametric, semiparametric, and nonparametric model specifications. Our results generally support our hypothesis. Especially large trading networks and dense household networks have a positive influence on a farm's productivity. Furthermore, our results indicate that transaction costs have a measurable impact on the productivity ranking of the farms.

Keywords Information networks · Transaction Costs · Semiparametric estimation · Nonparametric estimation · Productivity analysis

Geraldine Henningsen
Department of Management Engineering, Technical University of Denmark
Frederiksborgvej 399, 4000 Roskilde, Denmark
Tel.: +45-46775159, E-mail: gehe@dtu.dk

Arne Henningsen
Department of Food and Resource Economics, University of Copenhagen
Rolighedsvej 25, 1958 Frederiksberg C, Denmark
Tel.: +45-35332274, E-mail: arne@ifro.ku.dk

Christian H.C.A Henning
Institute of Agricultural Economics, Christian-Albrechts University Kiel
Olshausenstrasse 40, 24098 Kiel, Germany
Tel.: +49-431-8804444, E-mail: chenning@ae.uni-kiel.de

JEL codes D22, D23, D24, D85, L14, Q12

1 Introduction

Productivity, relating output quantities to input quantities, is a popular and widespread concept to measure the performance of firms, public services, and even nations. To explain the variation in observed productivity measures, it is common to assume that some enterprises use an inferior production technology and/or use production inputs inefficiently, e.g., because of managerial deficiencies or low qualifications.

However, we propose another—less obvious—source of variation in observed productivity measures, which is unrelated to the production process as such, but originates from the transaction costs the producer faces on input and output markets, and in the search for technological innovations (Williamson, 2000; Castilla et al, 2000). Transaction costs that originate from trade activities and knowledge acquisition consume resources often similar in type to those that enter into the production process. Hence, if transaction costs of a certain magnitude are present, we can distinguish between two usages of inputs: (a) inputs fed into the production process and (b) inputs consumed by transaction costs. As most production data combine these two usages into a single variable, productivity analyses based on such data can result in downward biased estimates of the true productivity of the production process. More importantly, if firms are unequally affected by transaction costs, the distortions of the productivity estimates also change the relative productivity differences between the analysed enterprises, i.e., the ranking of the observed productivity.

A crucial factor which determines the level of transaction costs a firm faces is access to information and contract enforcement assistance (Ménard, 2000; Levi, 2000; den Butter and Mosch, 2003). Therefore, we expect distortions to be particularly large for production data from enterprises with missing or limited access to formal institutions and public information channels—like firms in developing countries or in remote and rural areas in transition countries such as Poland. Earlier studies have shown that in the absence of well-functioning formal institutions and public information channels, social networks often supersede these formal facilitators (e.g., Nee, 1998; Fafchamps, 2001; Henning and Zuckerman, 2006). But in contrast to formal institutions, which (ideally) are equally accessible to everybody, the access to the benefits of a network varies with a firm's position in the network[1] (Buskens, 1999; Dekker, 2001). Based on these findings, we argue that in the presence of transaction costs, a firm's network will have a measurable impact on the firm's observed productivity, i.e. a firm's network will be a proxy for otherwise unobservable transaction costs.

A vast literature has quantified the effect of social networks on various firm performance measures (Stam et al, 2013), such as production efficiency (Lau and Bruton, 2011), labour and capital productivity (Di Matteo et al, 2005), or other broader non-monetary performance indicators (Bradley et al, 2012; Prajapati and Biswas, 2011).

[1] A firm's network position refers to the structural connectedness of the firm to other firms and other relevant actors.

Other studies (Luo, 2003; Beckmann et al, 2004; Henning et al, 2012) have empirically analysed the linkage between market uncertainty, as one element of transaction costs, and networks. A third strand of literature has examined the effect of networks as information channels on the process of innovation adoption (Jenssen and Koenig, 2002; Di Matteo et al, 2005; Bandiera and Rasul, 2006). All studies report that social networks have significant effects; in a meta-analysis Stam et al (2013) found that the magnitude of the effect of social networks on firm performance even exceeds the effect of human capital. These findings support our choice of proxy and confirm that networks have a positive effect on firm performance, market uncertainty, and information flow.

However, a coherent theoretical model linking social networks, transaction costs, and productivity together is still missing. Furthermore, to the best of our knowledge, no other study so far has explicitly modelled the effect of transaction costs on observed productivity. By doing so, we want to demonstrate that in cases where high transaction costs prevail, the interpretation of observed productivity measures must be based on a broader perspective than the current standard. In situations where high transaction costs must be expected, the observed productivity is a measure of firm productivity rather than a measure of the efficiency of the production process itself. This change in perspective can have implications for policy recommendations, e.g., policies concerning the promotion of farm and firm productivity in developing or transition countries.

We base our analysis on a cross-sectional dataset from a representative sample of 384 Polish farms. The dataset contains production data as well as data on ego-centred farm networks[2].

As earlier empirical studies find different functional relationships between social network parameters and measures of firm performance (Yu and Chiu, 2013; Stam et al, 2013) we apply and compare different parametric, semiparametric, and nonparametric specifications of the regression function in our analysis. The results show strong consistency over most estimation models and support our hypothesis and earlier findings that social networks have a significantly positive effect on farm productivity. In particular, large trading networks and dense household networks promote farm productivity. In contrast to studies reporting a non-linear relationship between measures of closure and firm performance (Uzzi, 1996; Yu and Chiu, 2013), our results generally support only a linear relationship between density and farm productivity. Furthermore, our results strongly indicate that by ignoring transaction costs, both the absolute observed productivity measures as well as the productivity ranking change considerably.

The remainder of the article is structured as follows: section two introduces the theoretical concept combining social networks, transaction costs, and productivity into a coherent model; section three gives a short description of the data; section four and five present the econometric specifications and the results, respectively; and section six discusses the findings.

[2] Ego-centred networks are networks sampled with open boundaries which are structured around one actor, the *ego*. In our particular case, the ego is the farmer. In contrast, a full network is a network with closed boundaries where the ties between all actors in the closed set are mapped.

2 Microeconomic Foundation

We assume that a firm uses a vector of n input quantities $\mathbf{x}^{PD} = (x_1^{PD}, \ldots, x_n^{PD})'$ to produce the output quantity y, where the transformation of the inputs into the output can be described by the production function:

$$y = f(\mathbf{x}^{PD}, T), \tag{1}$$

where T indicates the productivity level of the firm's production activities.[3]

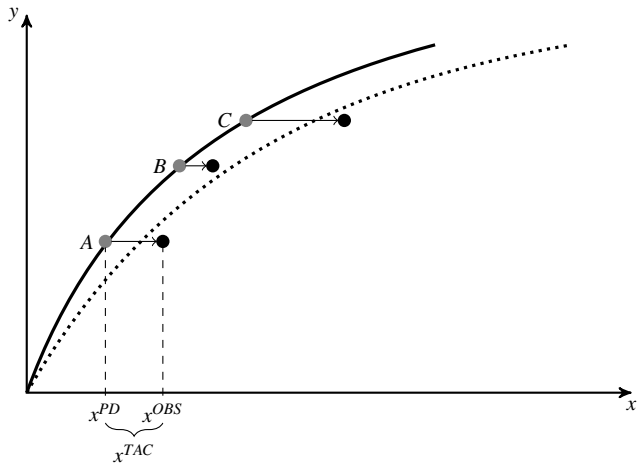


Fig. 1: Influence of transaction costs on the measurement of firm productivity

Figure 1 illustrates, by means of a simplified production process, the role of transaction costs in empirical production analysis. The solid black curve indicates the production function $f(x^{PD}, T)$. The three grey dots indicate the input quantities used for production (x^{PD}) and the corresponding output quantities (y) of three different firms (named A, B, and C). For simplicity, we assume that all three firms have an average productivity level (T) compared to other firms of the same size so that all three grey dots lie on the production function $f(x^{PD}, T)$. As argued above, firms also use inputs for trade activities and knowledge acquisition (x^{TAC}) so that the observed input quantities are $x^{OBS} = x^{PD} + x^{TAC}$. The observed input quantities of the three firms are indicated by black dots in Figure 1. Firms A and C have relatively large transaction costs, while firm B has relatively small transaction costs. The average relationship between the observed input quantities (x^{OBS}) and the output quantities (y) are indicated by the dotted curve. We call this relationship *augmented* production function and we claim that empirical production analysts usually estimate this augmented production function rather than the true production function $f(x^{PD}, T)$. The larger the

[3] The following derivations can also be calculated for multiple outputs, but for simplicity we only use a single output here.

input quantities that are used for trade activities and knowledge acquisition (x^{TAC}), the larger the difference between the true production function (solid curve) and the augmented production function (dotted curve). Our illustration also shows that differences in the firms' transaction costs affect the measured productivities of these firms. The production processes of all three firms in our example have an average productivity level compared to other firms of the same size, i.e. their points of production (grey dots) are located on the average true production function (solid line). However, an empirical production analyst will find that firms *A* and *C* that have relatively large transaction costs have a lower productivity compared to other firms of the same size, i.e. their observed points of production (black dots) are below the augmented production function (dotted curve), while firm *B* that has relatively small transaction costs has a higher productivity compared to other firms of the same size, i.e. its observed point of production (black dot) is above the augmented production function (dotted curve).

In the following, we derive a simple microeconomic model that takes transaction costs induced by innovation adoption and by trade activities into account.

2.1 Production technology and innovation

We assume that the adoption of productivity enhancing innovations is not cost free for the firm, but uses resources that can be of the same type as the inputs used for the production (e.g. labour, office supplies, IT technology, fuel). We denote these resources by $\mathbf{x}^{IN} = (x_1^{IN}, \ldots, x_n^{IN})'$, where the elements of $\mathbf{x}^{IN}$ correspond to the elements[4] of $\mathbf{x}^{PD}$ so that we can calculate the total input quantities that the firm uses for production and for improving the productivity of its production activities by $\mathbf{x}^{obs} = \mathbf{x}^{PD} + \mathbf{x}^{IN}$. The firm can utilise its network to improve the productivity of its production activities by gathering information from peers, which is otherwise difficult or costly to obtain or even unavailable. We assume that these relationships can be described by the function:

$$T = k(\mathbf{x}^{IN}, \mathbf{z}, \mathbf{u}), \tag{2}$$

where $\mathbf{z}$ is a vector of network parameters characterising the firm's networks and $\mathbf{u}$ is a vector of other factors that might affect the productivity of the firm's production activities or the resources that the firm needs to improve its productivity by a given level (e.g. the education of the management).

Datasets that are used for estimating production functions generally do not separate between input quantities used for the actual production ($\mathbf{x}^{PD}$) and input quantities used to improve the productivity of production activities ($\mathbf{x}^{IN}$). If no other transaction costs are included in the observed input quantities, they would be $\mathbf{x}^{obs} = \mathbf{x}^{PD} + \mathbf{x}^{IN}$. Therefore, the following approximation is necessary for empirical applications:

$$y = f(\mathbf{x}^{PD}, T) = f(\mathbf{x}^{PD}, k(\mathbf{x}^{IN}, \mathbf{z}, \mathbf{u})) \approx f^*(\mathbf{x}^{PD} + \mathbf{x}^{IN}, \mathbf{z}, \mathbf{u}) = f^*(\mathbf{x}^{obs}, \mathbf{z}, \mathbf{u}). \tag{3}$$

[4] Of course, some elements of $\mathbf{x}^{IN}$ might be zero (e.g. raw materials). If some inputs are only used for improving the productivity of the firm's production activities, but not in the actual production (e.g. advisory services or consulting), we can add further elements to the vector $\mathbf{x}^{PD}$ and set these elements to zero.

2.2 Transaction costs in trade

In addition to the resources required for the production ($\mathbf{x}^{PD}$) and for improving the productivity of the production activities ($\mathbf{x}^{IN}$), the firm needs further resources for trading goods, i.e. purchasing the inputs and selling the output. These resources can be of the same type as the inputs used for production (e.g. labour, capital, office supplies, IT technology, fuel). We denote the vector of resources used for trading goods by $\mathbf{x}^{TD} = (x_1^{TD}, \ldots, x_n^{TD})'$, where the elements of $\mathbf{x}^{TD}$ correspond to the elements[5] of $\mathbf{x}^{PD}$, $\mathbf{x}^{IN}$, and $\mathbf{x}^{obs}$. Hence, we can calculate the total input quantities that the firm acquires to produce the output, improve the productivity of its production activities, and to trade the goods by $\mathbf{x}^{OBS} = \mathbf{x}^{obs} + \mathbf{x}^{TD} = \mathbf{x}^{PD} + \mathbf{x}^{IN} + \mathbf{x}^{TD} = \mathbf{x}^{PD} + \mathbf{x}^{TAC}$. We expect that the quantities of the resources required for trading goods depend on the quantities of the traded goods. Furthermore, our considerations in the introductory section suggest that good networks can reduce the input quantities that are sacrificed for trading goods ($\mathbf{x}^{TD}$). We assume that the following set of (implicit) functions indicates the input quantities that are required for trading goods ($\mathbf{x}^{TD}$):

$$x_i^{TD} = g_i(\mathbf{x}^{OBS}, y, \mathbf{z}, \mathbf{v}) \; \forall \; i, \tag{4}$$

where $\mathbf{x}^{OBS} = \mathbf{x}^{PD} + \mathbf{x}^{IN} + \mathbf{x}^{TD}$ and y are the traded input and output quantities, respectively, $\mathbf{z}$ is—again—the vector of network parameters and $\mathbf{v}$ is a vector of other factors that might influence the resources required to trade the goods (e.g. heterogeneity of goods, distance to potential sellers and buyers).

In the following, we derive an augmented production function that only takes into account the observable input variables ($\mathbf{x}^{OBS}$). We start by substituting $x_i^{OBS} - x_i^{obs}$ for $\mathbf{x}^{TD}$ and $f^*(\mathbf{x}^{obs}, \mathbf{z}, \mathbf{u})$ for y in the set of equations (4) so that after rearranging we get:

$$x_i^{obs} = x_i^{OBS} - g_i(\mathbf{x}^{OBS}, f^*(\mathbf{x}^{obs}, \mathbf{z}, \mathbf{u}), \mathbf{z}, \mathbf{v}) \; \forall \; i. \tag{5}$$

By defining a set of functions $g_i^*(\mathbf{x}^{OBS}, \mathbf{x}^{obs}, \mathbf{z}, \mathbf{u}, \mathbf{v}) \equiv x_i^{OBS} - g_i(\mathbf{x}^{OBS}, f^*(\mathbf{x}^{obs}, \mathbf{z}, \mathbf{u}), \mathbf{z}, \mathbf{v})$ $\forall \; i$, we can rewrite the set of equations (5) to get a system of implicit functions for $\mathbf{x}^{obs}$:

$$x_i^{obs} = g_i^*(\mathbf{x}^{OBS}, \mathbf{x}^{obs}, \mathbf{z}, \mathbf{u}, \mathbf{v}) \; \forall \; i. \tag{6}$$

Then, we solve this system of implicit equations for $\mathbf{x}^{obs}$. We denote the resulting set of equations by:

$$x_i^{obs} \equiv h_i(\mathbf{x}^{OBS}, \mathbf{z}, \mathbf{u}, \mathbf{v}) \; \forall \; i. \tag{7}$$

Substituting these functions for $\mathbf{x}^{obs}$ in (3), we get

$$y = f^*(h(\mathbf{x}^{OBS}, \mathbf{z}, \mathbf{u}, \mathbf{v}), \mathbf{z}, \mathbf{u}) \equiv f^{**}(\mathbf{x}^{OBS}, \mathbf{z}, \mathbf{u}, \mathbf{v}), \tag{8}$$

where $\mathbf{x}^{OBS} = \mathbf{x}^{PD} + \mathbf{x}^{IN} + \mathbf{x}^{TD}$ corresponds to the input quantities that are usually observed in data sets used for empirical production analysis. Hence, the augmented production function $f^{**}(\mathbf{x}^{OBS}, \mathbf{z}, \mathbf{u}, \mathbf{v})$ not only includes the production process, but also the trading of goods and activities to improve the productivity of production activities. As we assume that firms with better networks need less resources for trading

[5] Of course, some elements of $\mathbf{x}^{TD}$ might be zero (e.g. raw materials).

goods and can improve the productivity of their production activities more easily and at less cost (see discussion in the introductory section), these firms should be able to produce the same amount of output (y) with smaller (total) input quantities ($\mathbf{x}^{OBS}$).

In order to empirically estimate the augmented production function $f^{**}(\mathbf{x}^{OBS}, \mathbf{z}, \mathbf{u}, \mathbf{v})$, we express the output and input quantities in equation (8) in logarithms and add an error term, ε, that accounts for unobserved factors:

$$\log y = f^{***}(\log \mathbf{x}^{OBS}, \mathbf{z}, \mathbf{u}, \mathbf{v}) + \varepsilon. \tag{9}$$

3 Data

In our empirical analysis, we use a cross-sectional dataset of 384 Polish farms. The data were collected in 2007 within the framework of the "Advanced-Eval" project financed by the European Union within the Sixth Framework Programme. The dataset includes detailed farm accountancy data from 2006 and information on the farms' ego-centred networks. We take the total value of all produced goods as output (in Złoty) and we distinguish between four inputs: labour (in working hours), land (in ha), capital (in Złoty), and intermediate inputs (in Złoty), where the intermediate inputs mainly consist of seeds, fertilisers, pesticides, purchased feed, fuel, and electricity.

Since Polish farms usually have a single farm manager, we do not have to model intra-firm networks, which can play an important role in information diffusion. Hence, our dataset has the advantage that we can neglect intra-firm networks when modelling networks.

The network data are collected through four different name generators[6] (Burt, 1984): trade network, information exchange network, social network, and farm-household network. To sample the *trade network*, we asked the farmer to name the most important trade partners on output as well as input markets; to sample the *information network*, we asked for the most important contacts with whom the farmer exchanges information on innovations or other important aspects of the business; to sample the *social network*, we asked for the contacts within the farmer's business network that were closer than mere business relations; finally, to sample the *farm-household network* we asked for the contacts that are mainly non-business related, like close family friends or contacts in unions and clubs. Of course, for some observations, the overlap between the networks can be considerable.

We apply two common network parameters for ego-centred networks to model the structure of each of the four farm networks, namely the number of *outdegrees* and the *density* of the network. The first network parameter refers to the total number of contacts (*alteri*) n that an *ego*—in our case the farm—has. The second network parameter, density, describes the degree of interconnectedness between *ego*'s *alteri*, $h \big/ [m(m-1)/2]$, where h is the actual number of ties between the *alteri* and $m(m-1)/2$ is the number of possible ties. Given the structure of ego-centred networks, the amount of structural information that can be derived is limited compared to what can

[6] Name generators are a sampling technique to map ego-centred networks through a battery of questions, e.g., 'Whom do you contact when looking for a job' or 'Name your most important trade partners'.

be ascertained from full network samples. However, the geographic distribution of the sampled farms prevented the sampling of a full network for our study. Furthermore, sampling ego-centred networks with a name generator tends to elicit strong ties[7] (Lin, 1999), which might affect the structural measures; this should be kept in mind when interpreting the results.

The variables that might affect the productivity of the firm's production activities or the resources that the firm needs to improve its productivity ($\mathbf{u}$) or that might influence the resources required to trade the goods ($\mathbf{v}$) include the region in which the farm is located as well as management characteristics, namely the farmer's level of education, work experience (in years) and attitude to risk. The latter is the average response to several lottery questions for assessing risk attitude, where larger positive values indicate higher risk aversion. Our data include farms from four different municipalities (Gminas). The municipalities Chotcza and Wieliszew are located close to urban areas, while Siemiątkowo and Kamieniec are located in remote areas. While Wieliszew and Kamieniec perform well economically, Chotcza and Siemiątkowo's economic performance is weak. Hence, the four municipalities cover all possible combinations of location and economic performance. Of course, this regional variable also accounts for differences in climate and soil, but we cannot differentiate between these effects. Whilst a separation of these effects would be interesting, it is not essential for our study. Descriptive statistics of the data set are given in Table 1.

4 Econometric Specification

If our considerations about transaction costs and networks are correct and we use a typical data set, where the input quantities include resources used for the production ($\mathbf{x}^{PD}$), resources used to improve the productivity of production activities ($\mathbf{x}^{IN}$), and resources used for trading goods ($\mathbf{x}^{TD}$), the production function should not only depend on the input quantities, but also on the firm's network position. Hence, we can test the hypothesis that networks influence transaction costs by estimating the augmented production function $\log y = f^{***}(\log \mathbf{x}^{OBS}, \mathbf{z}, \mathbf{u}, \mathbf{v})$ defined in (9) and testing if the network parameters $\mathbf{z}$ have a significant influence.

Given our microeconomic model derived above, the relationship between the total input quantities $\mathbf{x}^{OBS}$, the network parameters $\mathbf{z}$, the other factors $\mathbf{u}$ and $\mathbf{v}$, and the output quantity y is unknown and could be rather complex. To minimise the risk that our results depend on an unsuitable econometric specification, we estimate this augmented production function with different parametric, semiparametric, and nonparametric specifications of the regression function. While the parametric estimations may suffer from an unsuitable parametric specification, the fully nonparametric estimations may suffer from the curse of dimensionality. The semiparametric specifications take a middle ground between the fully parametric and the fully nonparametric specifications. We use the following model specifications:

[7] *Strong ties* are contacts that interact closely and frequently. An over-representation of strong ties in the sample can eventually lead to denser networks.

Table 1: Descriptive statistics

variable	n	mean	std dev	median	min	max
output	384	36083.60	45645.82	18150.00	240.00	323500.00
log(output)	384	9.79	1.30	9.81	5.48	12.69
labour	384	6402.36	8296.19	5305.25	704.00	99360.00
log(labour)	384	8.53	0.62	8.58	6.56	11.51
land	384	11.80	10.15	9.89	0.30	80.00
log(land)	384	2.15	0.84	2.29	-1.20	4.38
capital	384	323482.16	411592.21	188000.00	7500.00	3150000.00
log(capital)	384	12.15	1.05	12.14	8.92	14.96
intermed	384	41238.64	137703.93	15507.14	670.00	2515288.21
log(intermed)	384	9.79	1.14	9.65	6.51	14.70
municip	384					
Chotcza	109					
Kamieniec	99					
Siemiątkowo	75					
Wieliszew	101					
exper	384	24.94	12.00	24.00	1.00	79.00
education	374					
none	151					
apprenticeship	92					
vocational school	129					
university	2					
risk	384	-0.06	0.75	0.09	-2.04	1.34
outd Trade	384	3.54	1.42	3.00	0.00	9.00
dens Trade	384	0.35	0.29	0.42	0.00	1.00
outd Inf	384	0.26	0.61	0.00	0.00	4.00
dens Inf	384	0.02	0.13	0.00	0.00	1.00
outd Soc	384	0.14	0.45	0.00	0.00	3.00
dens Soc	384	0.02	0.15	0.00	0.00	1.00
outd HH	384	3.24	1.40	3.00	2.00	8.00
dens HH	384	0.47	0.43	0.50	0.00	1.00

Note: variable "output" indicates the aggregate output quantity measured in Złoty (PLN); "labour" indicates the total on-farm labour measured in hours/year; variable "land" is the farm's total utilised agricultural land area measured in ha; "capital" indicates the farm's capital stock measured in Złoty (PLN); "intermed" indicates the farm's use of intermediate inputs (e.g. feed, fertiliser, pesticides, fuel) measured in Złoty (PLN); "municip" indicates the municipality, in which the farm is located; "exper" indicates the farmer's farming experience in years; "education" indicates the farmer's agricultural education; "risk" indicates the farmer's attitude to risk, where larger positive values indicate higher risk aversion; the abbreviations "outd" and "dens" indicate the number of outdegrees and the density of the network, respectively, where "Trade," "Inf," "Soc," and "HH" indicate the firm's trade network, the firm's information network, the firm's social network, and the network of the corresponding household, respectively.

– **Parametric specifications:** Ordinary Least-Squares estimations of Cobb-Douglas and Translog production functions with location, management characteristics, and the network variables as additional (linearly modelled) explanatory variables.
– **Semiparametric specifications:** additive categorical regression spline models, where the continuous explanatory variables, i.e. input quantities, experience, risk attitudes, and the network variables, are modelled with B-splines and the categor-

ical explanatory variables, i.e. location and education, are modelled with kernel smoothing (Ma et al, 2012; Ma and Racine, 2012).[8]

– **Nonparametric specifications**: local-linear kernel estimations using the estimation method for both continuous and categorical explanatory variables (Li and Racine, 2004; Racine and Li, 2004).

We use the kernel proposed by Li and Racine (2004) for the unordered categorical explanatory variable, i.e. location, while the kernel proposed by Racine and Li (2004) is used for the ordered categorical explanatory variable, i.e. education, both in the semiparametric and in the nonparametric models.[9] The second-order Epanechnikov kernel is used for all continuous regressors in the nonparametric models. We make the frequently used assumption that the bandwidths of the kernels can differ between regressors, but are constant over the domain of each regressor. The optimal degree of smoothing—i.e. the optimal degrees and the optimal numbers of segments of the splines for the continuous explanatory variables and the optimal bandwidths for the categorical explanatory variables in the semiparametric models and the optimal bandwidths for the continuous and the categorical explanatory variables in the nonparametric models—is selected either by using least-squares leave-one-out cross-validation or according to the expected Kullback-Leibler criterion (Hurvich et al, 1998).

In the additive categorical regression spline models, we test the statistical significance of the explanatory variables with the significance test suggested by Ma and

[8] We also estimated other semiparametric models, but they were less suitable for our empirical application. For instance, we estimated smooth coefficient models, where the input quantities were modelled in parametric (Cobb-Douglas) form and the coefficients were allowed to vary based on location, education, experience, risk attitudes, and network parameters—and in some models also on the (logarithmic) input quantities in order to allow for the same flexibility in input quantities as a Translog production function. Unfortunately, our data set does not have a sufficient number of observations to obtain reasonable estimation results from this specification. Even in a fully parametric specification with a Cobb-Douglas production function, where the five smooth coefficients (intercept and the slope coefficients of the four logarithmic input quantities) are linear functions of the three location dummies, three education dummies, experience, risk attitudes, and eight network parameters, the model has 85 coefficients. Allowing for the same flexibility in input quantities as a Translog production function would require 16 additional coefficients. Modelling the smooth coefficients with a non-parametric approach further aggravates this problem. Therefore, a smooth coefficient model is unsuitable for our empirical application.
Furthermore, we estimated various partially linear models, where the (logarithmic) input quantities were modelled linearly (i.e. assuming a Cobb-Douglas functional form). These models had an extremely poor out-of-sample predictive performance (based on the procedure described in Section 5.2), which may not be surprising, as the results of most other models show that the effects of the (logarithmic) input quantities are nonlinear (see Section 5.1).
Finally, we estimated various partially linear models, where the network variables were modelled linearly, while the input quantities and the other explanatory variables were modelled nonparametrically. The results of these models were mostly similar to the results of the models that we present in Section 5. However, as these models are unable to detect nonlinearities in the effects of the network variables, we do not present the results of these models.

[9] While the kernels for ordered categorical variables of Wang and van Ryzin (1981) and Racine and Li (2004) are clearly different, the kernels for unordered categorical variables of Aitchison and Aitken (1976) and Li and Racine (2004) have exactly the same shape, but a different specification of the bandwidth parameter so that the choice between the kernels for unordered categorical variables does not make a difference if the bandwidth parameters are appropriately adjusted, e.g. by data-driven bandwidth selection (Czekaj and Henningsen, 2013).

Racine (2011), which is analogous to a simple t-test in a parametric regression setting, but which obtains the distribution of the test statistic under the null hypothesis by the 'residual' bootstrap method using independent identically distributed (iid) draws.

In the nonparametric models, we test the statistical significance of the regressors with the nonparametric test for irrelevant regressors that was initially suggested by Racine (1997) and later extended to categorical regressors by Racine et al (2006). This test checks if the sum (over all observations) of the squared (marginal) effects of a specific regressor on the dependent variable is significantly larger than zero, where the distribution of the test statistic under the null hypothesis is obtained by the bootstrap method using independent identically distributed (iid) draws.

In all semiparametric and nonparametric estimations, we take the logarithm of the output and all the input quantities. This makes it easier to compare the semiparametric and the nonparametric estimations with the parametric estimations, because the Cobb-Douglas and Translog functional forms also use logarithmic output and input quantities. Furthermore, the individual values of the logarithmic variables are more equally distributed within the range of observed values than the original (non-logarithmic) values. If we used original (non-logarithmic) input quantities in our data set, there would be many observations within the bandwidths for small values (farms), but very few observations within the bandwidth for large values (farms), which usually causes problems in nonparametric regression with fixed (constant) bandwidths. Finally, the unknown true augmented production function is probably more similar to a log-linear (Cobb-Douglas) function than to a linear function (which implies perfect substitutability between inputs) so that the use of a local-log-linear (local-Cobb-Douglas) specification using logarithmic quantities of the inputs and the output converges faster to the true augmented production function than a local-linear specification using original (non-logarithmic) quantities of the inputs and the output (Czekaj and Henningsen, 2013).

5 Results

All estimations were performed within the statistical software environment "R" (R Core Team, 2014) using the add-on packages "crs" (Nie and Racine, 2012) and "np" (Hayfield and Racine, 2008).

Initially, our model also included the farmer's education and his or her risk attitudes as explanatory variables, but we removed these explanatory variables, because they do not have a statistically significant effect in any of the estimated models.[10]

5.1 Degrees and segments of splines, bandwidths, and statistical significance

Table 2 presents the degrees and the numbers of segments of the splines and the bandwidths of the kernels for the explanatory variables that we obtained by least-squares cross-validation or according to the expected Kullback-Leibler criterion (Hurvich

[10] We were not able to test the statistical significance in the additive categorical regression spline models due to high multicollinearity.

Table 2: Degrees and segments of splines, bandwidths, and statistical significance

	labor	land	capital	intermed	municip	exper
models with all explanatory variables						
OLS CD	***	***	***	***	***	*
OLS TL	***	***		***	***	**
ACRS LS	6/1***	2/1***	6/1***	3/2***	0.41***	6/1***
ACRS AIC	4/2***	3/1***	5/1***	4/1***	0.21***	3/3***
NP LS	∞***	∞***	∞***	∞***	0.30**	∞*
NP AIC	∞***	∞***	∞***	∞***	0.33**	∞*
models with fewer explanatory variables						
OLS CD	***	***	***	***	***	*
OLS TL	***	***	*	***	***	*
ACRS LS	6/1***	2/1***	6/1***	3/2***	0.45***	1/3**
ACRS AIC	4/2***	1/1***	5/1***	4/1***	0.21***	3/3***
NP LS	0.94***	0.88***	0.93***	1.42***	0.21***	26.39**
NP AIC	0.92***	1.38***	1.03***	1.35***	0.17***	27.73**

	outd Trade	dens Trade	outd Inf	dens Inf	outd Soc	dens Soc	outd HH	dens HH
models with all explanatory variables								
OLS CD	***							***
OLS TL	***							***
ACRS LS	1/1***	0/1*	0/1	4/1	0/1	1/1*	0/1	1/1***
ACRS AIC	1/1***	1/1*	0/1	0/1	0/1	0/1	0/1	1/1**
NP LS	∞***	∞**	∞	∞*	∞	∞	∞	∞***
NP AIC	∞***	∞**	∞	∞*	∞	∞	∞	∞***
models with fewer explanatory variables								
OLS CD	***		—	—	—	—	—	***
OLS TL	***		—	—	—	—	—	***
ACRS LS	4/1***	1/1**	—	—	—	—	—	4/1***
ACRS AIC	1/1***	1/1**	—	—	—	—	—	1/1**
NP LS	4.98***	∞***	—	—	—	—	—	∞***
NP AIC	∞***	∞***	—	—	—	—	—	∞***

Note: in the upper panel, the first four columns indicate the degrees and segments of the splines or the bandwidths for the four logarithmic input quantities (labour, land, capital, intermediate inputs); column "municip" indicates the bandwidths for the location variable, and column "exper" indicates the degrees and segments of the splines or the bandwidths for the experience (in years). The lower panel indicates the degrees and segments of the splines or the bandwidths for the network variables (see explanations below Table 1). The abbreviations "CD OLS" and "TL OLS" indicate the Cobb-Douglas and Translog functional forms estimated by OLS; "ACRS" indicates additive categorical regression spline models and "NP" indicates a fully nonparametric local-linear model, where the additions "LS" and "AIC" indicate selection of the degrees and segments of the splines and the bandwidths of the kernels by least-squares cross-validation and according to the expected Kullback-Leibler criterion (Hurvich et al, 1998), respectively. In the rows for the additive categorical regression spline (ACRS) models, the first value indicates the degree of the spline and the second value indicates the number of segments; The infinity symbol (∞) indicates that the bandwidth that was chosen by the bandwidth selection procedure is at least 5 times as large as the standard deviation of the corresponding variable. A dash (—) indicates that the explanatory variable is not included in the model. Asterisks indicate the statistical significance of the explanatory variables where * = significant at 10%, ** = significant at 5%, and *** = significant at 1%.

et al, 1998). Furthermore, this table indicates the statistical significance of the explanatory variables. While the five network variables "outd Inf," "dens Inf," "outd Soc," "dens Soc," and "dens HH" are statistically insignificant at the 5% level in all six model specifications, all other explanatory variables are statistically significant at the 5% level in at least two model specifications.[11] Therefore, we re-estimated all six model specifications without the five insignificant network variables. In these smaller models, all explanatory variables are statistically significant at the 5% level in at least four out of the six model specifications.

The degrees and the numbers of segments of the splines in the additive categorical regression spline models (ACRS) allow for nonlinearities in nearly all logarithmic input quantities and the farmer's experience, while the effects of most network variables are either linear (one segment with degree one) or absent (one segment with degree zero). However, the least-squares cross-validation of the smaller ACRS model indicates that the effect of the outdegrees of the trade network ("outd Trade") and the density of the household network ("dens HH") are nonlinear.

In the nonparametric models (NP) with all explanatory variables, the bandwidths of all continuous explanatory variables are at least 5 times as large as their standard deviations so that these two models are linear in all continuous explanatory variables.[12] These large bandwidths are at least to some extent caused by the curse of dimensionality, because the large number of explanatory variables (13 continuous and one categorical explanatory variable) and the relatively small number of observations (384) make it difficult to detect nonlinearities in a local-linear estimation. After removing the five insignificant network variables, the bandwidths for the input quantities are all in the order of magnitude of their standard deviations, which allows for nonlinearities in the input variables. The bandwidths for experience in these two models are about twice the standard deviation of this variable, which allows for moderate linearity in experience. The bandwidths of all three remaining network variables are all at least 3.5 times as large as their standard deviations so that these two models are (nearly) linear in all network variables.

In all specifications of the additive categorical regression spline models (ACRS) and in the nonparametric models (NP), the bandwidth of the location (municipality) is clearly smaller than one, which indicates that the location has a noticeable influence on the augmented production function.

[11] The five individually insignificant network variables are also jointly insignificant in the OLS Cobb-Douglas model (*P*-value 0.145), in the OLS Translog model (*P*-value 0.255), in the nonparametric model with bandwidths obtained by least-squares cross-validation (*P*-value 0.189), and in the nonparametric model with bandwidths obtained according to the expected Kullback-Leibler criterion (*P*-value 0.180). We have not tested the joint significance in the ACRS models, because this feature is not yet available in the "crs" package.

[12] However, in contrast to a parametric linear regression (e.g. OLS), our nonparametric regression with large bandwidths still allows the marginal effects of the explanatory variables to differ between observations. In fact, each nonparametric estimation is similar to four linear (Cobb-Douglas) estimations for the four municipalities, where the estimation for each municipality also takes into account the observations of the three other municipalities using weights equal to the bandwidth of the location variable.

5.2 Evaluation of model specifications

In this section, we discuss, which model specification is the most suitable. Table 3 presents the *P*-values from specification tests of the parametric models. When comparing the parametric specifications, likelihood ratio tests clearly reject the Cobb-Douglas functional form in favour of the Translog production function. However, both the Cobb-Douglas and the Translog specifications are rejected by Ramsey's (1969) Regression Equation Specification Error Test (RESET) and by Hsiao et al's (2007) kernel-based specification test against the local-constant specification. These specification tests create some doubt about the suitability of the parametric specifications.

Table 3: Specification tests of the parametric models (*P*-values)

	Translog	RESET	NP LC	NP LL
models with all explanatory variables				
OLS CD	0.006	0.040	0.010	0.383
OLS TL		0.022	0.003	0.000
models with fewer explanatory variables				
OLS CD	0.003	0.035	0.003	0.336
OLS TL		0.016	0.005	0.749

Note: column "Translog" presents the *P*-values of likelihood ratio tests of the Cobb-Douglas models against the corresponding Translog models; column "RESET" presents the *P*-values of Ramsey's (1969) Regression Equation Specification Error Test (RESET) with squared and cubic fitted values as additional explanatory variables; columns "NP LC" and "NP LL" present the *P*-values from the kernel-based specification test of Hsiao et al (2007) against local-constant and local-linear alternatives, respectively, where we use the same kernel functions as in the nonparametric specifications, select the bandwidths by least-squares cross validation, and estimate the distribution under the null hypothesis by the bootstrap method with independent identically distributed (iid) draws.

In order to evaluate the semiparametric and nonparametric model specifications and to compare them with each other and the parametric specifications, we use the 'test for revealed performance' proposed by Racine and Parmeter (2014). This test assesses how close the different model specifications are expected to lie to the unknown 'true' data generating process. For the convenience of the reader, we briefly summarise the procedure here, while we refer the reader to Racine and Parmeter (2014) for a detailed description, the statistical background, and proofs:

1. All model specifications are estimated using data-driven methods to select the degrees and numbers of segments of the splines and the bandwidths of the kernels (see Section 5.1).
2. We randomly split our sample into two independent subsamples with sizes n_1 and n_2, respectively, where $n_1 + n_2 = n = 384$ is our total sample size.
3. We use the n_1 observations in the first subsample to re-estimate all model specifications using the same degrees and numbers of segments of the splines and the same degree of kernel smoothing as in step 1, i.e. the bandwidths of the kernel functions are adjusted to the smaller sample size and the different spreading of

the variables in the subsample so that the scaling factors are the same as in the model for the full sample.[13]
4. The models estimated with the first subsample and the explanatory variables of the n_2 observations in the second subsample are used to predict the values of the dependent variable in the second subsample.
5. The out-of-sample predictive performance of each model specification is evaluated by calculating the average squared prediction error (ASPE) with ASPE $= n_2^{-1} \sum_{j=n_1+1}^{n} (\hat{y}_j - y_j)^2$, where $\hat{y}_j$; $j = n_1 + 1, \ldots, n$ are the values of the dependent variable predicted by a model that was estimated without observations $j = n_j + j, \ldots, n$.
6. We repeat steps two to five 5,000 times and use the obtained ASPEs to construct the empirical distribution function of the variance of the expected true error for each model specification.
7. We compare the distributions of the ASPEs for the different model specifications.

Table 4 presents and compares the ASPEs of the estimated model specifications for evaluation samples of sizes $n_2 = 5$, $n_2 = 25$, and $n_2 = 50$. The OLS Cobb-Douglas model with all explanatory variables is the best performing parametric model for all three evaluation sample sizes, although it includes several statistically insignificant regressors and the Cobb-Douglas model was clearly rejected in favour of the Translog model. While the nonparametric models NP LS and NP AIC with all explanatory variables and the two ACRS AIC models clearly perform worse than the best parametric model, the nonparametric models NP LS and NP AIC with fewer explanatory variables and the two ACRS LS models outperform the best parametric model in three to five out of the six performance measures. According to the tests proposed by Racine and Parmeter (2014), the two ACRS LS models clearly have the best predictive performance, where the ACRS LS model with all explanatory variables has the lowest ASPE according to three test criteria and the ACRS LS model with fewer explanatory variables has the lowest ASPE according to two test criteria.

5.3 Effects of the explanatory variables on output

The median values of the marginal effects (gradients) of the explanatory variables on the (logarithmic) output quantity obtained by the 12 different model specifications are presented in Table 5. This table also shows the marginal significance levels (P-values) of the explanatory variables that we already presented in Table 2. The models generally give rather similar results, which indicates that most of our estimation results are robust to different model specifications.

The marginal effects of the logarithmic input quantities on the logarithmic output quantity coincide with the partial production elasticities of the inputs. However, in contrast to the classical definition of partial production elasticities, in our empirical

[13] It would also be desirable to adjust the number of segments and the degrees of the spline functions, but to our knowledge, no procedure for this exists and the number of segments and the degrees of the spline functions must be integers so that their adjustment by the same factors as the adjustments of the bandwidths of the kernel functions would probably in most cases not have been sufficient to adjust them (after rounding) to the previous or next integer value.

Table 4: Average mean prediction errors

	$n_2 = 5$		$n_2 = 25$		$n_2 = 50$	
	mean	trimmed mean	mean	trimmed mean	mean	trimmed mean
models with all explanatory variables						
OLS CD	0.714	0.616	0.718	0.685	0.718	0.696
OLS TL	0.733^{***}	0.622	0.733^{***}	0.695^{**}	0.736^{***}	0.711^{***}
ACRS LS	0.663_{***}	0.578_{***}	0.707	0.652_{***}	1.288^{*}	0.686_{***}
ACRS AIC	0.736^{***}	0.623	0.752^{***}	0.704^{***}	0.774^{***}	0.734^{***}
NP LS	0.731^{***}	0.631^{**}	0.734^{***}	0.701^{***}	0.733^{***}	0.712^{***}
NP AIC	0.731^{***}	0.631^{**}	0.734^{***}	0.701^{***}	0.733^{***}	0.712^{***}
models with fewer explanatory variables						
OLS CD	0.718^{***}	0.619	0.720^{***}	0.687	0.720^{***}	0.698
OLS TL	0.734^{***}	0.621	0.733^{***}	0.695^{**}	0.736^{***}	0.711^{***}
ACRS LS	0.670_{***}	0.589_{***}	0.687_{***}	0.657_{***}	0.971	0.679_{***}
ACRS AIC	0.745^{***}	0.623	0.758^{***}	0.706^{***}	0.777^{***}	0.734^{***}
NP LS	0.686_{***}	0.586_{***}	0.716	0.668_{***}	8.086^{*}	0.694
NP AIC	0.709_{**}	0.599_{**}	0.722	0.679_{*}	6.001^{*}	0.700

Notes: The values in the "mean" columns indicate the mean ASPEs over all 5,000 replications; the values in the "trimmed mean" columns indicate the mean ASPEs after removing the 5% of the replications with the largest ASPEs. Asterisks indicate the results of one-sided t-tests for the equality of the (trimmed) mean ASPE of each model specification with the (trimmed) mean ASPE of the best performing parametric model, i.e. the OLS Cobb-Douglas model with all explanatory variables, where * = significant at 10%, ** = significant at 5%, and *** = significant at 1%. The mean values are compared by paired t-tests, while the trimmed mean values are compared by non-paired t-tests (allowing for different variances in the ASPEs of the two compared model specifications), because in the different model specifications, different replications are 'trimmed'. If the (trimmed) mean value of a model specification is less [greater] than the (trimmed) mean value of the OLS Cobb-Douglas model with all explanatory variables, the alternative of the one-sided t-test is that the (trimmed) mean value of this model specification is less [greater] than the (trimmed) mean value of the OLS Cobb-Douglas model with all explanatory variables and the significance level is indicated in a subscript [superscript]. The abbreviations of the model specifications are the same as in Table 2.

application (and in most other empirical applications as well), the estimated partial production elasticities not only take into account the actual production process, but also activities for improving productivity and trading goods. The estimation results of all of our models indicate that intermediate inputs have the largest partial production elasticity, while land and capital have the lowest partial production elasticities.

All models indicate that farmers in Kamieniec and Wieliszew can produce much more output with the same amount of resources (for improving productivity, trading goods, and producing outputs) than farmers in Chotcza and Siemiątkowo.[14] However, the estimated size of the difference between the municipalities clearly differs between models: the estimated productivity differences are generally much larger in the parametric models (OLS) than in the semiparametric models (ACRS) and the nonparametric models (NP).

All models indicate that the median effect of the farmer's experience (exper) on productivity is negative. However, the four ACRS models indicate that the effect of

[14] Please note that in the ACRS and NP models, the significance levels refer to the significance of the location variable, not to the significance of the effect of the individual municipalities. Thus, for these models, a significance symbol for Siemiątkowo does not mean that the productivity significantly differs between Siemiątkowo and Chotcza, but that there are significant productivity differences between at least two municipalities.

Table 5: Median effects of explanatory variables on output

	labor	land	capital	intermed	kami	siem	wiel	exper
models with all explanatory variables								
OLS CD	0.291***	0.276***	0.165***	0.395***	0.605***	0.143	0.380***	-0.006*
OLS TL	0.244***	0.272***	0.130	0.442***	0.615***	0.138	0.457***	-0.007**
ACRS LS	0.427***	0.160***	0.304***	0.492***	0.097***	-0.024***	0.175***	-0.003***
ACRS AIC	0.371***	0.168***	0.235***	0.479***	0.167***	-0.072***	0.236***	-0.007***
NP LS	0.288***	0.216***	0.211***	0.437***	0.239**	-0.011**	0.105**	-0.008*
NP AIC	0.286***	0.217***	0.215***	0.436***	0.222**	-0.011**	0.097**	-0.008*
models with fewer explanatory variables								
OLS CD	0.284***	0.278***	0.170***	0.396***	0.602***	0.131	0.404***	-0.007*
OLS TL	0.235***	0.282***	0.137*	0.460***	0.609***	0.129	0.485***	-0.007*
ACRS LS	0.366***	0.195***	0.310***	0.470***	0.108***	-0.041***	0.090***	-0.012**
ACRS AIC	0.379***	0.245***	0.235***	0.448***	0.192***	-0.066***	0.229***	-0.008***
NP LS	0.312***	0.294***	0.240***	0.437***	0.219***	-0.042***	0.168***	-0.008**
NP AIC	0.303***	0.284***	0.249***	0.436***	0.276***	-0.029***	0.239***	-0.008**

	outd Trade	dens Trade	outd Inf	dens Inf	outd Soc	dens Soc	outd HH	dens HH
models with all explanatory variables								
OLS CD	0.146***	0.200	0.035	0.678	-0.040	-0.365	0.046	0.329***
OLS TL	0.136***	0.207	0.056	0.459	-0.014	-0.407	0.041	0.327***
ACRS LS	0.162***	0.000*	0.000	-0.388	0.000	-0.625*	0.000	0.318***
ACRS AIC	0.162***	0.469*	0.000	0.000	0.000	0.000	0.000	0.270**
NP LS	0.161***	0.451**	0.016	0.706*	-0.111	-0.467	0.044	0.395***
NP AIC	0.160***	0.444**	0.017	0.716*	-0.111	-0.466	0.043	0.396***
models with fewer explanatory variables								
OLS CD	0.148***	0.229						0.316***
OLS TL	0.138***	0.233						0.314***
ACRS LS	0.175***	0.436**						0.732***
ACRS AIC	0.163***	0.472**						0.265**
NP LS	0.114***	0.484***						0.281***
NP AIC	0.115***	0.484***						0.270***

Note: the columns of the four inputs (labour, land, capital, intermediate inputs) indicate the median values of their partial production elasticities; columns "kami," "siem," and "wiel" indicate the median differences in the logarithmic output of the municipalities Kamieniec, Siemiątkowo, and Wieliszew, respectively, compared to the municipality Chotcza (ceteris paribus); column "exper" indicates the median semi-elasticities of the experience (in years) on the (logarithmic) output. The abbreviations of the network variables are described below Table 1; the columns of these network variables indicate their median semi-elasticities on the (logarithmic) output. The abbreviations of the model specifications are explained below Table 2. Asterisks indicate the statistical significance of the corresponding variables (not the statistical significance of the median effects), where * = significant at 10%, ** = significant at 5%, and *** = significant at 1%.

experience is positive for many farmers with between 15 and 30 years' experience, but is mostly negative for farmers with less or more experience.

As management characteristics (exper) and the location of the farm (municip) are included both in vector **v** and in vector **u**, the above-mentioned total effects of these variables comprise their direct effect on the productivity of the production process ($\partial T/\partial \mathbf{u}$), their effect through resources required to improve the productivity of production activities ($\partial(\partial T/\partial \mathbf{x}^{IN})/\partial \mathbf{u}$), and their effect on resources required for trade ($\partial \mathbf{x}^{TD}/\partial \mathbf{v}$).

The estimated effects of the network parameters are mostly consistent between model specifications: the estimation results of all models agree that the number of outdegrees of the trading network of the farm (outd Trade) and the density of the household network (dens HH) have a rather large positive and highly statistically significant effect on the productivity. According to most models, an additional trading partner increases the farm output by around 15%, while increasing the density of the household network from zero (a totally loose network without any connection between the *alteri*) to one (a totally dense network with all *alteri* connected) would increase the output on average by 30–40%.[15] While most models indicate a linear effect of these two network variables, model ACRS LS with fewer explanatory variables indicates nonlinear effects. However, in this model, the effect of the outdegrees of the trade network (outd Trade) is nearly linear for two or more trading partners, whereas negative effects can only be observed for a very few observations with less than two trading partners. Furthermore, this model indicates considerable nonlinearities in the density of the household network (dens HH) with highest productivities at densities of around 0.15 and 0.9 and lowest productivities at a density of around 0.5. However, these nonlinearities are only based on about one third of the observations, because about two thirds of the observations have a density of either zero or one and thus, do not directly affect the shape of the regression line between the minimum and the maximum of this variable. Hence, also in model ACRS LS with fewer explanatory variables, there is only weak evidence for nonlinear effects of network variables.

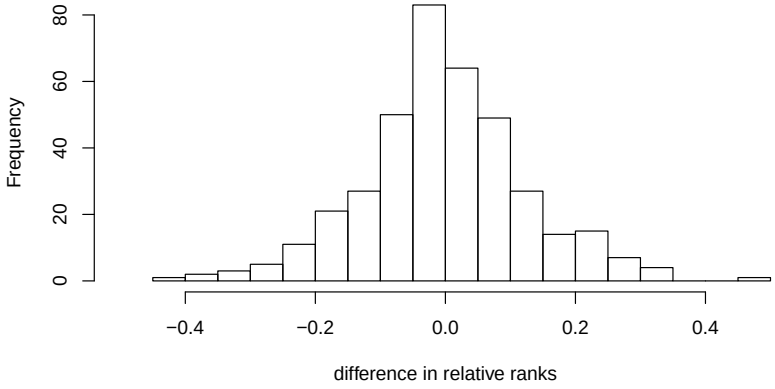
Several models indicate that the density of the trading network of the farm (dens Trade) also has a significant positive effect on productivity, but the estimated size of this effect varies considerably between models.

5.4 Effects of networks on the productivity ranking

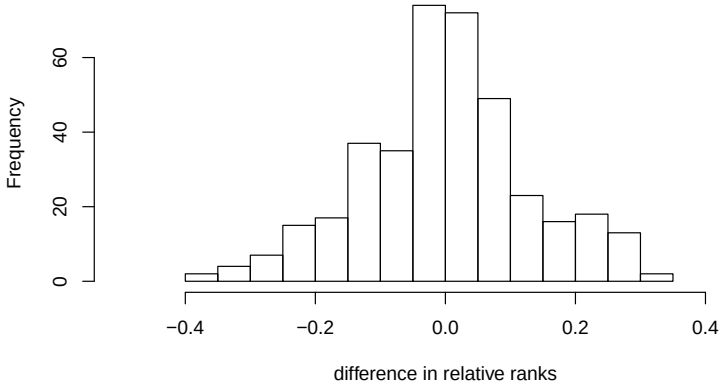
To test how the exclusion of networks, as a proxy for transaction costs, affects the productivity ranking of the farms, we compare the ranking of the residuals of the ACRS LS specification with fewer explanatory variables[16] with the ranking of the residuals of a corresponding model that omits all network variables, i.e. we compare the ranking of the residuals $\log y - \hat{f}^{***}(\log \mathbf{x}^{OBS}, \mathbf{z}, \mathbf{u}, \mathbf{v})$ with the ranking of the residuals $\log y - \hat{g}(\log \mathbf{x}^{OBS}, \mathbf{u}, \mathbf{v})$, where $\hat{f}^{***}(\log \mathbf{x}^{OBS}, \mathbf{z}, \mathbf{u}, \mathbf{v})$ is our estimated (logarithmic) augmented production function as defined in equation (9) and $\hat{g}(\log \mathbf{x}^{OBS}, \mathbf{u}, \mathbf{v})$ is a traditional production function estimated without network variables ($\mathbf{z}$). Figure 2a illustrates the distributions of the normalised differences, i.e. the differences between the rankings divided by the number of farms, between the rankings of both models. Although the differences in the ranking position are marginal or nil for many obser-

[15] Please note that a gradient of β means that the effect of a change from zero to one is a change of $100 \cdot (\exp(\beta) - 1)\%$, while the effect of a change from one to zero is a change of $100 \cdot (\exp(-\beta) - 1)\%$, because the dependent variable (output) is logarithmised and most of our models are (virtually) linear in the (majority of) network parameters.

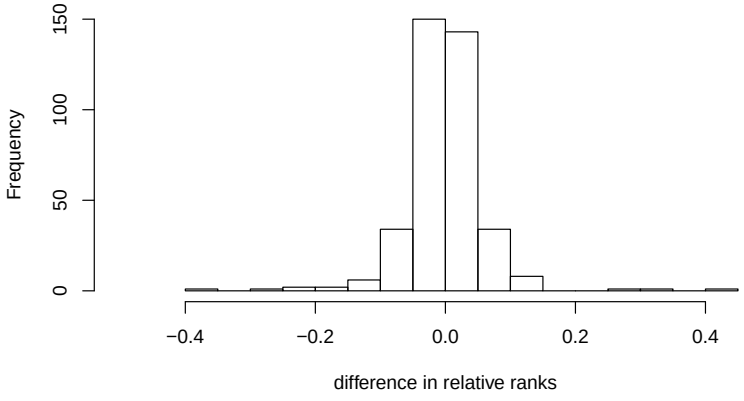
[16] We did the same analysis based on the ACRS LS specification with all explanatory variables, but we only present the analysis based on the ACRS LS specification with fewer explanatory variables, because the resulting figures only slightly differ between the two ACRS LS specifications.



(a) Networks vs. no networks



(b) Networks vs. no variation in networks



(c) Differences between both approaches

Fig. 2: Comparison of productivity rankings

vations, still more than a third of the observations exhibit a considerable shift in their ranking position at the boundary areas by up to 50 percentiles.

However, as we compare the residuals from estimations of two different model specifications, we may simply measure the effect of adding additional regressors to the model, which may affect the residuals even though their explanatory value is nil. To test the robustness of our findings, we repeat the exercise by comparing the ranking of the residuals from our estimated ACRS LS model with fewer explanatory variables with the ranking of artificially constructed residuals from this model when levelling out differences in the network structure. These artificially constructed residuals are taken to be the difference between the observed output of the farm and the predicted output at the observed input quantities with all network variables being equal to the sample medians, i.e. we compare the ranking of the residuals $\log y - \hat{f}^{***}(\log \mathbf{x}^{OBS}, \mathbf{z}, \mathbf{u}, \mathbf{v})$ with the ranking of the residuals $\log y - \hat{f}^{***}(\log \mathbf{x}^{OBS}, \bar{\mathbf{z}}, \mathbf{u}, \mathbf{v})$, where $\bar{\mathbf{z}}$ is a vector of the median values of the network variables. Figure 2b illustrates the distributions over the normalised differences in the ranking positions of both models. Although the overall change is not dramatic, the variance of the differences in Figure 2b is slightly larger than in Figure 2a.

Finally, we compare the ranking of the artificially constructed residuals with the ranking of the model estimated without network variables, i.e. we compare the ranking of the residuals $\log y - \hat{f}^{***}(\log \mathbf{x}^{OBS}, \bar{\mathbf{z}}, \mathbf{u}, \mathbf{v})$ with the ranking of the residuals $\log y - \hat{g}(\log \mathbf{x}^{OBS}, \mathbf{u}, \mathbf{v})$. Figure 2c shows that for the majority of the observations, the rankings only differ marginally. Hence, our results strongly indicate that omitting networks, or for that matter transaction costs, distorts the observed productivity ranking.

6 Discussion

Based on a sample of 384 Polish farms, we have analysed farm productivity under the aspect of transactions costs. As transaction costs are usually latent variables, we have chosen social networks as a proxy for transaction costs; earlier studies (e.g., Di Matteo et al, 2005; Henning et al, 2012) have shown that social networks can reduce transaction costs by positively affecting information acquisition or trade activities. Because the empirical relationship between structural parameters of social networks and productivity is possibly non-linear (Stam et al, 2013; Yu and Chiu, 2013), we have chosen to apply and to compare different parametric, semiparametric, and nonparametric specifications of the regression function to measure the effect of ego-centered farm networks on farm productivity.

The results show that social networks have a consistent and significantly positive influence on farm productivity. In particular, large trade networks and dense household networks seem to augment farm productivity. Furthermore, our results suggest that when transaction costs are ignored, observed productivity measures as well as their ranking change significantly for the majority of the observations.

Our study confirms earlier findings that investigate the effect of social networks on various measures of firm productivity (e.g., Stam et al, 2013). However, in contrast to Uzzi (1996) and Yu and Chiu (2013), our results generally indicate a linear

relationship between density and productivity. In addition, our results indicate that, in some situations, observed productivity measures may be flawed if environmental factors such as transaction costs are ignored. In the presence of transaction costs and by applying standard production data, one may rather measure firm productivity instead of the productivity of the production process itself. The new aspect in our findings is that besides the firm's ability to steer the production process, the firm's ability to deal with the market it is operating in also significantly affects observed firm productivity. These findings suggest that one should interpret standard productivity measures in a broader context, which may eventually affect the design of productivity enhancing policies and aid projects, e.g., in developing or in transition countries.

However, some limitations are worth noting. There remains—as in many social network applications—the question of the extent to which the network structure is endogenously determined. In our particular case, productivity may have a positive effect on the size of the trade network, or, more likely, a third unconsidered factor may have a positive influence on both network size and productivity. For example, unobserved personal traits such as openness, ambition, or a positive attitude, may potentially influence both the number of trading partners and the productivity. However, our data set includes variables that should at least in part capture personal traits, e.g., risk perception or education, but prove to be statistically insignificant. To finally resolve the question of causality, a dynamic network analysis would be necessary where networks are sampled repeatedly over time together with the other production variables in order to derive clear information on the interplay between all measures. Unfortunately, this option was beyond our means.

Another aspect is the limited number of structural parameters that can be derived from ego-centred networks, and the bias towards strong ties induced by the name generator sampling technique (Lin, 1999). To overcome these problems, future studies should limit the geographical scope to be able to sample a full network—which might return more precise and more interesting results.

Of course, as a proxy, social networks can only inaccurately represent the true transaction costs, hence, our results can merely indicate the impact of transaction costs on farm productivity. In addition, social networks are only one of many factors that can influence transaction costs at the firm level. Other factors such as market structure or product characteristics may play an equally important or even greater role, and could be included for reasons of comparison in future analyses. Furthermore, our research design prevents us from distinguishing the effect of social networks on different sources of transaction costs, like information acquisition or trade activities. Consequently, future research should, in particular, address the quantifiability of different sources of transaction costs with the aim of identifying more precise and more accessible proxies for different sources of transaction costs. However, these suggestions would require data that are difficult and costly to collect.

Acknowledgements The authors are grateful to Jeff Racine, Martin Browning, Subal Kumbhakar, Chris Parmeter, and two anonymous referees for their valuable suggestions regarding the econometric analysis and for improving the paper. Funding was provided by the European Union's sixth framework programme within the project Advanced-Eval. Arne Henningsen is grateful to the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) for financially supporting this research. Of course, all errors are the sole responsibility of the authors.

References

Aitchison J, Aitken CGG (1976) Multivariate binary discrimination by the kernel method. Biometrika 63(3):413–420

Bandiera O, Rasul I (2006) Social networks and technology adoption in northern mozambique. The Ecoonomic Journal 116(514):869–902

Beckmann CM, Haunschild PR, Phillips DJ (2004) Friends or strangers? firm-specific uncertainty, market uncertainty, and network partner selection. Organisation Science 15(3):259–275

Bradley SW, McMullen JS, Artz K, Simiyu EM (2012) Capital is not enough: Innovation in developing economies. Journal of Management Studies 49(4):684–717

Burt RS (1984) Network items and the general social survey. Social Networks 6:293–339

Buskens V (1999) Social networks and trust. Dissertation, Utrecht University

den Butter FAG, Mosch RHJ (2003) Trade, trust and transaction cost. Working Paper TI 2003-082/3, Tinbergen Institute, Amsterdam, URL `http://dare.ubvu.vu.nl/bitstream/1871/9575/1/03082.pdf`

Castilla EJ, Hwang H, Granovetter E, Granovetter M (2000) Social networks in silicon valley. In: Lee CM, Miller WF, Hancock MG, Rowen HS (eds) The Silicon Valley Edge: A Habitat for Innovation and Entrepreneurship, Stanford University Press, Stanford, pp 218–247

Czekaj T, Henningsen A (2013) Panel data specifications in nonparametric kernel regression: An application to production functions. IFRO Working Paper 2013/5, Department of Food and Resource Economics, University of Copenhagen

Dekker DJ (2001) Effects of positions in knowledge networks on trust. Tech. Rep. TI 2001-062/1, Tinbergen Institute

Di Matteo T, Aste T, Gallegati M (2005) Innovation flow through social networks: Productivity distribution in france and italy. The European Physical Journal B 47:459–466

Fafchamps M (2001) The role of business networks in market development in Sub-Saharan Africa. In: Aoki M, Hayami Y (eds) Community and Market in Economic Development, Oxford University Press, pp 186–214

Hayfield T, Racine JS (2008) Nonparametric econometrics: The np package. Journal of Statistical Software 27(5):1–32

Henning CHCA, Zuckerman EW (2006) Boon and bane of social networking in markets with imperfect information: Theory and evidence from Polish and Slovakian rural credit markets, christian Albrechts University Kiel, Institute of Agricultural Economics

Henning CHCA, Henningsen G, Henningsen A (2012) Networks and transaction costs. American Journal of Agricultural Economics 94(2):377–385

Hsiao C, Li Q, Racine J (2007) A consistent model specification test with mixed discrete and continuous data. Journal of Econometrics 140(2):802–826

Hurvich CM, Simonoff JS, Tsai CL (1998) Smooting parameter selection in nonparametric regression using an improved Akaike information criterion. Journal of the Royal Statistical Society Series B 60:271–293

Jenssen JI, Koenig HF (2002) The effect of social networks on resource access and business start-ups. European Planning Studies 10(8):1039–1046

Lau CM, Bruton GD (2011) Strategic orientations and strategies of high technology ventures in two transition economies. Journal of World Business 46:371–380

Levi M (2000) When good defenses make good neighbors: A transaction cost approach on trust, the absence of trust and distrust. In: Ménard C (ed) Institutions, Contracts, and Organizations: Perspective from New Institutional Economics, Chichester, UK: Edward Elgar, chap 12, pp 137–157

Li Q, Racine JS (2004) Cross-validated local linear nonparametric regression. Statistica Sinica 14(2):485–512

Lin N (1999) Building a network theory of social capital. Connections 22(1):28–51

Luo Y (2003) Industrail dynamics and mangerial networking in an emerging market: The case of china. Strategic Management Journal 24:1315–1327

Ma S, Racine JS (2011) Inference for regression splines with categorical and continuous predictors, unpublished Working Paper, Department of Economics, McMaster University

Ma S, Racine JS (2012) Additive regression splines with irrelevant categorical and continuous regressors. Department of Economics Working Papers 2012-07, McMaster University, URL `http://ideas.repec.org/p/mcm/deptwp/2012-07.html`

Ma S, Racine JS, Yang L (2012) Spline regression in the presence of categorical predictors. Department of Economics Working Papers 2012-06, McMaster University, URL `http://ideas.repec.org/p/mcm/deptwp/2012-06.html`

Ménard C (2000) Enforcement procedures and governance structures: What relationship? In: Ménard C (ed) Institutions, Contracts and Organizations. Perspectives from New Institutional Economics, Cheltenham, Edward ldgar Pub., chap 17, pp 235–253

Nee V (1998) Norms and networks in economic and organizational performance. American Economic Review 88(2):85–89

Nie Z, Racine JS (2012) The crs package: Nonparametric regression splines for continuous and categorical predictors. The R Journal 4(2):48–56

Prajapati S, Biswas S (2011) Effect of entrepreneur network and entrepreneur self-efficacy on subject performance: a study of handicraft and handloom cluster. Journal of Entrepreneurship 20(2):227–247

R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`

Racine JS (1997) Consistent significance testing for nonparametric regression. Journal of Business and Economic Statistics 15:369–379

Racine JS, Li Q (2004) Nonparametric estimation of regression functions with both categorical and continuous data. Journal of Econometrics 119(1):99–130

Racine JS, Parmeter CF (2014) Data-driven model evaluation: A test for revealed performance. In: Racine JS, Su L, Ullah A (eds) The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics, Oxford Handbooks in Economics, Oxford University Press, pp 308–345

Racine JS, Hart J, Li Q (2006) Testing the significance of categorical predictor variables in nonparametric regression models. Econometric Reviews 25:523–544

Ramsey JB (1969) Tests for specification errors in classical linear least-squares regression analysis. Journal of the Royal Statistical Society Series B (Methodological) 31(2):350–371

Stam W, Arzlanian S, Elfring T (2013) Social capital of entrepreneurs and small firm performance: A meta-analyis of contextual and methodological moderators. Journal of Business Venture (in press)

Uzzi B (1996) The sources and consequences of embeddedness for the economic performance of organizations - the network effect. American Sociological Review 94(4):674–698

Wang MC, van Ryzin J (1981) A class of smooth estimators for discrete distributions. Biometrika 68:301–309

Williamson OE (2000) The new institutional economics: Taking stock, looking ahead. Journal of Economic Literature 38:595–613

Yu SH, Chiu WT (2013) Social networks and corporate performace: The moderating role of technical uncertainty. Journal of Managerial Issues 25(1):26–45