Received: 11 May 2022 | Revised: 8 October 2022 | Accepted: 30 November 2022

DOI: 10.1111/1477-9552.12522

Journal of Agricultural Economics

ORIGINAL ARTICLE

Analysing inefficiency in a non-parametric spatial-dynamic by-production framework: A *k*-nearest neighbour proposal

Ioannis Skevas[1] | **Alfons Oude Lansink**[2] | **Theodoros Skevas**[3]

[1]Department of International and European Studies, University of Macedonia, Thessaloniki, Greece

[2]Business Economics Group, Wageningen University, Wageningen, The Netherlands

[3]Division of Applied Social Sciences, University of Missouri, Missouri, USA

Correspondence

Ioannis Skevas, Department of International and European Studies, University of Macedonia, Thessaloniki, Greece.
Email: gskevas@uom.edu.gr

Abstract

This paper accounts for spatial effects by benchmarking farms against their *k*-nearest neighbours (KNN) and measuring their inefficiency in a non-parametric dynamic by-production setting. The optimal number of neighbours *k* against which farms are compared corresponds to the value of *k* that maximises the Moran I test for spatial autocorrelation of the good and the bad output of the farms' two sub-technologies. The inefficiency scores for farms' good output, variable inputs, investments and bad outputs are then computed and compared with those calculated based on a global technology, which benchmarks all farms together. The application focuses on an unbalanced panel of specialised Dutch dairy farms over the period 2009–2016 that contains information on their exact geographical locations. The results suggest that the inefficiency scores exhibit statistically significant differences between the KNN and the global model. Specifically, the inefficiencies are generally deflated when a KNN technology is considered, suggesting that ignoring spatial effects can overestimate inefficiency.

KEYWORDS

data envelopment analysis, Dutch dairy farms, dynamic by-production inefficiency, k-nearest neighbours

JEL CLASSIFICATION

C14, C21, D22, Q12

1 | INTRODUCTION

Close proximity between farms is crucial in developing networks that can help to reduce production costs and increase availability of inputs, as highlighted in the literature on agglomeration economies. Further neighbourhood benefits include knowledge spillovers (Skevas, 2020), technology diffusion (Case, 1992) and R&D spillovers (Mairesse & Mulkay, 2007). However, spatial concentration can also create tension between nearby farms through increased competition for labour or land (Weiss, 1999). In either case, given the important role that location plays in farms' production processes, farm-level benchmarking should account for spatial proximity of other farms.

In parametric efficiency studies, there are two main approaches to account for spatial proximity. The first approach is a spatial autoregressive (SAR) frontier model, which includes a spatial lag of the dependent variable on the right-hand side of the production/cost frontier (Glass et al., 2016), and the second approach directly includes a spatial lag of the inefficiency component (Skevas, 2020; Skevas & Skevas, 2021).[1] With regard to the non-parametric efficiency literature, studies account for spatial effects either in two-stage or in single-stage models. In the former, Skevas and Grashuis (2020), Schneider et al. (2021) and Skevas and Oude Lansink (2020) compute the inefficiency scores using data envelopment analysis (DEA), and then account for spatial spillovers in the traditional Simar and Wilson (2007) truncated bootstrap approach. However, a growing literature questions the validity of the assumption of separability, which is typically made in two-stage models between variables used for defining the frontier and variables that explain the distance to the frontier. For this reason, another stream of literature includes spatial aspects directly in the DEA model.

The direct incorporation of spatial factors in DEA models has its roots in the seminal papers of Cazals et al. (2002) and Daraio and Simar (2007) who introduced the order-m frontier that limits the analysis to a subset of m firms in order to mitigate the impact of outliers, and the conditional order-m efficiency approach that includes environmental factors that can influence firms' production processes, and therefore their inclusion in a certain subset m. Apart from individual characteristics, these environmental factors can include spatial indicators. However, as Vidoli and Canello (2016), and Fusco et al. (2020) argue, a selection of an erroneous or an incomplete set of spatial variables can significantly influence the inefficiency scores, for which ex-post validation processes are not available. To overcome this limitation, they propose using the location of each firm as the conditional (spatial) environmental variable, such that each firm is compared only against its neighbouring peers. This approach captures the global spatial trend, which is difficult to identify and/or measure by simply using a set of spatial variables. The immediate question is then about the optimal number of neighbours against which to benchmark. The answer depends heavily on the data at hand. Neighbouring peers can be the ones that belong to the same region or municipality if this information is available. However, given firms' exact location information (i.e., latitude and longitude), Vidoli and Canello (2016) propose the following procedure: (1) estimation of the optimal distance in terms of spatial autocorrelation;[2] (2) identification of neighbours falling within this distance; and (3) solution of the DEA problem, where individual firms are only compared against their optimal number of neighbours.

[1]Although not applied to efficiency analysis yet, an alternative approach called eigenvector spatial filtering (ESF) exists that incorporates spatial information in a model by computing spatial eigenvectors, which are defined by the spatial structure associated with a specific variable (Murakami & Griffith, 2019).

[2]Vidoli and Canello (2016) use the so-called semivariogram which studies the relationship between a random variable (i.e., inefficiency) and the location, so that it reveals how the variability of the random variable changes with increasing distance.

We propose an alternative optimal neighbour identification strategy that combines the widely used Moran I test for spatial autocorrelation and the *k*-nearest neighbour (KNN) concept. Using values of the number of neighbours, we construct several KNN matrices for different values of neighbours k, and test which value yields the maximum Moran I test statistic for the utilised variables. Subsequently, farms are benchmarked relative to their optimal KNN. As opposed to Vidoli and Canello (2016), our approach is less computationally demanding because the semiovariogram approach involves multiple estimation steps including (i) the estimation of the semiovariogram using a modified version of the General Additive Model, and (ii) the subsequent calculation of the optimal number of spatial peers based on the local maximum found from the previous step. We further depart from the existing literature by applying the proposed method to the dynamic by-production model building on the static model of Murty et al. (2012), and extended to the dynamic context by Dakpo and Oude Lansink (2019). Specifically, we derive farms' inefficiencies with respect to good output, variable inputs, investments and bad output once using a model that benchmarks farms only against their optimal KNN (i.e., KNN technology), and once using a model where farms are benchmarked against all farms in the sample (i.e., global technology). Subsequently, we use the non-parametric adapted Li test (Li, 1996; Simar & Zelenyuk, 2006) to compare the distributions of the inefficiencies' estimates across the global and the KNN technologies. By doing so, we provide information on whether dynamic inefficiency differs when accounting for spatial aspects.

The remainder of this paper is organised as follows. Section 2 presents the utilised dynamic by-production framework. Section 3 provides details on the estimation of the global and KNN dynamic by-production inefficiencies. Section 4 describes the data used in the application. Section 5 presents the results and discusses the performed robustness checks. Section 6 concludes.

2 | DYNAMIC BY-PRODUCTION TECHNOLOGY

This study considers a dynamic by-production technology, based on Murty et al. (2012), that consists of two sub-technologies: (1) a good output sub-technology $\Psi_g(t)$ that produces a vector of intended outputs $y(t)$, where $y \in \mathcal{R}_+^Q$; and (2) a bad output sub-technology $\Psi_b(t)$, that yields a vector of unintended outputs $b(t)$, where $b \in \mathcal{R}_+^R$. The term t represents time. We denote the vector of non-polluting fixed inputs as $f^{np}(t)$, with $f^{np} \in \mathcal{R}_+^S$, the vector of polluting fixed inputs as $f^p(t)$, with $f^p \in \mathcal{R}_+^W$, the vector of quasi-fixed inputs as $k(t)$, with $k \in \mathcal{R}_+^L$, the vector of gross investments as $i(t)$, with $i \in \mathcal{R}_+^L$, and the vector of variable inputs as $v(t)$, where $v \in \mathcal{R}_+^J$. The dynamic by-production technology $\Psi(t)$ is defined as the intersection of the good output dynamic sub-technology and the bad output dynamic sub-technology:

$$\Psi(t) = \Psi_g(t) \cap \Psi_b(t) \tag{1}$$

where $\Psi_g(t) = \Big\{(v(t), i(t), k(t), f^{np}(t), f^p(t), y(t), b(t)) : (v(t), i(t))$ can produce $y(t)$ given $k(t)$, $f^{np}(t)$ and $f^p(t)\Big\}$ and $\Psi_b(t) = \Big\{(v(t), i(t), k(t), f^{np}(t), f^p(t), y(t), b(t)) : (v(t), i(t))$ can yield $b(t)$ given $k(t)$, $f^{np}(t)$, and $f^p(t)\Big\}$. The properties of the good output dynamic sub-technology $\Psi_g(t)$ are described in Silva and Stefanou (2003):

g_1	No free lunch and inactivity
g_2	Input essentiality and attainability
g_3	Non-emptiness and closeness
g_4	Boundedness
g_5	Positive monotonicity in $v(t)$: if $v(t) \in \Psi_g(t)$ and $v'(t) \geq v(t)$ then $v'(t) \in \Psi_g(t)$
g_6	Negative monotonicity in $i(t)$: if $i(t) \in \Psi_g(t)$ and $i'(t) \leq i(t)$ then $i'(t) \in \Psi_g(t)$
g_7	Free disposability of good outputs: if $y(t) \in \Psi_g(t)$ and $y'(t) \leq y(t)$ then $y'(t) \in \Psi_g(t)$
g_8	Reverse nestedness in $k(t)$, $f^{np}(t)$, and $f^{p}(t)$: if $k(t) \in \Psi_g(t)$ and $k'(t) \geq k(t)$ then $k'(t) \in \Psi_g(t)$. If $f^{np}(t) \in \Psi_g(t)$ and $f^{np'}(t) \geq f^{np}(t)$ then $f^{np'}(t) \in \Psi_g(t)$. Similarly, if $f^{p}(t) \in \Psi_g(t)$ and $f^{p'}(t) \geq f^{p}(t)$ then $f^{p'}(t) \in \Psi_g(t)$.
g_9	Convexity in $(v(t), i(t), f^{np}(t), f^{p}(t), y(t))$

Given the properties presented above and assuming a flexible variable returns to scale (VRS) technology, the DEA problem for N firms for the good output sub-technology is:

$$\begin{gathered} \Psi_g(t) = (v(t), i(t), k(t), f^{np}(t), f^{p}(t), y(t), b(t)) \ : \\ y_o(t) \leq \sum_{n=1}^{N} \mu_n^g y_n(t), \\ v_o(t) \geq \sum_{n=1}^{N} \mu_n^g v_n(t), \\ f_o^{np}(t) \geq \sum_{n=1}^{N} \mu_n^g f_n^{np}(t), \\ f_o^{p}(t) \geq \sum_{n=1}^{N} \mu_n^g f_n^{p}(t), \\ i_o(t) - \delta k_o(t) \leq \sum_{n=1}^{N} \mu_n^g (i_n(t) - \delta k_n(t)), \\ \sum_{n=1}^{N} \mu_n^g = 1, \forall n, \\ (v(t), i(t), k(t), f^{np}(t), f^{p}(t), y(t), b(t)) \in \mathcal{R}_+^{J+L+L+S+W+Q+R} \end{gathered} \tag{2}$$

where δ stands for the depreciation rate with respect to the quasi-fixed input k, and $i_o(t) - \delta k_o(t)$ represents net investments (i.e., gross investments net of the depreciation rate). The inequality signs in the constraints of the good output sub-technology imply that all inputs increase good output. For the case of dairy farms, increasing variable inputs such as veterinary expenses and feed can improve cow's health/diet and raise milk production. Regarding fixed inputs, more animals and land imply higher production volume, whereas more labour can mean more milking hours, leading to increased milk production. However, investments in quasi-fixed assets decrease good output in the year of investment because they cause adjustment costs (financial and/or learning) that hinder production.

Moving to the bad output dynamic sub-technology $\Psi_b(t)$, its properties are opposite to those of the good output sub-technology, and as in Dakpo and Oude Lansink (2019), are expressed as:

b_1	Negative monotonicity in $v(t)$: if $v(t) \in \Psi_b(t)$ and $v'(t) \leq v(t)$ then $v'(t) \in \Psi_b(t)$
b_2	Positive monotonicity in $i(t)$: if $i(t) \in \Psi_b(t)$ and $i'(t) \geq i(t)$ then $i'(t) \in \Psi_b(t)$
b_3	Negative monotonicity in $k(t)$: if $k(t) \in \Psi_b(t)$ and $k'(t) \leq k(t)$ then $k'(t) \in \Psi_b(t)$
b_4	Negative monotonicity in $f^p(t)$: if $f^p(t) \in \Psi_b(t)$ and $f^{p'}(t) \leq f^p(t)$ then $f^{p'}(t) \in \Psi_b(t)$
b_5	Positive monotonicity in $f^{np}(t)$: if $f^{np}(t) \in \Psi_b(t)$ and $f^{np'}(t) \geq f^{np}(t)$ then $f^{np'}(t) \in \Psi_b(t)$
b_6	Positive monotonicity in $b(t)$: if $b(t) \in \Psi_b(t)$ and $b'(t) \geq b(t)$ then $b'(t) \in \Psi_b(t)$
b_7	Convexity in $(v(t), i(t), k(t), b(t))$
b_8	Polluting inputs essentiality $[v(t), k(t), f^p(t)]$
b_9	Boundedness

The above properties highlight the costly disposability of pollution as expressed in Murty et al. (2012), and one can consult Dakpo and Oude Lansink (2019) for a discussion of each property. Based on the above properties and assuming again a flexible VRS technology, the DEA problem for the bad output sub-technology is:

$$\begin{aligned}
\Psi_b(t) = (v(t), i(t), k(t), f^{np}(t), f^p(t), y(t), b(t)) \; : \\
b_o(t) \geq \sum_{n=1}^{N} \mu_n^b b_n(t), \\
f_o^{np}(t) \geq \sum_{n=1}^{N} \mu_n^g f_n^{np}(t), \\
f_o^p(t) \leq \sum_{n=1}^{N} \mu_n^g f_n^p(t), \\
v_o(t) \leq \sum_{n=1}^{N} \mu_n^b v_n(t), \\
i_o(t) - \delta k_o(t) \geq \sum_{n=1}^{N} \mu_n^b (i_n(t) - \delta k_n(t)), \\
\sum_{n=1}^{N} \mu_n^b = 1, \forall n, \\
(v(t), i(t), k(t), f^{np}(t), f^p(t), y(t), b(t)) \in \mathcal{R}_+^{J+L+L+S+W+Q+R}
\end{aligned} \tag{3}$$

The inequality signs in the constraints of the bad output sub-technology for non-polluting fixed inputs and investments in quasi-fixed assets suggest that they mitigate pollution. Notably, there is no consensus in the literature regarding the inclusion or not of such non-polluting inputs in the bad output sub-technology. However, in line with Dakpo and Oude Lansink (2019), we argue that although non-polluting inputs do not directly generate pollution, they can impact pollution generation indirectly. For example, and considering again the case of dairy farms, increasing fixed inputs such as labour can lead to better manure management and as a result reduce emissions. Furthermore, investments can mitigate pollution because there is a substitution between the associated adjustment costs and emissions (i.e., the resources used to learn a new technology will not be used to produce emissions).[3] The inequality signs in the constraints of variable inputs and polluting fixed inputs imply that they increase pollution because increasing such inputs is related to a larger herd size that can contribute to higher emissions due to higher manure production and rumen and intestinal fermentation.

[3]Note that in the properties of the bad output dynamic sub-technology, the disposability assumptions for investments and for quasi-fixed inputs have opposite signs. This is because a higher quantity for a quasi-fixed input (e.g., capital) can be related to a bigger herd size in the farm, which can in turn imply higher emissions.

Finally, following Murty et al. (2012) the DEA problem for the overall technology $\Psi(t)$ is:

$$
\begin{aligned}
\Psi(t) = (v(t), i(t), k(t), f^{np}(t), f^{p}(t), y(t), b(t)) \; : \\
y_o(t) \leq \sum_{n=1}^{N} \mu_n^g y_n(t), \\
v_o(t) \geq \sum_{n=1}^{N} \mu_n^g v_n(t), \\
f_o^{np}(t) \geq \sum_{n=1}^{N} \mu_n^g f_n^{np}(t), \\
f_o^{p}(t) \geq \sum_{n=1}^{N} \mu_n^g f_n^{p}(t), \\
i_o(t) - \delta k_o(t) \leq \sum_{n=1}^{N} \mu_n^g (i_n(t) - \delta k_n(t)), \\
\sum_{n=1}^{N} \mu_n^g = 1 \\
b_o(t) \geq \sum_{n=1}^{N} \mu_n^b b_n(t), \\
f_o^{np}(t) \geq \sum_{n=1}^{N} \mu_n^g f_n^{np}(t), \\
f_o^{p}(t) \leq \sum_{n=1}^{N} \mu_n^g f_n^{p}(t), \\
v_o(t) \leq \sum_{n=1}^{N} \mu_n^b v_n(t), \\
i_o(t) - \delta k_o(t) \geq \sum_{n=1}^{N} \mu_n^b (i_n(t) - \delta k_n(t)), \\
\sum_{n=1}^{N} \mu_n^b = 1, \forall n, \\
(v(t), i(t), k(t), f^{np}(t), f^{p}(t), y(t), b(t)) \in \mathcal{R}_+^{J+L+L+S+W+Q+R}
\end{aligned} \tag{4}
$$

In order to account for interdependence in the two sub-technologies $\Psi_g(t)$ and $\Psi_b(t)$, we follow Dakpo and Oude Lansink (2019) and further impose the following constraints which equalise the optimal values of the common variables of the different sub-technologies (i.e., variable inputs, fixed inputs and investments):

$$
\begin{aligned}
\sum_{n=1}^{N} \mu_n^g v_n(t) = \sum_{n=1}^{N} \mu_n^b v_n(t) \\
\sum_{n=1}^{N} \mu_n^g f_n^{np}(t) = \sum_{n=1}^{N} \mu_n^b f_n^{np}(t) \\
\sum_{n=1}^{N} \mu_n^g f_n^{p}(t) = \sum_{n=1}^{N} \mu_n^b f_n^{p}(t) \\
\sum_{n=1}^{N} \mu_n^g (i_n(t) - \delta k_n(t)) = \sum_{n=1}^{N} \mu_n^b (i_n(t) - \delta k_n(t))
\end{aligned} \tag{5}
$$

The interdependence constraints ensure that projections towards the different frontiers reach consistent benchmarks across the different sub-technologies (Dakpo & Oude Lansink, 2019).

3 | ESTIMATION OF GLOBAL VERSUS KNN DYNAMIC BY-PRODUCTION INEFFICIENCY

3.1 | Global dynamic by-production inefficiency

Measuring technical inefficiency in the dynamic by-production framework presented in the previous section can be achieved using several approaches, such as the radial and the hyperbolic distance functions. However, these two approaches may not be suitable for the specified dynamic by-production model because radial distance functions cannot handle zero values for one or more variables, which is typically the case for gross investments in dynamic models, while the hyperbolic distance function makes the unrealistic assumption that good and bad outputs should be expanded and contracted, respectively, by the same proportion. The above restrictions, though, are not in place when using a directional distance function (Chambers et al., 1998). Hence, in this study we consider the two sub-technologies presented in the previous section and use a general formulation of the non-radial form of the directional distance function to measure inefficiency, as in Zhang and Choi (2014). Specifically, the global directional distance function for the overall technology is written as:

$$
\begin{aligned}
&\overrightarrow{D}_t^{global}\left(v(t), i(t), k(t), f^{np}(t), f^{p}(t), y(t), b(t); \overrightarrow{g_y}, \overrightarrow{g_v}, \overrightarrow{g_i}, \overrightarrow{g_b}\right) \\
&= \max_{\beta,\mu^g,\mu^b} \frac{1}{N_{\vec{g}}}\left[\beta_y + \beta_v + \beta_i + \beta_b\right] \\
&\text{s.t.} \\
&y_o(t) + \beta_y \overrightarrow{g_y} \le \sum_{n=1}^{N} \mu_n^g y_n(t) \\
&v_o(t) - \beta_v \overrightarrow{g_v} \ge \sum_{n=1}^{N} \mu_n^g v_n(t) \\
&f_o^{np}(t) \ge \sum_{n=1}^{N} \mu_n^g f_n^{np}(t) \\
&f_o^{p}(t) \ge \sum_{n=1}^{N} \mu_n^g f_n^{p}(t) \\
&i_o(t) - \delta k_o(t) + \beta_i \overrightarrow{g_i} \le \sum_{n=1}^{N} \mu_n^g (i_n(t) - \delta k_n(t)) \\
&b_o(t) - \beta_b \overrightarrow{g_b} \ge \sum_{n=1}^{N} \mu_n^b b_n(t) \\
&v_o(t) - \beta_v \overrightarrow{g_v} \le \sum_{n=1}^{N} \mu_n^b v_n(t) \\
&f_o^{np}(t) \ge \sum_{n=1}^{N} \mu_n^g f_n^{np}(t) \\
&f_o^{p}(t) \le \sum_{n=1}^{N} \mu_n^g f_n^{p}(t) \\
&i_o(t) - \delta k_o(t) + \beta_i \overrightarrow{g_i} \ge \sum_{n=1}^{N} \mu_n^b (i_n(t) - \delta k_n(t)) \\
&\sum_{n=1}^{N} \mu_n^g v_n(t) = \sum_{n=1}^{N} \mu_n^b v_n(t) \\
&\sum_{n=1}^{N} \mu_n^g f_n^{np}(t) = \sum_{n=1}^{N} \mu_n^b f_n^{np}(t) \\
&\sum_{n=1}^{N} \mu_n^g f_n^{p}(t) = \sum_{n=1}^{N} \mu_n^b f_n^{p}(t) \\
&\sum_{n=1}^{N} \mu_n^g (i_n(t) - \delta k_n(t)) = \sum_{n=1}^{N} \mu_n^b (i_n(t) - \delta k_n(t)) \\
&\sum_{n=1}^{N} \mu_n^g = 1 \\
&\sum_{n=1}^{N} \mu_n^b = 1
\end{aligned}
\tag{6}
$$

where $N_{\overleftarrow{g}}$ stands for the number of decision variables in the problem's objective function, implying that we give the same weight to each inefficiency score. Generalising the non-radial approach, each of the β inefficiencies are individualised with the different corresponding variables β_{yq}, β_{vj}, β_{il} and β_{br}, (i.e., good output, variable inputs,[4] investments and bad output inefficiencies). As its name suggests, in the global directional distance function each firm is compared against all other firms. Following Chung et al. (1997), the directional vectors for the good output, variable inputs and bad output are set equal to the corresponding observed values. The directional vector for investments equals 20% of capital, as in Dakpo and Oude Lansink (2019) and Skevas and Oude Lansink (2020). Additionally, the DEA problem is solved separately for each year recognising that technology can differ across different time periods. Finally, the interdependence constraints that appear in Equation (6) and equalise the optimal values for variable inputs, fixed inputs and gross investments for the two sub-technologies, make sure that these two sub-technologies are imbedded in the same Euclidean sub-space.

3.2 | Identification of optimal number of neighbours k

Benchmarking firms in a KNN framework requires the identification of the optimal number of nearest neighbours k. Since in DEA the best-practice frontier is formed based on the available data (and not based on functional forms that shape their relationship, and noise), the optimal number of neighbours is identified based on the observed spatial autocorrelations in the specified variables. The Moran I test is used to test for spatial autocorrelations. We apply the Moran I test to the two outputs of these two technologies, tha is, the good and the bad output. Besides, given that the inputs of production are transformed into outputs, as production theory suggests, any spatial autocorrelations in the inputs should eventually be reflected in the outputs produced. Note, however, that we also apply the test to all remaining variables in the study's robustness checks. In a general form, the Moran I test measures the spatial autocorrelation of the variable of interest based on the following formula:

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(x_i - \overline{x}\right)\left(x_j - \overline{x}\right)}{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2} \tag{7}$$

where N is the number of firms indexed by i and j, x is the variable of interest with a bar over it representing its average value, and w_{ij} is the spatial weight between observations i and j. Typically, in the spatial econometrics literature w_{ij} is presented in its matrix form $\mathbf{W}$ of $N \times N$ dimension, which in our case is constructed based on the KNN concept. Specifically, a weight of 1 is assigned to each firm's KNN and a weight of 0 otherwise. We construct several $\mathbf{W}$ matrices for all possible values of k and test which value gives the maximum (significant) Moran I test statistic. The Moran I test takes values between -1 and $+1$ and its statistical significance can be evaluated based on a z-score and its associated p-value.[5]

[4]We measure inefficiency separately for each variable input because this is more informative than simply presenting an aggregate inefficiency score for all variable inputs.

[5]Note in passing that we do not row-normalise the spatial weights matrix $\mathbf{W}$, as this is mainly needed when estimating a parametric spatial model and not when conducting the Moran I test. Besides, in the subsequent application the outcomes of the Moran I test are the same irrespective of whether or not we row-normalise the spatial weights matrix.

3.3 | KNN dynamic by-production inefficiency

Using the optimal value k from the above step, the KNN dynamic by-production model limits the inefficiency analysis to a subset of the optimal k firms, with each firm being compared with its KNN. Hence, the KNN directional distance function for the overall technology reads as:

$$
\begin{aligned}
&\overrightarrow{D}_t^{KNN}\left(v(t), i(t), k(t), f^{np}(t), f^{p}(t), y(t), b(t); \overrightarrow{g_y}, \overrightarrow{g_v}, \overrightarrow{g_i}, \overrightarrow{g_b}\right) \\
&= \max_{\beta,\mu^g,\mu^b} \frac{1}{N_{\vec{g}}}\left[\beta_y + \beta_v + \beta_i + \beta_b\right] \\
&\text{s.t.} \\
&y_o(t) + \beta_y \overrightarrow{g_y} \le \textstyle\sum_{k=1}^{K} \mu_k^g y_k(t) \\
&v_o(t) - \beta_v \overrightarrow{g_v} \ge \textstyle\sum_{k=1}^{K} \mu_k^g v_k(t) \\
&f_o^{np}(t) \ge \textstyle\sum_{k=1}^{K} \mu_k^g f_k^{np}(t) \\
&f_o^{p}(t) \ge \sum_{k=1}^{K} \mu_k^g f_k^{p}(t) \\
&i_o(t) - \delta k_o(t) + \beta_i \overrightarrow{g_i} \le \sum_{k=1}^{K} \mu_k^g (i_k(t) - \delta k_k(t)) \\
&b_o(t) - \beta_b \overrightarrow{g_b} \ge \sum_{k=1}^{K} \mu_k^b b_k(t) \\
&v_o(t) - \beta_v \overrightarrow{g_v} \le \sum_{k=1}^{K} \mu_k^b v_k(t) \\
&f_o^{np}(t) \ge \sum_{k=1}^{K} \mu_k^b f_k^{np}(t) \\
&f_o^{p}(t) \le \sum_{k=1}^{K} \mu_k^b f_k^{p}(t) \\
&i_o(t) - \delta k_o(t) + \beta_i \overrightarrow{g_i} \ge \sum_{k=1}^{K} \mu_k^b (i_k(t) - \delta k_k(t)) \\
&\sum_{k=1}^{K} \mu_k^g v_k(t) = \sum_{k=1}^{K} \mu_k^b v_k(t) \\
&\sum_{k=1}^{K} \mu_k^g f_k^{np}(t) = \sum_{k=1}^{K} \mu_k^b f_k^{np}(t) \\
&\sum_{k=1}^{K} \mu_k^g f_k^{p}(t) = \sum_{k=1}^{K} \mu_k^b f_k^{p}(t) \\
&\sum_{k=1}^{K} \mu_k^g (i_k(t) - \delta k_k(t)) = \sum_{k=1}^{K} \mu_k^b (i_k(t) - \delta k_k(t)) \\
&\sum_{k=1}^{K} \mu_k^g = 1 \\
&\sum_{k=1}^{K} \mu_k^b = 1
\end{aligned}
\tag{8}
$$

where $k = 1, \ldots, K$ represents the subset of the firm's KNN. The same directional vectors are used as in the case of the global technology, whereas the DEA problem is solved separately for each year. This is done because the technology as well as the optimal k can differ across years.

3.4 | Comparison of inefficiency densities

We use the non-parametric adapted Li test (originally developed by Li, 1996 and subsequently extended to the DEA framework by Simar & Zelenyuk, 2006), to compare the densities of the inefficiencies' estimates under the global and the KNN technologies. The test recognises the bounded support of the inefficiency estimates and uses the estimated inefficiency scores instead of the true ones.[6] Assuming that u_g and u_k are two vectors of random variables, the test's null and alternative hypothesis are:

$$\begin{aligned} H_0 &: f_{g(u_g)} = f_{k(u_k)} \\ H_1 &: f_{g(u_g)} \neq f_{k(u_k)} \quad \text{for a set of positive measures} \end{aligned} \tag{9}$$

where f denotes the densities of the random variables. Given that in this study the estimated inefficiency scores are bounded from below at 0, we follow Simar and Zelenyuk (2006) and use a smoothing procedure:

$$\hat{\beta}_n^* = \begin{cases} \hat{\beta}_n + \epsilon_n, & \text{if} \quad \hat{\beta}_n = 0 \\ \hat{\beta}_n & \text{otherwise} \end{cases} \tag{10}$$

with $\epsilon_n \sim \text{Uniform}\left(0, \min\left\{N^{-2/D}, a\right\}\right)$, where N stands for the number of observations, D is the dimension of the convergence rate of the DEA model (i.e., the total number of specified variables), and α is the 0.05 quantile of the empirical distribution of $\hat{\beta}_n > 0$. Li (1996) shows that the asymptotic distribution of the test-statistic is standard normal. We use the algorithm presented in Simar and Zelenyuk (2006) to obtain the value of the test-statistic and the corresponding p-values.

4 | APPLICATION

The data used in the application are obtained from the Dutch Farm Accountancy Data Network (FADN) collected by Wageningen Economic Research in the Netherlands. The dataset contains information on specialised Dutch dairy farms observed over the period 2009–2016, and is an unbalanced panel of 2103 observations. Based on the FADN definition, specialised dairy farms are those whose revenues from sales of milk, milk products and turnover and growth of cattle comprise at least 66% of their total revenues.

The model distinguishes one good output, three fixed inputs (two non-polluting and one polluting), one quasi-fixed input with its associated investments, two variable inputs and one bad output. The good output is farms' total output and includes milk, milk products, turnover and growth of cattle, crop and other products. The three fixed inputs are total land measured in hectares, total labour expressed in hours, and animals measured in livestock units. Total labour is considered as a fixed input because it mostly concerns family labour. Animals are also specified as a fixed input because changing herd size can involve adjusting a large part of capital (i.e., milking machines, animal housing etc.). Total land and total labour are treated as non-polluting fixed inputs, whereas animals are considered as a polluting input as they are responsible for methane (CH_4) emissions. The quasi-fixed input is capital stock of buildings and machinery, and gross investments in this component are also considered. These gross investments are measured as the end value of capital stock in year t minus the beginning value of capital stock in the same year (which is essentially the end value of capital stock in year $t-1$), plus the value of depreciation in year t. Note also that this is an aggregate variable

[6]Due to the inherent bias of the estimated inefficiency scores, other non-parametric tests such as the Kolmogorov–Smirnov or the Mann–Whitney are avoided because they have an incorrect size and, as a result, wrong p-values (Kenjegalieva et al., 2009; Ohene-Asare et al., 2017; Simar & Wilson, 2002).

consisting of gross investments in buildings, and gross investments in machinery. Although for one of the sub-components a negative value is observed for some observations, the aggregate variable is for all observations non-negative. Hence, we do not observe any disinvestments in the sample. In any case, these are still accommodated through the use of the directional vector that is positive for all observations (20% in our case), which allows for each farm to be projected on the part of the frontier with positive investments.

The two variable inputs are costs of intermediate inputs (excluding purchased feed) and purchased feed costs. Intermediate input costs consist of veterinary expenses, crop-specific costs, energy expenses, contract work and other variable costs. The bad output is farms' total CH_4 emissions that come from (1) cows' manure production and storage (amount × emission factor), and (2) from their rumen and intestinal fermentation (the ration, the ration composition and the emission factor). These CH_4 emissions are measured in tons of carbon dioxide (CO_2) equivalent. Specifically, the factor with which CH_4 emissions are multiplied in order to yield the corresponding CO_2 equivalent is 28, as laid down in the most recently published standard of the Intergovernmental Panel on Climate Change. All the above monetary variables are measured in 2010 constant prices using Eurostat price indices, thus creating implicit quantities. The Eurostat price indices, which are the same for all farmers, are used because our dataset does not contain information on farm-specific prices. However, we do not expect that prices vary across space because farmers are price-takers and the Netherlands is a relatively small country. Summary statistics of all variables appear in Table 1.[7]

Finally, our dataset provides information on farms' latitudes and longitudes. These are used to calculate the distance of each farm to all others, and based on them, construct the different spatial weight matrices **W** as described in Section 3.2. Note that we allow the optimal number of neighbours to change over time due to both technical and theoretical reasons. From a technical perspective, the optimal number of neighbours *k* changes over time because the employed dataset is unbalanced, with farms rotating in and out of the sample over the years. From a theoretical perspective, the optimal number of neighbours changes because of farm exit and entry, changes in farmers' social networks over time due to their changing preferences, priorities, targets or friendships, the establishment of new study groups that can change the spatial interactions, or because farmers retire and new generations with different social relationships take over.

TABLE 1 Summary statistics of utilised variables

Variable	Mean	Std. dev.	5%	95%
Good output (€ in 2010)	336,423.400	239,921.600	78,653.460	818,078.200
Land (hectares)	68.227	40.939	22.156	145.448
Labour (hours)	4645.135	2595.074	2101.200	8492.000
Animals (livestock units)	235.346	434.449	56.700	521.700
Capital (€ in 2010)	484,042.600	411,158.100	62,033.380	1,399,308.000
Gross investments (€ in 2010)	117,133.500	172,278.400	7447.563	419,612.600
Intermediate inputs (€ in 2010)	76,817.560	58,383.620	18,226.380	182,489.800
Purchased feed (€ in 2010)	79,513.310	65,754.830	14,114.630	207,650.300
Bad output (CH_4 emissions in tons of CO_2-eq)	664.686	450.840	174.745	1562.514

[7]Note that Table 1 only presents the lowest and highest 5% values of the utilised variables, as reporting minimum and maximum values is not allowed by the data provider.

5 | RESULTS

5.1 | Main empirical findings

Table 2 presents the optimal values of k (i.e., the ones that yield the maximum significant Moran I test statistic) for good and bad output for each year. The optimal k values are the same for both variables in each year, highlighting a consistency in terms of spatial autocorrelation across the outputs of the two sub-technologies. Additionally, we observe that the optimal k values differ only slightly across years ranging from 14 to 19. Based on these results, the KNN DEA model is computed separately for each time period and with the corresponding number of optimal k. Obviously, the model is solved with a small number of observations in each year. Although this fact implies that the model can suffer from the 'curse of dimensionality', there does not exist a certain rule of thumb for the minimum number of observations in DEA, particularly for the dynamic by-production model. Certainly, it would be preferable to have a bigger dataset as it would probably yield a larger number of optimal neighbours. However, this is not the case with the available FADN data. Note as well that it is difficult to ameliorate the 'curse of dimensionality' by either changing the number of neighbouring peers k or by removing some variables. This is because the optimal value of k is calculated based on the Moran I test, and all the specified variables are already aggregated as much as possible. Also, the small number of observations is due to the fact that the DEA problem is solved separately for each year recognising that technology can differ across different time periods. Hence, there exists a trade-off between 'curse of dimensionality' and model specification.

Tables 3 and 4 present the inefficiency estimates across time periods for the global and the KNN technology, respectively. The inefficiencies with respect to good output, investments and bad output are lower in the KNN technology. Specifically, farms can increase their good output by 1.7% under the global technology and by 1.1% under the KNN technology, on average. Investments can on average be expanded by 99.52% and 13.38% (inefficiency $\times 0.2 \times 100$) based on the global and KNN technology, respectively. Bad output can be contracted by 19.3% according to the global technology and by 1.9% based on the KNN technology, on average. Regarding variable inputs, slightly lower inefficiency scores are observed under the global technology. Intermediate inputs can on average be reduced by 5.9% and 15.6% under the global and the KNN technology, respectively. Regarding purchased feed, a 5.4% average decrease in the global technology and a 11.7% average decrease in the KNN technology is possible.

The above differences in the inefficiency estimates are visualised through Figure 1, which presents the time evolution of all inefficiencies (averaged across farms) for the global and the KNN technolo-

TABLE 2 Optimal values of k for good and bad output across years

Variable	2009	2010	2011	2012	2013	2014	2015	2016
Good output	15	18	15	14	19	15	15	16
Bad output	15	18	15	14	19	15	15	16

TABLE 3 Global inefficiency estimates across time periods

Inefficiencies	2009	2010	2011	2012	2013	2014	2015	2016	Average
Good output	0.015	0.005	0.014	0.012	0.017	0.013	0.039	0.017	0.017
Intermediate	0.079	0.041	0.054	0.045	0.088	0.079	0.053	0.033	0.059
Feed	0.081	0.016	0.040	0.066	0.094	0.028	0.055	0.053	0.054
Investments	7.439	5.822	4.530	3.648	4.324	4.835	5.166	4.042	4.976
Bad output	0.178	0.285	0.181	0.170	0.185	0.181	0.125	0.235	0.193
	$N=227$	$N=241$	$N=275$	$N=277$	$N=281$	$N=275$	$N=267$	$N=260$	

JAE Journal of Agricultural Economics

TABLE 4 KNN inefficiency estimates across time periods

Inefficiencies	2009	2010	2011	2012	2013	2014	2015	2016	Average
Good output	0.011	0.010	0.005	0.010	0.009	0.008	0.017	0.021	0.011
Intermediate	0.167	0.168	0.158	0.148	0.138	0.160	0.154	0.158	0.156
Feed	0.141	0.123	0.108	0.108	0.104	0.109	0.127	0.116	0.117
Investments	1.153	0.925	0.514	0.413	0.536	0.559	0.690	0.559	0.669
Bad output	0.020	0.033	0.021	0.017	0.019	0.014	0.013	0.018	0.019
	$K = 15$	$K = 18$	$K = 15$	$K = 14$	$K = 19$	$K = 15$	$K = 15$	$K = 16$	

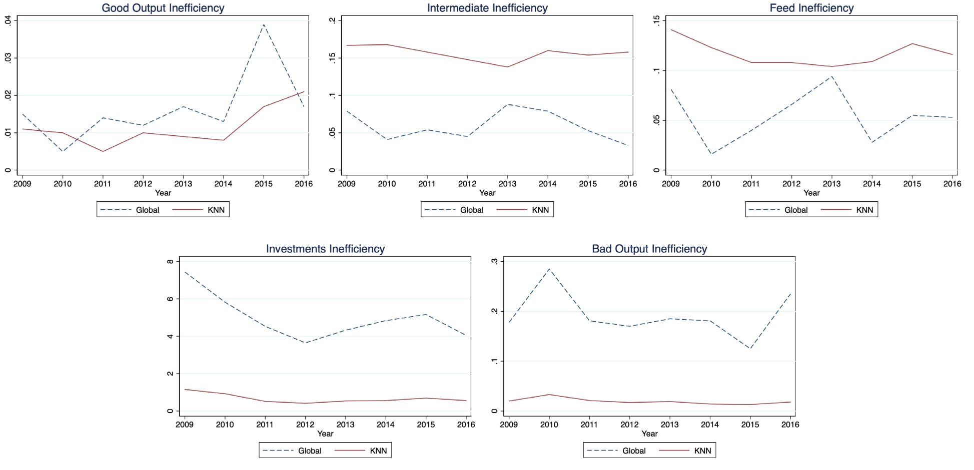

FIGURE 1 Time evolution of global and KNN inefficiencies

gies. Furthermore, Figures A1–A8 present kernel densities of all estimated inefficiencies for all years for the global and the KNN technologies.

In general, the reported technical inefficiency estimates are consistent with the findings of other related studies. Very low output inefficiency scores in the global technology for Dutch dairy farms are also reported by Skevas et al. (2018) and Skevas (2020). With respect to emissions, Dakpo and Oude Lansink (2019) report that GHG emissions from French cows can be reduced by an average of 7.1%, which is closer to our finding from the KNN technology. Dakpo and Oude Lansink (2019) also find high investment inefficiency, as the present study does. This is because in dairy farming we typically observe producers who invest very small amounts and others that invest a lot, with this high heterogeneity explaining the corresponding high investments inefficiency. Regarding variable inputs' inefficiencies, the estimates from the KNN technology are closer to the findings of other studies as the one of Skevas and Oude Lansink (2020) who report an average inefficiency of 22% for Dutch dairy farms' variable inputs.

Moving to the bigger picture, the finding that lower inefficiency scores are observed for good output, investments and bad output under the KNN technology is an expected result also reported by Vidoli and Canello (2016) and Fusco et al. (2020). This is because in a KNN setting, firms are only compared relative to their neighbouring peers given that their close proximity can result in employing similar production practices/technologies. This leads to a reduction of the heterogeneity in the sample that prevents an overestimation of inefficiency that is observed when firms are benchmarked under a global technology. This is particularly true for the inefficiency with respect to investments. This heterogeneity may not only be due to different farm networks that arise across space but also due to climatic differences (the growing season differs by 3 weeks between farms in the north and the south

TABLE 5 Li-test for each inefficiency component

Variable	Mean global inefficiency	Mean KNN inefficiency	Li test	*p*-value
Good output	0.017	0.011	11.212	0.000
Intermediate	0.059	0.156	219.624	0.000
Feed	0.054	0.117	183.334	0.000
Investments	4.976	0.669	329.664	0.000
Bad output	0.193	0.019	449.525	0.000

TABLE 6 Optimal values of k for all remaining variables for each year

Variable	2009	2010	2011	2012	2013	2014	2015	2016
Land	15	15	17	16	19	15	15	16
Labour	16	18	17	17	NS	NS	NS	NS
Animals	NS	NS	NS	14	19	15	14	15
Capital	17	15	NS	17	18	15	15	14
Intermediate	16	14	15	14	15	14	14	16
Feed	15	15	15	14	16	14	15	15
Investments	16	NS	NS	18	NS	NS	14	NS
Minimum	**15**	**14**	**15**	**14**	**15**	**14**	**14**	**14**
Maximum	**17**	**18**	**17**	**18**	**19**	**15**	**15**	**16**

Note: NS means that the Moran I test is not significant for all k.

TABLE 7 Robustness checks estimates of inefficiency scores under a KNN technology

Variable	2009	2010	2011	2012	2013	2014	2015	2016	Average
Good output	0.013	0.010	0.007	0.011	0.006	0.008	0.015	0.022	0.012
Intermediate	0.160	0.177	0.149	0.145	0.143	0.162	0.152	0.168	0.157
Feed	0.134	0.122	0.098	0.110	0.105	0.112	0.128	0.122	0.116
Investments	1.246	0.773	0.620	0.447	0.408	0.564	0.674	0.489	0.653
Bad output	0.022	0.019	0.023	0.020	0.013	0.013	0.013	0.016	0.017
	$K = 17$	$K = 14$	$K = 17$	$K = 18$	$K = 15$	$K = 14$	$K = 14$	$K = 14$	

of the Netherlands) or differences in soil conditions as dairy farms can be found on sandy, peat, clay and loss soils (van den Berg et al., 2017). However, inefficiencies with respect to variable inputs exhibit lower values when a global technology is considered. This occurs because the method removes all slacks from the model. Hence, the optimal solution may entail a larger contraction of one of the inputs in a sub-sample compared to the entire sample.

Finally, Table 5 presents the values of the Li test and the corresponding p-values for comparing the densities of the different inefficiency components across the global and the KNN technologies. The null hypothesis of equal densities is rejected for all inefficiencies, manifesting statistically significant differences between the global and the KNN inefficiencies. This result suggests that ignoring spatial effects across farms significantly changes the inefficiency scores.

5.2 | Robustness checks

We perform robustness checks with respect to different values of k in order to identify the sensitivity of the results from the KNN technology. For this purpose, we first compute the optimal k values for all the remaining variables and report their range (minimum and maximum values) for each time period. The results are presented in Table 6.

We observe that for some variables, the Moran I test is insignificant for certain years for all k. However, in most cases the Moran I test is significant and reaches its maximum for k between 14 and 19. This result is very close to the optimal k values reported for the good and the bad outputs.

Then, we compute for each year the inefficiency scores under a KNN technology using the extremes of optimal k's (i.e., minimum and maximum) for all involved variables (including the good and the bad output). However, given that for all years the minimum or the maximum optimal k from Table 6 coincides with that of our base KNN model from Table 2 (for instance in 2009 the minimum k from Table 6 coincides with the optimal k presented in Table 2), the minimum or maximum k from Table 6 is only used dependent on the case. In this way, the inefficiencies under a KNN technology are computed based on the lowest and largest value of the range of optimal k's of all involved variables. The results from the robustness checks are presented in Table 7. The inefficiency scores for all variables are quite similar to those from the base KNN model presented in Table 4. This finding suggests that inefficiencies are not sensitive across the range of optimal k values for all utilised variables.[8] This result is even more clear when looking at Table A1, which presents the inefficiency scores for the minimum and the maximum values of k for each year.

6 | CONCLUSIONS

This study proposes an optimal neighbour identification strategy that combines the Moran I test for spatial autocorrelation and the k-nearest neighbour (KNN) approach to benchmark farms against their optimal KNN and measure their inefficiencies in a non-parametric dynamic by-production setting. The optimal number of KNN corresponds to the number of neighbours that maximise the Moran I test statistic for the outputs of farms' two sub-technologies (good and bad output). The results from the KNN model are compared with those from a global model that benchmarks all farms together. The models are applied to specialised dairy farms from the Netherlands and measure their inefficiencies in terms of their good output (milk and other products) production, variable inputs (costs of intermediate inputs and purchased feed) use, investments levels and bad output (methane emissions) production as in Dakpo and Oude Lansink (2019).

According to the Moran I test, the optimal numbers of neighbours (i.e., those that maximise the Moran I test) range from 14 to 19 dependent on the year under consideration. The results from DEA suggest that technical inefficiency is generally deflated when using a KNN approach. This holds for good output, investments and bad output inefficiencies. The benefits from using a KNN model are especially highlighted in the case of investments inefficiency, which is severely inflated when considering a global technology. This is because farms are compared against very heterogeneous ones, which is less pronounced in the KNN model that compares farms only against their neighbouring peers. However, inefficiencies with respect to variable inputs are slightly higher when using a KNN technology, because the method removes all slacks from the model, and therefore the optimal solution results in a larger contraction for the variable inputs in the KNN sub-sample. Robustness checks with respect to different KNN reveal that the results are not sensitive across the range of optimal KNN for all utilised variables. Additionally, the adapted Li-test (Simar & Zelenyuk, 2006) suggests that the densities of all inefficiency components are significantly different among the global and the KNN models, highlighting that practitioners should be cautious regarding the peers against which farms are benchmarked, in order to provide more accurate efficiency estimates that can be used by both farmers and policy-makers in their efforts to improve efficiency.

[8]Furthermore, the adapted Li test suggests that the inefficiency scores from the robustness checks of the KNN technology are again statistically different from the ones derived from the global technology.

An interesting avenue for future research would be to decompose inefficiency changes over the years from the KNN model into inefficiency changes due to changes in the number of optimal neighbours k across years, frontier shifts, and farm-specific inefficiency changes. Also, the proposed approach can be adapted to the estimation of total factor productivity growth and its components. Given the significant differences of the inefficiency scores among the global and the KNN technology reported in the current study, an analogous result can be expected for productivity. Furthermore, the KNN benchmarking approach can be particularly useful when analysing the inefficiencies/productivities of farms located in less developed countries that are characterised by high degree of farm heterogeneity where a global benchmarking can severely distort the performance estimates. Additionally, the Moran I test that is used to derive the optimal number of neighbours k is conducted separately for the good and the bad outputs. We argue that this approach is not so restrictive as the good and the bad outputs are already correlated by definition, given that the latter is a by-product of the former. This inherent correlation is most probably responsible for the fact that the optimal number of neighbours k found from the Moran I test are the same for both variables, irrespective of the year considered. However, future research could focus on deriving the optimal number of neighbours k, while taking into account the inherent correlation between the good and the bad outputs when conducting the Moran I test. Finally, in cases where farms are unevenly distributed across space, which is more probable in large countries rather than in small countries like the Netherlands where farms are spread across space, one can relax the assumption of a homogeneous spatial pattern and allow the optimal number of neighbours to vary across farms.

ACKNOWLEDGEMENTS

We would like to thank two anonymous reviewers for constructive comments on an earlier draft. The publication of the article in OA mode was financially supported by HEAL-Link.

ORCID

Ioannis Skevas https://orcid.org/0000-0003-4595-4396

REFERENCES

Case, A. (1992) Neighborhood influence and technological change. *Regional Science and Urban Economics*, 22(3), 491–508.

Cazals, C., Florens, J. & Simar, L. (2002) Nonparametric frontier estimation: a robust approach. *Journal of Econometrics*, 106(1), 1–25.

Chambers, R.G., Chung, Y. & Färe, R. (1998) Profit, directional distance functions, and Nerlovian efficiency. *Journal of Optimization Theory and Applications*, 98(2), 351–364.

Chung, Y.H., Färe, R. & Grosskopf, S. (1997) Productivity and undesirable outputs: a directional distance function approach. *Journal of Environmental Management*, 51(3), 229–240.

Dakpo, K.H. & Oude Lansink, A. (2019) Dynamic pollution-adjusted inefficiency under the by-production of bad outputs. *European Journal of Operational Research*, 276(1), 202–211.

Daraio, C. & Simar, L. (2007) Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach. *Journal of Productivity Analysis*, 28(1–2), 13–32.

Fusco, E., Vidoli, F. & Rogge, N. (2020) Spatial directional robust benefit of the doubt approach in presence of undesirable output: an application to Italian waste sector. *Omega*, 94, 102053.

Glass, A.J., Kenjegalieva, K. & Sickles, R.C. (2016) A spatial autoregressive stochastic frontier model for panel data with asymmetric efficiency spillovers. *Journal of Econometrics*, 190(2), 289–300.

Kenjegalieva, K., Simper, R., Weyman-Jones, T. & Zelenyuk, V. (2009) Comparative analysis of banking production frameworks in eastern European financial markets. *European Journal of Operational Research*, 198(1), 326–340.

Li, Q. (1996) Nonparametric testing of closeness between two unknown distribution functions. *Econometric Reviews*, 15(3), 261–274.

Mairesse, J. & Mulkay, B. (2007) An exploration of local RD spillovers in France. *Annales d'Économie et de Statistique*, (87/88), 145–166. https://www.jstor.org/stable/27650046

Murakami, D. & Griffith, D.A. (2019) Eigenvector spatial filtering for large data sets: fixed and random effects approaches. *Geographical Analysis*, 51(1), 23–49.

Murty, S., Russell, R.R. & Levkoff, S.B. (2012) On modeling pollution-generating technologies. *Journal of Environmental Economics and Management*, 64(1), 117–135.

Ohene-Asare, K., Turkson, C. & Afful-Dadzie, A. (2017) Multinational operation, ownership and efficiency differences in the international oil industry. *Energy Economics*, 68, 303–312.

Schneider, K., Skevas, I. & Oude Lansink, A. (2021) Spatial spillovers on input-specific inefficiency of Dutch arable farms. *Journal of Agricultural Economics*, 72(1), 224–243

Silva, E. & Stefanou, S.E. (2003) Nonparametric dynamic production analysis and the theory of cost. *Journal of Productivity Analysis*, 19, 5–32.

Simar, L. & Wilson, P.W. (2002) Non-parametric tests of returns to scale. *European Journal of Operational Research*, 139(1), 115–132.

Simar, L. & Wilson, P.W. (2007) Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics*, 136, 31–64.

Simar, L. & Zelenyuk, V. (2006) On testing equality of distributions of technical efficiency scores. *Econometric Reviews*, 25(4), 497–522.

Skevas, I. (2020) Inference in the spatial autoregressive efficiency model with an application to Dutch dairy farms. *European Journal of Operational Research*, 283(1), 356–364.

Skevas, I. & Oude Lansink, A. (2020) Dynamic inefficiency and spatial spillovers in Dutch dairy farming. *Journal of Agricultural Economics*, 71(3), 742–759.

Skevas I. and Skevas T. (2021). A generalized true random-effects model with spatially autocorrelated persistent and transient inefficiency. *European Journal of Operational Research*, 293, 1131–1142 doi: https://doi.org/10.1016/j.ejor.2021.01.004.

Skevas I., Zhu X., Shestalova V., and Emvalomatis G. (2018). The impact of Agri-environmental policies and production intensification on the environmental performance of Dutch dairy farms. *Journal of Agricultural and Resource Economics*, 43 (1835-2018-3859),423–440.

Skevas, T. & Grashuis, J. (2020) Technical efficiency and spatial spillovers: evidence from grain marketing cooperatives in the US Midwest. *Agribusiness*, 36(1), 111–126.

van den Berg, F., Tiktak, A., Hoogland, T., Poot, A., Boesten, J., van der Linden, A. et al. (2017) *An improved soil organic matter map for GeoPEARL_NL: model description of version 4.4. 4 and consequence for the Dutch decision tree on leaching to groundwater*. Tech. rep. Wageningen/the Netherlands: Wageningen Environmental Research.

Vidoli, F. & Canello, J. (2016) Controlling for spatial heterogeneity in nonparametric efficiency models: an empirical proposal. *European Journal of Operational Research*, 249(2), 771–783.

Weiss, C.R. (1999) Farm growth and survival: econometric evidence for individual farms in Upper Austria. *American Journal of Agricultural Economics*, 81(1), 103–116.

Zhang, N. & Choi, Y. (2014) A note on the evolution of directional distance function and its development in energy and environmental studies 1997–2013. *Renewable and Sustainable Energy Reviews*, 33, 50–59.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Skevas, I., Lansink, A.O. & Skevas, T. (2022) Analysing inefficiency in a non-parametric spatial-dynamic by-production framework: A *k*-nearest neighbour proposal. *Journal of Agricultural Economics*, 00, 1–19. Available from: https://doi.org/10.1111/1477-9552.12522