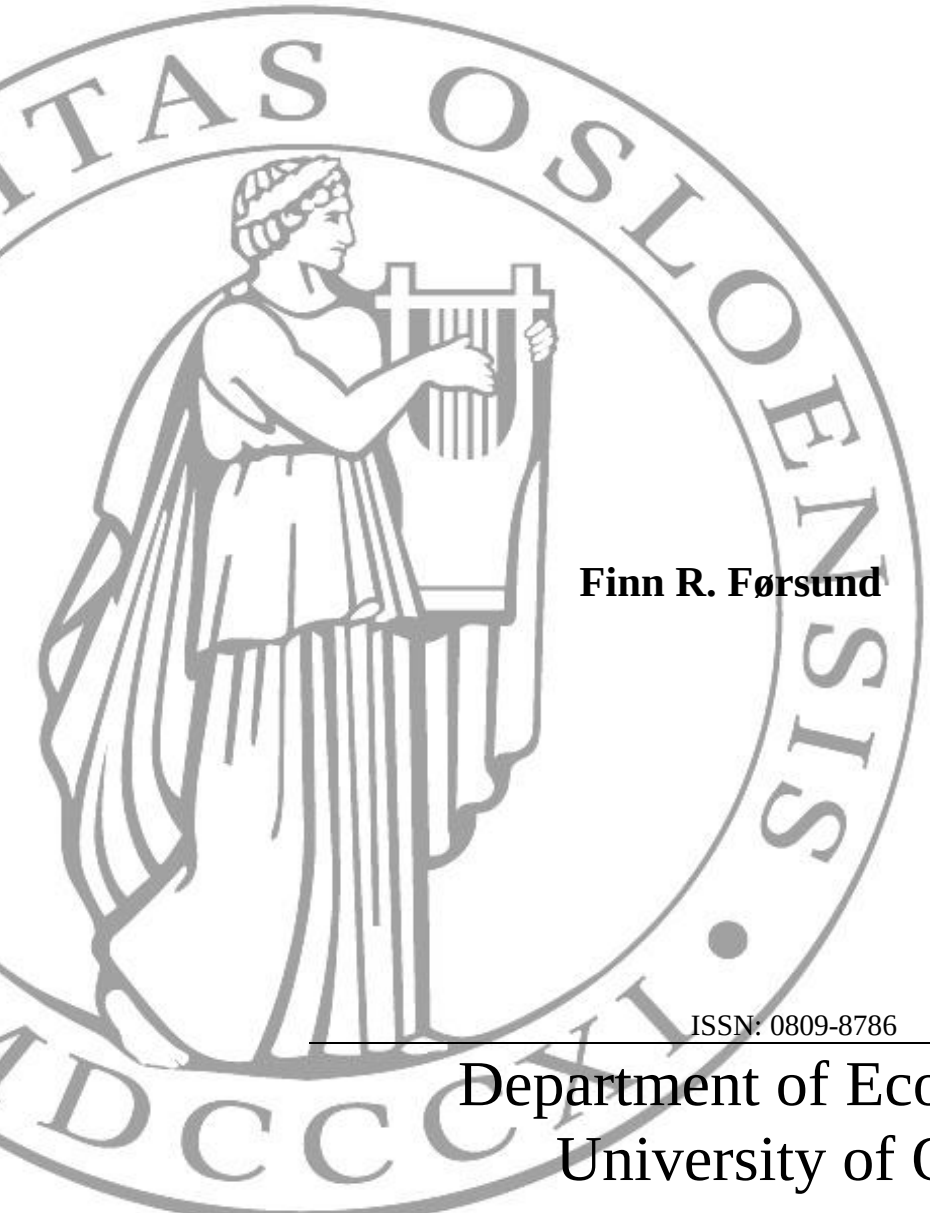
MEMORANDUM

No 09/2019
November 2019

Environmental Performance Measurement: The Rise and Fall of Shephard-inspired Measures



Finn R. Førsund

ISSN: 0809-8786

Department of Economics
University of Oslo

This series is published by the
University of Oslo
Department of Economics

P. O.Box 1095 Blindern
N-0317 OSLO Norway
Telephone: + 47 22855127
Fax: + 47 22855035
Internet: http://www.sv.uio.no/econ
e-mail: econdep@econ.uio.no

In co-operation with
The Frisch Centre for Economic Research

Gaustadalleén 21
N-0371 OSLO Norway
Telephone: +47 22 95 88 20
Fax: +47 22 95 88 25
Internet: http://www.frisch.uio.no
e-mail: frisch@frisch.uio.no

Last 10 Memoranda

No 08/2019	Inés Hardoy and Tao Zhang The long and winding road – Labour market integration of refugees in Norway
No 07/2019	Karine Nyborg No Man is an Island: Social Coordination and the Environment
No 06/2019	Kristine Wika Haraldsen, Ragnar Nymoen, Victoria Sparrman Labour market institutions, shocks and the employment rate
No 05/2019	Ragnhild C. Schreiner Unemployed or Disabled? Disability screening and Labor Market Outcomes of Youths
No 04/2019	Vegard Sjurseike Wiborg, Kjell Arne Brekke, Karine Nyborg Collaboration, Alphabetical Order and Gender Discrimination. - Evidence from the lab
No 03/2019	Steinar Holden A new model for wage formation in Iceland
No 02/2019	Karine Nyborg Humans in the perfectly competitive market
No 01/2019	Bjorn Dapi, Ragnar Nymoen, Victoria Sparrman Robustness of the Norwegian wage formation system and free EU labour movement. Evidence from wage data for natives.
No 08/2018	Olav Bjerkholt Ragnar Frisch (1895-1973)
No 07/2018	Geir B. Asheim, Stéphane Zuber Rank-discounting as a resolution to a dilemma in population ethics

Previous issues of the memo-series are available in a PDF® format
at:http://www.sv.uio.no/econ/english/research/unpublished-works/working-papers/

Environmental Performance Measurement:

The Rise and Fall of Shephard-inspired Measures*

by

Finn R Førsund

Department of Economics, University of Oslo

Abstract The generation of unintended residuals when producing intended outputs is the key factor behind our serious problems with pollution. The way this joint production is modelled is therefore of crucial importance for our understanding and empirical efforts to change economic activities in order to reduce harmful residuals. Estimation of efficiency and productivity when producing both intended and unintended outputs has emerged as an important research strand. The most popular models in the field are based on weak disposability between the two types of outputs and null jointness introduced by Shephard. The purpose of the paper is to show that these model types are seriously flawed. An alternative model based on the production theory of Frisch introduces technical jointness for the case when the unintended output is unavoidable. The materials balance based on physical laws tells us that when material inputs are used unintended outputs are unavoidable. The modelling of joint production must therefore reflect this. A key feature is that the two types of outputs should be separated using different production relations. This facilitates estimating two independent frontiers and calculating efficiency scores and Malmquist productivity changes for the two types using a non-parametric DEA model.

Keywords Intended and unintended outputs; Joint production; Materials balance; Technical jointness; Pollution; Weak disposability

JEL Classification C14, D24, D62, Q50

* Preliminary versions of the paper were presented at the World Congress of Environmental and Resource Economists WCERE 2018 in Gothenburg, and as keynote presentation at the European Workshop on Efficiency and Productivity Analysis EWEPA 2019 in London. I am indebted to the discussant R. Robert Russell at EWEPA, and Rolf Färe for suggestions improving the paper.

1. Introduction

A crucial building block in environmental economics is the phenomenon of joint generation of intended outputs and unintended ones[1] in production and consumption activities. The discharge of undesirable outputs causes the ubiquitous environmental problems facing humankind today. Estimation of efficiency and productivity when producing both desirable and undesirable outputs has emerged as an important research strand. However, mainstream environmental economics has hardly considered inefficiency issues. The literature spawned by the Porter hypothesis (Porter (1991); Porter and van den Linde 1995) is an exception (empirical studies and critique of the hypothesis are extensively reviewed in Brännlund and Lundgren (2009); Lanoie et al (2011); Ambec et al 2013).

Based within the inefficiency research strand Färe et al (1986); (1989) pioneered the issue of measuring inefficiency empirically when producers generate both desirable and undesirable outputs based on theoretical schemes presented in Shephard (1970) introducing jointly weak disposability of intended and unintended outputs, and null jointness (formally, null jointness was defined in Shephard and Färe 1974). Färe and Grosskopf (1983) developed the theoretical ideas into explicit efficiency measures.[2] Up to now, the Shephard- inspired models has completely dominated the literature on efficiency when producing simultaneously intended and unintended outputs.

The introduction of a directional distance function in Chung et al (1997) lead to the widespread adoption of this approach in the literature and replaced the hyperbolic efficiency measure used in Färe et al (1989). The output-oriented radial distance function of Shephard (1970) was generalised using a distance function that adds to the desirable output of an inefficient observation and

[1] In the related literature, intended outputs are also called desirable outputs, good outputs, or just goods. Unintended outputs are also called not desirable, undesirable, waste, and bads. Bad output means that the consumers' willingness to pay for a reduction of the bad is positive. The neutral or generic name for waste is residuals. I will use these terms interchangeably.

[2] DEA was used to calculate two measures, one based on assuming strong disposability and the other weak disposability and then the ratio was interpreted as the loss of specifying weak disposability. The approach was followed up empirically in Färe et al (1986).

subtracts from the observed inefficient level of the undesirable output in order for a projection of the observation to be on the frontier. The calculation of the added/subtracted values where done using the same scalar factor multiplied with the observed values of both types of outputs, using as the direction of scaling the observed output values.

The rise of the Shephard-inspired models of joint production of intended and unintended outputs in journals has been rather spectacular, with Färe et al (1989) having 805 citations in Web of Science per 29.08.2019[3] and Chung et al (1997) having 829 citations in Web of Science per 29.08.2019. In comparison, Førsund (2009)[4] has 106 citations in Web of Science per 29.08.2019, and Murty et al (2012) 99 citations in Web of Science per 29.08.2019. However, the tide may turn.

A crucial feature of the technology specification is the unavoidable generation of residuals. I am mainly interested in the residuals that cause environmental damage, identified as positive willingness to pay to reduce the damage, and residuals are then called pollutants. The materials balance, introduced in environmental economics in Ayres and Kneese (1969), expresses the essential insight that the material content of inputs cannot disappear, but must be part of the intended outputs or become residuals discharged to the natural environment.[5] The materials balance reflects the two thermodynamic laws of conservation of matter and energy. Due to entropy, there is a minimum of energy and materials that will not be contained in the intended output. The pervasiveness of residuals generation then follows. The general situation building on materials balance insights is that unintended outputs are function of materials and energy inputs and not a function of intended outputs.

The seminal papers Färe et al (1989); Chung et al (1997) answered the question of how to calculate efficiency measures when both intended and unintended outputs are produced. This achievement was impressive. However, the Shephard-inspired approach they used are not without serious flaws. The main purpose of the paper is to expose the problems of Shephard-inspired measures, and to offer an approach based on a specific form of joint production satisfying the materials balance

[3] Färe et al (1986) had a more modest impact of 82 citations in Web of Science per 29.08.2019.
[4] To the best of my knowledge, Førsund (1998) (journal version in 2009) was the first paper to criticise the weak disposability assumption.
[5] In Ayres and Kneese (1969) the materials balance was explored assuming fixed relationships between material inputs and outputs. The use of linear relationships with fixed coefficients served their purpose of demonstrating the pervasiveness of residuals generation, but lacked flexibility regarding technology. Leontief (1970); Leontief and Ford (1972) extended the input-output model introducing fixed coefficients between residuals and intended outputs. However, I will not pursue models with fixed coefficients here (see Førsund 1985).

identity. The alternative approach is based on a separation into two types of production functions, one for intended outputs and another for unintended outputs.[6]

My critique is not based so much on technical or mathematical insights as to an understanding of how to model the production relationships for an intended output when the creation of unavoidable unintended products causing negative externalities are also produced simultaneously. For the analysis, I will use mostly production functions with continuous partial derivatives of first and second order, and assuming that the implicit function theorem is valid. Of course, these are stricter assumptions than necessary. Starting out with some reasonable assumptions or axioms about production sets and then deriving their properties will yield richer results as to the generality of the analysis, and may be required for disentangling disposal properties of multiple equations (Murty and Russell 2017; Murty and Russell 2018). However, it is not necessary for my purpose to go for maximal generality.

The plan of the paper is to discuss the materials balance in Section 2, and to present types of joint production in Section 3. The seminal approaches of Shephard (1970); Baumol and Oates (1988) are presented and discussed in Section 4. Two recent alternatives to the Shephard-inspired models are presented and discussed in Section 5. The key model development in the paper is based on the factorially determined multi- output production functions of Frisch (1965). Section 6 is summing up the critique of Shephard-inspired models. Section 7 discusses the efficiency concept and shows how to estimate efficiency measures in the case of both intended and unintended outputs using a non-parametric DEA model. Section 8 concludes.

2. The materials balance

The mass of material inputs appears in the materials balance relation, and it is therefore convenient to operate with two classes of inputs (Ayres and Kneese 1969, p. 289); material inputs (tangible raw materials) x_M and non-material inputs that I will call service inputs x_S. These latter inputs are not "used up" or transformed in the production process. The materials balance tells us that mass

[6] This separation was done - in an environmental economics context - already in Førsund (1972); (1973), based on the classification of systems of production functions in Frisch (1965) termed factorially determined multi-output production functions. I return to this in Section 5.

contained in material inputs cannot disappear, but must be contained in the products y or end up as residuals z. The residuals are discharged to the natural environment. The variables in the materials balance relation must be expressed in the same unit of measurement. Weight of mass is a natural unit of measurement. The weight of the different inputs containing a specific substance k can then be summed over the number of material inputs $j = 1,\ldots, n_M$. Part of this substance ends up in intended outputs $i = 1,\ldots, m$ if they are of the material kind. The difference between the mass of substance k in the material inputs and the mass of substance k contained in the m types of outputs is the amount of substance k discharged to Nature, measured in the same weight unit as the substance in material inputs and in intended outputs. However, the residual may be discharged to Nature in different forms, e.g. CO_2, CO, tar, ash, etc., that can be classified as different types $r = 1,\ldots, R$. For example, coal used in producing electricity contains carbon, but in the combustion process, oxygen is picked up and CO_2 is emitted to air. A coefficient c_{rk} measures the amount of the substance k contained in residual of type r per unit of total discharged residual z_k. The weights a_{jk}, b_{ik}, c_{rk} convert the unit of measurements commonly used for the variables (piece, length, area, volume, etc.) into weight. The general materials balance can then be written:

$$\begin{aligned} &\sum_{j=1}^{n_M} a_{jk} x_{Mj} \equiv \sum_{i=1}^{m} b_{ik} y_i + \sum_{r=1}^{R} c_{rk} z_k \,(k = 1,\ldots, K), \\ &\sum_{k=1}^{K}\sum_{j=1}^{n_M} a_{jk} x_{Mj} \equiv \sum_{k=1}^{K}\sum_{i=1}^{m} b_{ik} y_i + \sum_{k=1}^{K}\sum_{r=1}^{R} c_{rk} z_k . \end{aligned} \tag{1}$$

The coefficient a_{jk} in front of material inputs x_{Mj} tells us the mass of substance k in a unit of x_{Mj}, the coefficient b_{ik} in front of intended output y_i is the mass of substance k contained in a unit of the output y_i, and the coefficient c_{rk} in front of the residual z_k contains the mass of substance k in type r of the emitted residual. If it is the type of residual r that is used as the definition of the residual, then the carbon in coal must be converted to units of CO_2, etc.[7]

The first line in (1) shows the mass balance for one type of substance k (see Baumgärtner and de Swaan Arons 2003, footnote 5, p. 121). However, the balance is here extended to cover the different types of residuals r containing the substance k. The second line shows the total mass

[7] Notice that the parameters a_{jk} and c_{rk} are not emission coefficients of standard definition; an emission coefficient for a material input tells us the amount of the emitted residual of type r (e.g. CO_2) that is created per unit of the input x_{Mj} (e.g. coal).

balance for a production unit. In the case k is only appearing in a single type of residual, i.e., r =1, then $c_{rk} = c_k$. However, the distribution on different types r for substance k may change, as when a combustion process transforms the material inputs, and temperature, pressure, supply of oxygen, etc., vary. Variable mix of types of emissions all containing a common substance implies inefficiency in some of the operations.

The creation of residuals during the production process also contain materials provided free by nature: oxygen for combustion processes and oxygen used to decompose organic waste discharged to water (biological- oxygen and chemical demand, BOD and COD), nitrogen oxides created during combustion processes, and water for pulp and paper that adds to the weight of residuals discharged to the environment. Such substances must either be added to the left-hand side as material inputs - and then contained in the residuals z - or we can focus on the actual materials in inputs and redefine z accordingly, like calculating the carbon content in weight for all three types of variables and not measure residuals as CO_2 or CO, etc. This is what we have done in (1).

For each production unit we have an *accounting identity* for the use of materials contained in the input x_{Mj}. The relation holds as an identity meaning that it must hold for any accurately measured observation, being efficient or inefficient. The relation should not be regarded as a production function, but serves as a restriction on specifications of these[8].

The importance of the materials balance is the insight that generation of unintended residuals cannot be avoided. However, measuring all the factors involved in the materials balance accurately may not be so easy, especially on the more aggregated level that is commonly used in efficiency analyses. If we accept that residuals are measured accurately, we know that all observations of production units, efficient as well as inefficient units, must obey the materials balance as an identity. If we do not have observations, but data that are theoretical it may not be feasible to assign the materials balance accurately to hypothetical observations based on observed ones.

[8] Several authors, among them Pethig (2006); Ebert and Welsch (2007) extend the materials balance to make what are their production functions.

3. Joint production

A problem with the Shephard-inspired approaches is that the nature of joint production when dealing with unintended outputs is not discussed in any of the papers following the Shephard (1970) approach, including Shephard himself. The lack of clarification of the nature of joint production when desirable and undesirable outputs are produced is a key reason for the Shephard inspired approaches developing unsatisfactory modelling of efficiency for intended and unintended outputs.

Frisch (1965, Chapter 14a-d, pp. 269-281) stated that in the case of multi-output production "…the production law cannot be studied separately for each separate product, but must be considered simultaneously for all connected products." He introduced already in the introductory Chapter 1 three types of joint production defined generally as having "… some kind of technical connection between several products…" (p.11).

The types of joint production are:

a) Assorted production: Inputs can be applied alternatively to produce different products; agricultural land can be used for different crops, a wood cutting machine can be used to making different objects. An assortment of outputs is produced. The inputs are then output-specific. The technical connection between outputs making it joint production is that the same type of inputs are used to produce the outputs.[9]

b) Technical jointness: Standard classical examples are given by agricultural production; sheep yield mutton and wool, hens yield eggs and poultry, growing wheat also yields straw, and coke and gas is gotten from coal as input, to name a few classical examples. The connections between outputs is also based on common inputs as for assorted production. The main difference to assortment is that the inputs are not product specific; it is not possible to reallocate amount of inputs on different outputs. However, the mix of output can change if the mix of inputs change;

[9] In Murty and Russell (2017, p. 3) assorted production is called rival production, and they distinguish this from joint production.

examples in Frisch (1965) are change of feed to hens changing the mix of eggs and poultry meat, and changing types of sheep from a high share of wool compared with meat to the opposite.

c) Extreme jointness: Fixed proportions between outputs independent of inputs; as in distillates of crude oil, and pure factor bands, i.e. relations between factors independent of outputs. The former case is called *complete* [*product*] *coupling* in Frisch (1965, p. 273). If we assume fixed input-output coefficients as in the Leontief input – output case this case belongs to the category of extreme jointness.

However, unintended outputs are not mentioned in Frisch (1965). Examples from today's industrial activities using material inputs generating residuals are ubiquitous, e.g., pulp and paper industry, steel production industry, cement, oil refineries, fossil fuel-based electricity generation to mention just a few.[10]

The classical writers[11] introduced three types of outputs; intended outputs that have positive prices in a market, by-products that also have positive prices, but contribute rather less to the revenue, and waste that has no economic value. Jevons (1883, p. 142)[12] considered the case of joint production typical, stating: "… I shall point out that these cases of joint production, far from being 'some peculiar cases' form the general rule, to which it is difficult to point out any clear or important exceptions."

The examples above do not connect waste to intended products. However, Jevons (1883, p. 144) remarks "The waste products of a chemical works, for instance, will sometimes have a low value; at other times it will be difficult to get rid of them without fouling the rivers and injuring the neighbouring estates; in this case they are discommodities and take the negative sign …". He included many forms of industrial production as examples of all three types of outputs.

In the case of assorted production, resources can be reallocated among outputs. If this reallocation is without limits unintended outputs will, of course, be set to zero by an efficient producer (Førsund

[10] Førsund and Strøm (1974) extended the multi-sectoral model (MSG model) in Johansen (1960), using 38 types of waste from 26 production sectors based on data from 1970 in a projection exercise from 1970 to 2000. Førsund and Strøm (1976) used 35 types of waste from 86 production sectors for data from 1970. Førsund (1985) used 37 types of waste from 123 production sectors based on data from 1978.

[11] An extensive survey of joint production in classical texts is found in Kurz (1986).

[12] The first edition was published in 1881. The third 1883 edition is available on the internet. The latter edition is identical to the second edition concerning the main text.

2009). We must have the case of technical jointness (including extreme jointness) when unintended outputs are generated.

The consequence of generating an unintended output is thus that a firm operating a technology efficiently, will by definition generate as little as possible of the unintended output; the minimum dictated by the technology used given the input quantities. The material inputs are used to produce intended outputs, and materials contained in the residual come at the expense of producing them. Thus, to be efficient for given inputs there is a minimum of an unintended output that is unavoidable. The materials balance (1) shows the split of material inputs on intended and unintended outputs.[13] It is meaningless to split non-material inputs on intended and unintended outputs because the generation of intended and unintended outputs take place simultaneously. There is only a single common process. Unintended residuals cannot be generated in separate processes from intended outputs. When formulating production relations this feature must be taken seriously. Introducing unintended outputs makes this to be a special case of Frisch (1965) factorially determined multi-output production within the category of technical jointness. We will come back to this in Section 5.

4. Early models for production of intended and unintended outputs

4.1 The Shephard model

The theoretical model in Shephard (1970) for producing simultaneously intended and unintended outputs[14] based on assuming *weak disposability* has up to now completely dominated the empirical

[13] In the case of non-material output like electricity a given amount of material inputs used (e.g. coal) will generate a specific amount of residuals independent of intended output but residual mix may change if there is inefficiency in production. My assumption will then be that to realise the frontier function generation of electricity is done using the installed technology efficiently.

[14] Cf. Shephard (1970, Chapter 9, p. 178): "Here we are concerned with technologies which yield several different joint products for a given input vector of the factors of production. For the most general treatment, all of these products need not be desirable or have positive economic or social value. In particular, waste products, which lead to pollution of air, stream and land and cost society for their control, may be explicitly treated as part of the joint outputs of the technology."

literature on efficiency for that case. The general point of departure in the literature is to characterise the technology by formulating the general production possibility set *T*:

$$T = \{ (y,z,x) \,|\, y \geq 0 \text{ and } z \geq 0 \text{ can be produced by } x \geq 0 \}. \quad (2)$$

We regard the variables as vectors. Here *y* is the intended output vector, *z* is the unintended output vector and *x* is the vector of inputs. The production possibility set is conventionally defined as containing all possible ways of producing given outputs. Assumptions about specific properties, presumably based on a combination of how the real world functions, and the practical and analytical needs for simplifications, are stated so many times in the literature that this is skipped here. Suffice it to say that the set is assumed to be convex, closed, and allowing no free lunch. (See Coelli et al (2005) for an elementary introduction and Cooper et al (2007) for a more advanced treatment.) It is rather obvious that if no material inputs are consumed, no material residuals will be generated.[15]

The technology set (2) can equivalently be characterised by the output set $P(x)$ or input set $L(y, z)$:

$$\begin{aligned} P(x) &= \{ (y,z) \,|\, (y,z) \text{ can be produced by } x, (y,z,x) \in T \}, \\ L(y,z) &= \{ x \,|\, \text{at least } x \text{ is required to produce } (y,z), (y,z,x) \in T \}. \end{aligned} \quad (3)$$

These sets are bounded. The border of the sets represents efficient operations. If the efficient operations could be formulated by a function this function would represent the frontier production function, and the output- and input isoquants would belong to this frontier function. Points in the interior of the production possibility sets are inefficient per definition.

It is obvious that the general characterisations of the production possibility set *T* and output- and input sets are not meant to tell us about the nature of the joint production involved. However, output- or input distance functions are introduced as description of technology. Since the formulations of technology using distance functions do not exclude assorted production, restrictions must be introduced ruling out such a form of joint production, as will be explained in Section 5.

[15] However, we also have non-material residuals stemming from energy use, like noise. Undesirable outputs functioning as public bads belong to a subclass of outputs generating what is termed negative externalities in the literature.

The weak disposability assumption

A way out of the assorted production problem was introduced in Shephard (1970) formulating weak disposability between the intended and unintended outputs:

$$\text{If } (y,z) \in P(x), \text{then } (\theta y, \theta z) \in P(x) \text{ for } 0 \leq \theta \leq 1 \tag{4}$$

This condition is adapted in the subsequent literature. However, although there is an extensive discussion of joint production also involving undesirable outputs in Shephard (1970, Section 9.5), assorted production is not mentioned or recognised as a problem; the concern is about disposability properties of the two types of outputs. The condition (4) says that *if* realisations of the two types are reduced proportionally, then the new points will belong to the production possibility set $P(x)$.[16] The consequence of assumption (4) is illustrated by the original Figure 32(a) in Shephard (1970, p. 188), here Fig. 1. The solidly drawn frontier segments are efficient parts of the output sets, and

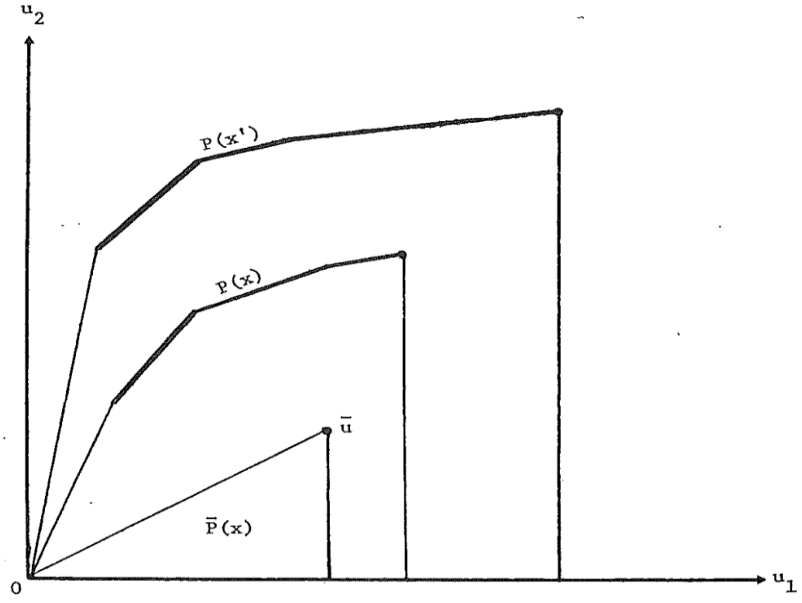


FIGURE 32 (a): OUTPUT SETS FOR A PRODUCTION CORRESPONDENCE WITH WEAK DISPOSAL (u_1 NOT DESIRABLE), $x' \geqq x$

***Figure 1**. Output sets P(x) obeying weak disposability for outputs u_1 (undesirable) and u_2 (desirable) for given levels of inputs x*
Source: Shephard (1970, p. 188)

[16] This is not exactly the same as saying that (4) imposes that the two types must change *proportionally* as is often done in the literature (see e.g. Dakpo et al 2016, p. 351). Taking the piecewise linear frontier isoquants in Fig. 1 at face value it is easy to see that the change in outputs along the segments is not proportional.

the thin lines starting on the left down to the origin show that the intended output (here u_2) and the unintended one (u_1) have what in Shephard and Färe (1974) is called null-jointness. We see that any point on the efficient parts satisfies the condition (4) due to convexity of the sets.

The point $\bar{u}$ and the set $\bar{P}(x)$ represents the case of a constant relationship between the intended and unintended outputs and thus conforms to extreme jointness as defined in Section 3. There is just a single efficient point $\bar{u}$ in the set $\bar{P}(x)$ (the thin line from the point to the origin does not belong to an efficient face) (Shephard 1970, p. 188).[17] Shephard did not expand on how to measure efficiency. He used distance functions to identify efficient border sets. As mentioned in Section 1 the use of the theoretical model of Shephard was first implemented empirically in Färe et al (1986); (1989). Distance functions are used in estimating efficiency scores within the strand of Shephard-inspired modelling.

The output oriented directional distance function $\vec{D}_o$ has been the preferred model after Chung et al (1997):

$$\vec{D}_o(x, y, z; g_y, -g_z) = \max\{\beta : (y + \beta g_y, z - \beta g_z) \in P(x)\}, \vec{D}_o(x, y, z; g_y, -g_z) \geq 0 \quad (5)$$

Instead of a radial direction of projections to the frontier of inefficient points, a projection point is found by adding to the observed intended output following a chosen direction g_y and a subtracting from the observed unintended output following the direction g_z. It is most common to set $g_y = y$ and $g_z = z$, i.e., equal to the observations. However, the projection to the frontier is crucially dependent on the existence of a trade-off isoquant between intended and unintended outputs. Assuming differentiability, as is often done (Färe et al 2013, p. 111), then $(\partial \vec{D}_o(x, y, z; g_y, -g_b) / \partial z) / (\partial \vec{D}_o(x, y, z; g_y, -g_b) / \partial y)$ is the rate of transformation between the good and the bad for given inputs. This ratio is used for estimating shadow price of the residual (Färe et al 2013, Eq. (12) p.111), and the trade-off curve is illustrated there and in numerous papers by Färe et al and other authors of similar models.

[17] The case of extreme jointness implies proportionality between the two types of output independent of input levels for frontier functions as drawn in Fig. 1 with only the point $\bar{u}$ in set $\bar{P}(x)$ being efficient, but this property does not seem to follow from (4) concerning the efficient border of the sets.

4.2 The externality model of Baumol and Oates

In Baumol and Oates (1988) (first edition 1975) that has been an influential book on environmental economics, both desirable and undesirable outputs were introduced in the context of an environmental externality model. Although inefficiency aspects and efficiency measures were not discussed, their externality model is interesting because it led to a discussion later whether unintended outputs are inputs instead. A production possibility set was specified by using a single transformation function relation $F(y,z,x) \leq 0$ extended with residuals vector *z* where *y* is the intended output vector, and *x* the input vector.[18] The relation $F(y,z,x) = 0$ defines the border of the set and is called the transformation relation. This is an implicit representation of the efficient production technology that we call the frontier. Inefficient points yield function values $F(y,z,x) < 0$. However, Baumol and Oates (1988) do not study inefficiency, but are only interested in the frontier. In economics, it is commonly assumed that the transformation function is differentiable and have continuous partial derivatives of first and second order. This is also the case in Baumol and Oates (1988). In addition, it is also common to assume that the implicit function theorem is valid. A standard convention is that increasing an output at a frontier point will increase the function value, and increasing an input from a frontier point will decrease the function value. We then have $F'_y > 0, F'_x < 0$. This signing conforms to regarding *y* and *x* as being freely disposable.[19] The question is how to sign the partial derivative of the residual. Differentiating the transformation function w.r.t. *y*, *z* and *x*, assuming for simplicity single variables of each type, yields:

[18] The model of Baumol and Oates (1988, pp. 37- 40) also include consumer utility functions in intended consumer goods with positive marginal utility and unintended residuals being pollutants with negative marginal utilities. The purpose of the modelling was to find maximum utility given the resources.

[19] For a frontier point or an inefficient point inside the production possibility set $F(y,z,x) \leq 0,$ reducing *y* for a given *x* or increasing *x* for given *y*, both moves reduce the function value.

$$
\begin{aligned}
&F_y'(y,z,x)dy + F_x'(y,z,x)dx = 0 \Rightarrow \frac{dy}{dx} = -\frac{F_x'(y,z,x)}{F_y'(y,z,x)} > 0 \\
&F_z'(y,z,x)dz + F_x'(y,z,x)dx = 0 \Rightarrow \frac{dz}{dx} = -\frac{F_x'(y,z,x)}{F_z'(y,z,x)} \\
&F_y'(y,z,x)dy + F_z'(y,z,x)dz = 0 \Rightarrow \frac{dy}{dz} = -\frac{F_z'(y,z,x)}{F_y'(y,z,x)}
\end{aligned}
\tag{6}
$$

The first relation defines the standard positive marginal productivity of the input *x*. If $F_z' > 0$ for the unintended output in the second line an increase in the input *x* will also give an increase in the unintended output. However, this implies that the value of *F*(.) increases, in spite of the consumers valuing this output negatively (given that the unintended output is an environmental pollutant). Thus, having $F_z' > 0$ is not a property our model should have. Assuming assorted production, this problem is solved reallocating all resources to producing the intended output *y* and zero unintended output (Førsund 2009). However, it clearly goes against the main problem with joint production of intended and unintended outputs that generation of the unintended outputs is unavoidable

Assuming that the partial derivative of *F*(.) with respect to the unintended output is negative, i.e. *as if z* is an input, we see in the second line that this implies that there is a substitution between the input *x* and the variabel *z*; increasing *x* reduces *z*. However, if *x* is a material input *z* cannot be reduced if *x* increases. This goes against the materials balance that tells us that *z* increases if material input increases.

Furthermore, adopting the positive sign of the unintended output the third relation shows a trade-off between the intended and the unintended outputs, if one of them increases, the other has to decrease. But this is what happens when assuming assorted production and then optimality implies that *z* is set to zero, and this is impossible given that *z* is unintended.

The residual *z* is not only unintended, but also unavoidable. The firm has no choice but to produce the pollutant. The negative trade-off appearing when both partial derivatives of *y* and *z* are positive cannot be realised except in the case of assorted production. If this is the case, then reallocating resources can reduce the residual *z* in order to producing more of the intended output *y*. But this is per definition the type of joint production that is not possible in the case of unintended outputs; the joint production cannot be assorted production when an output is unintended, but must either be the type technical jointness or extreme jointness.

However, Baumol and Oates (1988) do not discuss the implication of the type of jointness. They "solve" the dilemma - without informing the reader - simply by assuming that the partial derivative of the residual is negative; $\partial F / \partial z < 0$ (see Table 4.1, in Baumol and Oates (1988, p. 39), *as if* the residual is an input. Then we have $-F_z' / F_y' > 0$. An increase (decrease) in *z* now increases (decreases) *y*. To reduce the residual generation *z* at a frontier point is costly in terms of reduced intended output *y*. However, the residual is definitely an output and not an input. What is missing here is the fact that there is no direct substitution between the two types of outputs when we have technical jointness. The generation of both types of outputs occurs simultaneously by use of a given set of inputs. There is no interaction possible between the two types of outputs for fixed inputs. Assuming that our three variables are all single, then we have the classical definition of efficient production that for given input *x* output *y* is maximised. To treat the residual just as a normal output does not make sense, because the production cannot be efficient if the pollutant is to be maximised. The opposite is the case; efficient production implies that the residual has to be as small as the technology allows for given resources in order to maximise intended output. Regarding *z* as an input does not work because substitution between *x* and *z* as inputs for given intended output is impossible according to the materials balance.

There is a confusion here in the literature. A standard mistake is to disregard the micro setting of production and thinking at a more aggregated level implying the resources can be used to abate pollution and thus take resources away from production of intended output. However, at the micro level, a firm's use of resources must be explicitly specified, and this is not the case in literature claiming the unintended output is an input (see Førsund (2009) for a critique of the assumption that the unintended output can be treated as an input).

5. Production functions satisfying technical jointness

5.1 Factorially determined multi-output production functions

In order to represent the generation of the unavoidable residual in a way conforming with technical jointness, and solving the dilemma posed by assuming that the unintended output is an input, two

separate equations can be introduced; one equation for the intended output and another for the unintended one. The crucial point is that both outputs are produced simultaneously, and are functions of the same set of inputs; i.e., the inputs are not specific for each type of output. It is the analyst that introduces two production functions in order to formulate a model getting a grasp on the situation. However, physically there is only one activity with simultaneous generation of both intended and unintended outputs.

Since the process of generating both types of output is a simultaneous process, it seems rather obvious that the same inputs must be specified in both functions. When joint production was discussed in Frisch (1965, pp. 270-276), he introduced just a type of technical jointness that fits our case. He named the type as *factorially determined multi-output production*. Each output has a *separate* production function, but the inputs are the same for all functions. Frisch underlined that having this type of technical jointness the mix of outputs is not necessarily fixed, but can vary with varying mix and level of inputs.

Frisch specified only intended outputs with positive demand and specified traditional production functions for them. However, the separation property can also be extended to the joint production of intended and unintended outputs. As stated previously, residuals are generated simultaneously with the intended outputs and stem from the raw materials employed as inputs. It seems important to satisfy these physical realities arising from use of material inputs in any sound modelling of the interaction of economic production activity and generation of pollutants.

The model from the production theory of Frisch (1965) of product separability, the factorially determined multi-output model, seems tailor-made for capturing the physical process of generation of residuals simultaneously with desirable outputs. Single-output production functions for each unintended residual are added to the single-output functions for intended outputs:

$$
\begin{aligned}
&y = f(x_M, x_S),\ f'_{x_M},\ f'_{x_S} > 0\ ,\ f''_{x_M},\ f''_{x_S} \le 0 \\
&z = g^*(x_M, x_S),\ g'^*_{x_M} > 0, g''^*_{x_M} \ge 0,\ \ g'^*_{x_S} = 0 \text{ for all } x_S, x_M \ge 0 \\
&\Rightarrow z = g(x_M),\ g'_{x_M} > 0, g''_{x_M} \ge 0
\end{aligned}
\tag{7}
$$

The production functions for intended outputs are assumed to have the standard properties of a

neoclassical production function with positive (but decreasing) marginal productivities of inputs implying substitution possibilities. Notice that the two functions in (7) are frontier functions.[20] Regarding the unintended output the service input x_S has no influence on the level of the unintended output z resulting from any (non-negative) value of the service input in the $g^*(.)$ function. This property make the production function specification a special subcase of the Frisch system of factorially determined multi-output production.[21] When considering substitution along an isoquant of $f(.)$ there is a relation – the input isoquant – between the inputs keeping the value of the $f(.)$ function constant. Regarding e.g. cost minimisation as economic adaptation of the firm implies that the choice of input levels depends on the price ratio between the inputs. The choice of x_M in the production of y then determines the level of residual z.

These two production function types represent efficient functions. The intended output y is maximal for given inputs, and the unintended output is minimal for the given material input x_M. The function for the residual z is a function of materials input only because it is the mass incorporated in the material input that constitutes the unintended residual when mass from input contained in intended outputs are taken into consideration.[22] The materials balance in (1) shows the distribution of mass on material inputs, intended-, and unintended outputs. There are the usual substitution possibilities between the material and the non-material inputs when producing intended outputs. However, the unintended residual is just the mass of the input x_M that is not contained in the intended output. This is different from the example of technical jointness of wool and mutton where both outputs are desirable and have a positive market demand. The residual in that example may be the sheep excrements.[23] Another classical example of joint production is that

[20] Regarding disposability properties of the functions the intended output and the two inputs are freely disposable, but this is not the case for the unintended output and material input in the $g(.)$ function; for given x_S, z can only be reduced by reducing x_M and this is costly because y is then also reduced.

[21] Unfortunately the function $g^*(.)$ used in Førsund (2018 a,b,c) using also x_S as an input assuming the marginal product of x_S being negative, and the figure illustrating isoquants for both types of outputs in the three publications, are not correct. The relation for the unintended output is correctly specified in Murty et al (2012); Murty and Russell (2018). It may seem that specifying only the material input in the residuals relation in (7) goes against the definition of factorially determined multi-output production. However, this is not the case; the point is that residuals contain the substances present in the material input (disregarding here oxygen taken from the air) and not used in the output, implying that service inputs have no additional impact on the residuals generated by the $g(.)$ function.

[22] In coal-fired electricity generation often used in empirical studies all mass is contained in the residuals.

[23] I assume that the excrements are not used as fertiliser if this type of z is to remain without positive economic value. Anyway, excrements are unavoidable.

a cow gives milk and also meat and hide; all three marketable goods,[24] but the emission of methane gas during digestion is a pollutant with climate change effects. This output is unintended and unavoidable.

Separating a general single transformation function as used in Baumol and Oates (1988) into two functions also solves the problem of maintaining the classical property of a production function that intended output is maximised for given inputs. However, if a single transformation function is used with several outputs as arguments, maximising one intended output at a time keeping the other intended outputs constant for given inputs results in a different production function for each intended output, complicating the usefulness of a single transformation relation.[25]

The production possibility sets can be written:[26]

$$\begin{aligned} y &\leq f(x_M, x_S), \\ \bar{z} &\geq z \geq g(x_M). \end{aligned} \tag{8}$$

Here $\bar{z}$ is the total material contained in x_M. If we consider only one type of substance for convenience, we have from the materials balance (1) that $\bar{z} = ax_M$. Obviously, the maximal amount of residual cannot be greater than this amount, but will be less if the intended output contains materials. Inefficient use of resources producing the intended output results in less intended output than realised on the efficient frontier $f(.)$. At the same time more of the unintended output is produced than what will be produced on the efficient frontier $g(.)$. Efficient use of resources implies that both intended output and unintended output are at efficient levels simultaneously.

Fig. 2 shows the production possibility sets for two outputs, the intended y and the unintended z, and one input in the same diagram, the material one x_M, because the input is common in both production functions. The measuring units are generally different for the outputs, so the placement of the production functions is arbitrary. For ease of illustration the borders of the set, the frontier

[24] Of course, you only get meat and hide after slaughter, while alive the cow gives milk and emits the unintended methane gas during the digestion process.
[25] Shephard (1970) uses output- and input distance function in the case of multiple outputs to define efficient subsets when the values of the functions are 1, see Proposition 65 and 66, p. 214. See also Russell (1998) for difficulties expressing joint production functions with many outputs.
[26] The number of each type of output can easily be extended using single equations as in (7).

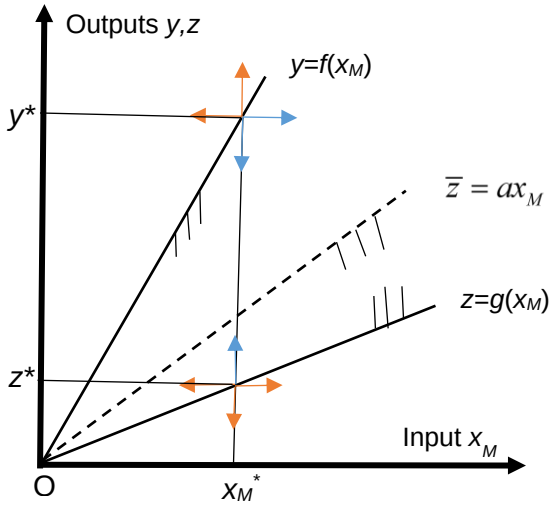


Figure 2. *The two production possibility sets, and efficient and inefficient points*

production functions have constant returns to scale The production possibility set for y is below its border $f(x_M)$ down to the horizontal axis. The frontier point is (y^*, x_M^*). The production possibility set for the unintended output is the area between the frontier $g(x_M)$ and the broken ray that is the upper limit $\bar{z}$ in (8). The frontier point is (z^*, x_M^*). The arrows show movements in y and x_M, and z and x_M, respectively. The horizontal blue arrow to the right from the frontier point y^* shows an increase in input x_M for a constant y, and this point is in the interior of the set and is inefficient. The vertical movement downwards keeping the input x_M^* constant but reducing y is also in the interior of the set and is inefficient. The two red arrows show new points that are outside the set and thus infeasible.

The vertical movement upwards from the frontier point (z^*, x_M^*) following the blue arrow increases the unintended output for a fixed x_M^*, thus creating an inefficient point. However, the other three points are all infeasible; input cannot be decreased (then efficiency increases!), and increasing input for constant z^*, and decreasing z for constant input x_M^* are infeasible.

We have discussed the two sets in isolation. However, the variables are related through the same x_M. Increasing the input to create an inefficient unit in the y set will increase the unintended output. It is only the opposite vertical movements of outputs that are feasible keeping the input at the frontier value of x_M^*. The reduction in y is connected to the increase in z that must satisfy the materials balance. Regarding disposability we see that neither z nor x_M has standard properties, but as expressed in Murty et al (2012, p. 119) "…violates standard disposability with respect to goods that cause (or affect) pollution generation and exhibits costly disposability with respect to

pollution." Without abatement the only option to reduce the unintended output is to reduce x_M and thereby also reducing the intended output.

Shephard (1970, p. vii) has the following statement about production functions: "… the central topic [of production functions] being an understanding of the possibilities of substitution between factors of production to achieve a given output." Accordingly, I will focus on the isoquants in the factor space for efficient production functions. It is reasonable to assume that the production function $f(.)$ for intended output y has the traditional properties with positive marginal productivities that are decreasing, as is standard in textbooks on production functions. This results in substitution possibilities between material inputs x_M and service inputs x_S. The marginal rate of substitution for the production of the intended output is:

$$dx_S / dx_M = -\frac{f'_{x_M}(x_M, x_S)}{f'_{x_S}(x_M, x_S)} < 0 \Rightarrow dx_M = -\frac{f'_{x_S}(x_M, x_S)}{f'_{x_M}(x_M, x_S)} dx_S . \quad (9)$$

Increasing the service input on the frontier function isoquant keeping the intended output constant will reduce the use of the material input. Increasing a service input like labour results in more efficient use of raw materials thus needing less of them.[27]

As explained earlier the production function $g(.)$ for the unintended output is only a function of the materials inputs. The marginal productivity for the material input is positive and the production functions maybe exhibiting constant or decreasing returns (cf. Panel 1 and Panel (a) in Fig. 4 in the next Subsection 5.2). The latter two properties are empirical questions.

The substitution between material and service inputs for a given level of intended output results in a decrease in the unintended output, as seen from (9). As exhibited in Fig. 3, I draw only the part of isoquants within the substitution region. The blue curved lines are traditional textbook isoquants for the intended good y with typical curvature. The level of the intended output increases in the northeast direction. There is no trade-off between material and service input in the production function $g(x_M)$ for the unintended output z, so there are no traditional isoquants, just vertical lines from the materials input axis having the same level of the unintended output for all points on the

[27] Cf. the chocolate example in Frisch (1935) (retold in Førsund 1999) of *ex post* substitution where more labour reduced the waste of chocolate production by picking out rejects and returning the chocolate mass back to the process.

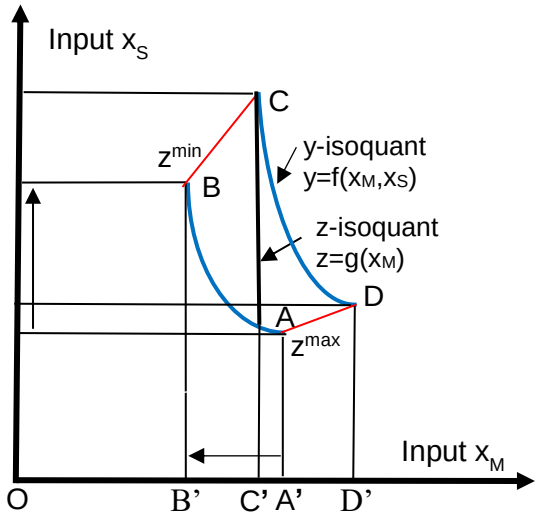


Figure 3. *Isoquants for the production of y and max-min values of z*

same line, as on the vertical line CC'. The thick vertical line from C to the intersection point on the BA isoquant intersects with all *y* isoquants moving from isoquant CD to isoquant BA, decreasing the level of the service input keeping constant material input.[28]

Obviously, dealing with material inputs there must be limitations on the substitution possibilities. It follows from the materials balance that the possibility for substitution between the material inputs x_M and the service inputs x_S as shown in Eq. (9) must be limited for a given level of the intended output. This means that the length of the isoquants may be rather short compared with textbook illustrations, where isoquants often cover the entire first quadrant. I have tried to capture this by setting limits for intended output isoquants by the levels z^{min} and z^{max} for the unintended output.[29] By definition, if we consider points B and C in Fig. 3, the intended output isoquants must be vertical at these points; the partial derivative of the service input is then zero: $f'_{x_S} = 0$. It is not possible to produce more intended output by increasing the service input. At the other end of the isoquants the partial derivatives of the material input is zero, $f'_{x_M} = 0$ and the isoquants are

[28] Frisch (1965, Fig. (14b.2), p. 272) points out that if the isoquants are separable then it is possible to choose producing more of one output than the other by changing the input mix. This situation is illustrated in Fig. 3 for the general case of input substitution when producing a fixed level of *y*. Holding the level of intended output fixed the substitution of service input x_S for material input x_M moving from A to B implies a reduction in *z* for a constant *y*.

[29] $\bar{z}$ in (8), defined as the mass in the material input, is greater than z^{max} in Fig. 3 if the intended output requires mass.

horizontal at these points.[30] It is not possible to produce more intended output by increasing the material input. The "min" and "max" values of the unintended output delimits the substitution region of the isoquants of the intended output as indicated by the straight red lines.[31] The limits are not dictated by the intended or unintended outputs as such; the outputs *y* and *z* are independent of each other, but both are determined by the inputs that are chosen. I have assumed in Fig. 3 that the length of isoquants increase with the amount of material inputs. This seems to be reasonable given the signing of derivatives, but is not essential for my story.

Using the notation z^{min} should not be misunderstood to mean that this is the minimum of the unintended output for all realisations of the amount of the intended output. The amounts z^{min} and z^{max} give the individual range of the unintended output for each isoquant for the intended output. Remember that it is assumed that we are at the frontier function of both types of outputs, any *z*-value at a point on a *y* isoquant is the minimum value for the chosen amounts of inputs and output. All points on vertical lines for the material input lying between these limits shown in Fig. 3, exhibit a minimum amount of the unintended output generated by using the combinations of inputs within the substitution region of the intended output. It is not of economic interest to consider points outside the substitution region. Without any regulation of the generation of residuals, profit maximisation or cost minimisation are solely based on determining the intended output and the two inputs (in the cost minimisation case only the level of inputs needs to be determined).

Let us start at point A with $f'_{x_M} = 0$ in Fig. 3. The efficient amount of the unintended output (i.e. the minimum of *z* for the level of inputs at A) is given by the z^{max} level at this point. Moving to point B along the intended output isoquant utilising the substitution possibilities, the use of material input decreases from A' to the smallest possible level at B' with $f'_{x_S} = 0$ at B. The service input has increased considerably more to realise the minimal generation of the unintended output while keeping the level of the intended output constant (see the two arrows indicating the changes along the axes). Point B has the minimal amount of the unintended output for the given level of

[30] In Frisch (1965, p.272) isoquants are exhibited as continuous contour curves as we have in a map of a mountain with a distinct maximum point. However, free disposability of the intended output implies that the isoquants are vertical continuing above a point like B and horizontal from point A.

[31] Since the detailed shape of the borderlines of the substitution region does not really matter in our context within the limited window of isoquants shown in Fig. 3, for simplicity I have chosen the lines to be linear.

the intended output. All levels of the unintended output along the isoquant for the intended output are minimal for the varying mix of inputs.

Point D (with $f'_{x_M} = 0$) exhibits a larger z^{max} than point A and a higher level of the service input x_S (it seems reasonable when the intended output increases to increase both inputs). The isoquant ends at point C (with $f'_{x_S} = 0$) that has a larger z^{min} than at point B. The vertical line CC' represents the same minimum value of *z* at all points on the thick line. The thick part of the line is within the substitution region for the intended output.

The material input is essential in the production functions in (7): zero material inputs imply zero production both of the indented output and the unintended one:

$$f(x_M, x_S) = g(x_M) = 0 \text{ for } x_M = 0 \text{ and } x_S > 0 \tag{10}$$

However, I am not trying to exhibit this in Fig. 3.

The situation in the output space can be illustrated in Fig. 4 using the points exhibited in the factor

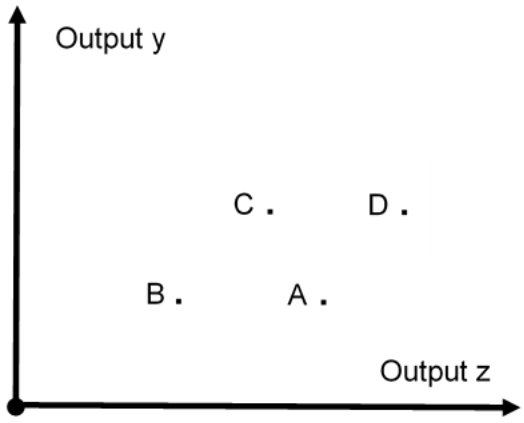


Figure 4. *Points in output space corresponding to the points in Fig. 2*

space in Fig. 3. All four points have different levels and mix of inputs. As can also be seen in Fig.3 points A and B have the same level of intended output *y*, and C and D have the same higher level of the intended output. As to the levels of the unintended output *z* all points have different levels, starting with the lowest level at B and then successively increasing levels at C, A and D. The consequence of having technical jointness as the type of joint production implies that there are just

output *points* in the output space. To make connection lines between the output points in the form of isoquants or trade-off curves for given inputs is not possible. According to our special variant of technical jointness, inputs have to change to generate different level of outputs.[32] The trade-off curve in Fig. 1 due to Shephard (1970) copied in so many papers show changing levels of both intended and unintended outputs in the same direction for given inputs.

5.2 The by-production model

Starting out with a single transformation relation similar to the Baumol and Oates (1988) model in Subsection 4.2, using the implicit function theorem it is stated in Murty et al (2012, p.120) that there seems to be some inconsistencies concerning the relationship between z and y, and between z and x_M. This correspond to the discussion of Baumol and Oates (1988) in Subsection 4.2, Eq. (6).

Murty et al (2012)[33] then introduced a model with separate production possibility sets for intended outputs and unintended outputs and called it the *by-production approach*.[34] In the most simple case for the intended output they operate with one transformation relation involving two inputs; the non-material input x_S and the material input x_M (using my notation), and two outputs; the intended (traded) output y and an intended abatement output y^a for internal use. The second relation is for the generation of the *net* pollutant z using two inputs; the material input x_M and the abatement output y^a from the first process. The functional representation is:

$$\begin{aligned} &f(x_S, x_M, y, y^a) = 0, \;\; f_i'(x_S, x_M, y, y^a) \le 0, \; i = S, M \\ &f_y'(x_S, x_M, y, y^a) \ge 0, \;\; f_a'(x_S, x_M, y, y^a) \ge 0 \\ &z = g(x_M, y^a), \;\; g_{x_M}'(x_M, y^a) > 0, \; g_a'(x_M, y^a) < 0 \end{aligned} \tag{11}$$

We notice that the value of the $f(.)$ function is independent of the level of z, and that the $g(.)$ function is independent of the level of y.

[32] Technical jointness means that increasing the intended output from the intersection point of CC' and isoquant BA is impossible without increasing input x_S keeping input x_M fixed. It is impossible to reallocate a bundle of resources that is fixed to the different product. If one could, then we have assorted production.

[33] The theoretical part of this paper was originally published as a working paper (Murty and Russell 2002).

[34] As pointed out in Førsund (2018a), by-production is in general economic literature, starting with classical economics on joint production (see Section 3), used for outputs sold on markets, but bringing in modest revenue compared with main products.

In Murty and Russell (2018), the same type of model is used (abatement output y^a is called a in the Murty and Russell (2018) paper, I keep the notation y^a here). An illustration of the connections between inputs and outputs in the case of two outputs and a single input is provided in Fig. 5.[35]

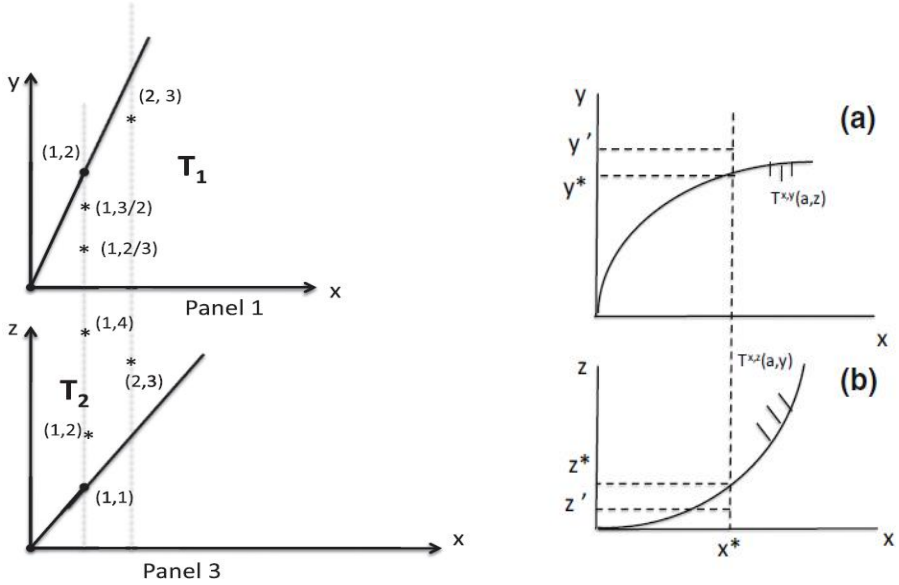


Figure 5. *Intended production sets (T_1 and (a)) and unintended ones (T_2 and (b))*
Source: Murty et al (2012, Fig. 1, p. 126 Panel 1 and 3); Murty and Russell (2018, Fig. 2, p. 13)

Panel 1 and 3 to the left show the production possibility sets T_1 and T_2 with CRS borders. The point (1, 2) (x=1, y=2) is on the border of the set T_1, and point (1, 1) (x=1, z=1) is on the border of set T_2. Panel (a) to the right shows for the intended output y a decreasing returns to scale production function when the feasible set of the abatement level y^a and unintended output z are held fixed. The level of the intended output y^* is on the border of its production possibility set $T^{x,y}(y^a, z)$. Panel (b) with increasing returns shows the relation between emission-causing input x and emissions level of the unintended output z when abatement level and intended output are held fixed. The point z^* is on the border of its feasible production possibility set $T^{x,z}(y^a, y)$.[36]

Radial efficiency measures are calculated for each type of function in Murty and Russell (2018),

[35] Notice that it is the shapes of production functions that are shown in Fig. 5 and not isoquants in the input space, as in Fig. 3 in Subsection 5.1.
[36] The points y' and z' appear outside their respective sets and are thus not feasible. The points belong to a discussion of what (x, y^a) can produce of y and z.

but emphasis is also put on a unified efficiency score based on the intersection of the sets T_1 and T_2, or (a) and (b) in Fig. 5.

The type of the first relation $f(.) = 0$ in (11) may be a case of assorted production;[37] the resources can be reallocated to the two types of intended outputs; output for sale and abatement output for internal use (see Murty et al (2012, Fig. 2, Panel 1, p. 134) for a confirmation of the assumption of assorted production). Assuming a single raw material used for both intended outputs the residual generation is the same. (However, it may be more realistic that the two outputs are using different raw materials. Then two equations for unintended outputs have to be specified.) When this is the case, it should be stated if the generation of residuals is the same per unit of the two outputs or different. The production function in the third line of (11) for the residual z is influenced only by material input x_M and abatement output, now in the role as input. In the transformation function for the intended outputs there is a substitution possibility for inputs due to the assumption about derivatives in the first line of Eq. (11).

Dakpo et al (2016) review the by-production approach. The approach is implemented empirically in Dakpo et al (2017); (2018); Arjomandia et al (2018) (however, all papers without specifying abatement).

Murty and Nagpal (2019) present a comprehensive review of the by-product model, and apply this model to an empirical study of Indian electricity coal-fired electricity producers. They have key critical remarks about Shephard-inspired technologies. However, I miss an explicit mentioning of joint production types as given in Section 3.

The by-production model is quite close to the factorially determined multi-output model in Subsection 5.1 on the splitting into two types of production functions. However, the factorially determined multi-output model can easily be extended to multiple outputs due to the separability property. It is not clear how the by-product model should be extended to multiple outputs of both types of output.

However, the papers Murty et al (2012); Murty (2015); Murty and Russell (2017); (2018) have rigorous mathematical treatment of assumptions or axioms, and in the last three papers prove

[37] In Walheer (2018), such a mix of type of production relations seems to be assumed.

theorems about necessary and sufficient requirements for an emission-generating technology to be of the by-production type. Thus, the analyses give more general results and insights than the specification in Subsection 5.1.

5.3 Abatement

There are two main possibilities for how to abate, the first one being internal technical changes not regarded as major changes, as discussed in Porter and van den Linde (1995). Some measures are short-run measures like improved process control, small-scale re-engineering, introducing more internal recycling of waste, etc. All such measures lead to improved efficiency of utilising material inputs and thereby reducing pollutants (Førsund (2018a,b,c). The second possibility is to introduce end-of-pipe technologies most popular in environmental economics (see e.g. Førsund (2009, pp. 28-30); (2018a, pp. 80-82); (2018b, pp. 58-61); (2018c, pp. 299-300)).[38] Although the first possibility may be the most used one in practice, it is usually very difficult to get data for internal abatement activities of the types mentioned. Short-run changes in technology may be mistaken for more long-run changes, and allocation of inputs such as labour on activities may not be recorded or even not be possible to distinguish. On the other hand, there are better possibilities to get data for end-of-pipe abatement due to the distinct separation of activities. However, end-of-pipe is not necessarily a unit separated from the main production equipment. I regard e.g. scrubbers and electrostatic filters on smokestacks as end-of-pipe because primary pollutants are inputs in these processes, and capital equipment and inputs like lime or chemicals do not interfere with the production of intended outputs, or play any role in that production. End-of-pipe abatement transforms varying shares of primary pollutants into usually harmless residuals and sometimes to by-products that have market value (Porter and van den Linde 1995).

Polluting firms often have capital equipment with embodied technologies, When reducing environmental pollution became a policy priority in the early 70ties, adding end-of-pipe equipment was seen as a more realistic and economic alternative for existing firms than requiring development of new equipment reducing waste. However, in the long run technology changes focussing on

[38] In the literature one can find that the first possibility is termed prevention and the second treatment, see e.g. Jaraite-Kazukauske et al (2014); Bostian et al (2016).

reducing generation of pollutants (i.e. prevention) would often be the most effective and the most economic measure.

Internal measures typically change technology. However, a popular measure due to regulation imposed by policy makers not changing the technology is to substitute cleaner inputs for more polluting ones, like using lighter oil for heating purposes, and using natural gas instead of coal in electricity generation.

The nature of the abatement is rather hidden in Murty et al (2012); Murty and Russell (2018). What I have called internal abatement (called prevention in the literature) is not mentioned, but the role of y^a in (11) appears as prevention. As far as I know the internal type of abatement in the by-production model has not been implemented empirically in the literature. The papers mentioned above have relevant examples of end-of-pipe abatement, but it is not easy to see that the formulation in (11) of two types of production functions can be turned into three separate equations as required introducing end-of-pipe abatement proper (Førsund 2018a,b,c).

In the environmental economics literature substitution between inputs as mentioned previously and end-of-pipe are the typical abatement options modelled. The latter option distinguishes between primary and secondary pollutants (or uncontrolled and controlled pollutants as used by EPA, or gross and net used in Murty and Russell 2018). In end-of-pipe abatement, primary pollutants are used as inputs. This feature seems to be absent in the abatement specification in Murty et al (2012); Murty and Russell (2017); (2018).

6. The critique of the Shephard-inspired literature

I use the Shephard Fig. 1 with two outputs and one input as the departure for my critique. Two restrictions are put on the technology; weak disposability and null jointness. The latter restriction drives the shape of the trade-off curves in the figure. The trade-off curves must start at the origin, thus giving the positive slopes of the efficient output isoquants (thick-line segments in Fig. 1). However, these isoquants are the border of the output sets and by definition efficient, and efficiency is based on producing maximal quantity of the intended output. The shape of the isoquants is in direct conflict both with the efficiency requirement and with the materials balance.

Taking Fig. 1 at face value goes against the materials balance assuming that the single input is material; input is constant along each output isoquant in Fig.1. It is not possible to reduce both intended and unintended output keeping inputs constant, containing a certain fixed amount of materials. This is obvious if the inputs are fully used in the production of the two outputs producing less of both for a given amount of inputs, moving from the right to the left along the trade-off curves.

The property of null jointness between outputs in Shephard and Färe (1974, p. 80) is introduced as a definition, and it is difficult to see the basis in a real-life joint production. To claim null jointness between the intended and unintended outputs does not reflect the basic relationships of technical jointness; the point is that each output will be zero simultaneously if the material inputs are zero, as stated in Eq. (10). Furthermore, null jointness between y and z as portrayed by the output isoquants in Fig. 1 definitely breaks with the materials balance having positive input at the origin; it makes no sense to have $y = z = 0$ with $x > 0$.

It does not help to assume that part of the inputs are used to abate the unintended output.[39] This proposal cannot be taken seriously when there is no abatement activity modelled. You cannot draw curves assuming a given level of input along the curve and then say that the input is actually reduced when moving along the trade-off curve. This is not in accordance with the basic definition of an isoquant. The abatement process must be explicitly modelled. If inputs are reallocated to abatement, then the input cannot be constant along the output isoquants of the production possibility sets. If it is the case that some of the inputs are actually reallocated this does not show up in Fig.1. In order to satisfy the definition of isoquants as based on keeping the input level constant, a part of input cannot at the same time be removed.

Weak disposability does not appear as a technical restriction in an engineering sense concerning the shape of the border of the output sets. It just tells us that reducing the quantities of an output point on the borders or in the interior of the set proportionally with a factor in the interval [0, 1], then the new point also belongs to the output set. We are only interested in the efficient points on the border, and it is clear from Fig.1 that going from the right to the left along the border the change

[39] In Färe and Grosskopf (1983 p. 1071) it is stated: "… If a reduction in emissions is desired, one could also divert some of the constant input vector to the "clean-up" of those emissions, which implies that less input would be available for the production of electricity, resulting in a simultaneous decline in good and bad output."

is not proportional, as stated in several papers. The ratio between the outputs change continuously along each frontier segments in Fig. 1.

Although joint production is discussed in Shephard (1970, Chapter 9.5, pp. 212-220) there is no discussion of the implication of having a simultaneous production of intended and unintended outputs for how to model joint production. Neither the concept of assorted production nor the concept of technical jointness are used. Introducing weak disposability as in (4) takes care of the problem with assorted production, but this is done without commenting on the existence of this form of joint production. The connection between weak disposability and joint production is not explained.

As stressed in Section 3 the generation of intended and unintended outputs takes place simultaneously. There is only a single common process. Unintended residuals cannot be generated in separate processes from intended outputs per definition. As illustrated in Fig. 3 and 4 in Subsection 5.1 the very nature of unintended production implies that efficient utilisation of inputs to produce given levels of intended outputs will *unavoidably* generate positive minimums of unintended outputs.[40]

Fig. 1 presenting the figure in Shephard (1970, p. 188) has been reproduced in one form or another in almost all papers using the Shephard-inspired model. This type of figure postulates a positively sloped connecting curve, or a trade-off curve, between the intended and the unintended outputs in the case of one of each for given inputs. As demonstrated in Subsection 5.1, this is impossible taking technical jointness seriously.

By the nature of technical jointness and the thermodynamic laws there will be a positive minimum of residuals generated on the frontier for given inputs and given the applied production technology. There is no such minimum formulated in Shephard-inspired literature as far as I know. The Shephard-inspired literature on intended and unintended outputs all use a trade-off isoquant between the two types of outputs for given inputs. However, this is not possible given that the joint production is of the type technical jointness.

[40] As shown in Fig. 3 to realise a specific level of intended output the point on the corresponding isoquant of the frontier function $f(.)$ implies that the level of the material input, and thereby the unintended output, is the minimal for the situation.

The use of directional distance functions (Chung et al 1997) is also based on a trade-off isoquant between intended and unintended outputs. Therefore, this approach has all the weaknesses of the Shephard-inspired models. In addition, the assumption that the frontier point is found by adding/subtracting values using the same scalar factor multiplied with the observed values of both types of outputs, constrains the calculation of efficiency and productivity in a way that is difficult to accept as giving valid measures. In the single equation approach using distance functions the argument for this special treatment is based on giving 'credit' for intended outputs and 'penalise' unintended output.[41] Lastly, the choice of direction influences the results, and this seems rather arbitrary, especially when using efficiency scores for productivity measures like the Malmquist productivity index with varying choice of directions for each period (Chung et al 1997).

The Shephard-inspired models using the distance function being a single equation imposes a straitjacket on the estimation of efficiency and productivity. In Murty and Russell (2017, p. 12) it is stated: "…a single functional relation is not sufficient to capture all the complex trade-off among inputs and outputs involved in the production of economic outputs and the generation of emissions." However, the problems with null jointness and weak disposability used in the Shephard-inspired single-equation models all disappear when introducing two types of production functions, one for each type of output, as shown in Section 5.

7. Measuring inefficiency in a nonparametric multi-equation model

7.1 Defining inefficiency

The efficiency literature is in general focussed on measuring efficiency. However, the causes of inefficiency are rarely researched (see e.g. Førsund (2010) for a review of reasons for inefficiency). Inefficiency arises in general when the potential engineering or blueprint technology, the frontier for short, is not achieved when transforming inputs into outputs, assuming that this is feasible.[42]

[41] In Färe et al (1989, p. 90) it is stated: "When evaluating the performance of producers it makes sense to credit them for their provision of desirable outputs and penalize them for their provision of undesirable outputs."

[42] In the case of the presence of embodied technology or vintage capital, a distinction should be made between efficient utilisation of the mix of existing technologies and the efficiency of the most modern technology available (Førsund 2010).

For given desirable outputs too many resources of raw materials and service inputs are used. For a given amount of inputs containing physical mass, it means that at the frontier more outputs could have been produced. In terms of the materials balance (1) the implication is that the amount of residuals z for constant inputs x_M at inefficient operation will be reduced if the frontier is achieved. Inefficiency in the use of service inputs means that with better organisation of the activities more output could be produced if the frontier is realised for constant x_S.

The materials balance also holds for inefficient observations (as pointed out in Section 2). It is the amount of residuals and outputs that have potentials for change, while the a, b, c coefficients and the variables in Eq. (1) remain the same. The combustion process may be run less efficient in converting the raw material into heat, and a different mix of combustion substances may be produced than at efficient operation. In thermal electricity production based on coal, the mix of substances such as CO_2, CO, particles, NO_x and ash may differ between inefficient and efficient operations. Another source of inefficiency is the occurrence of rejects of intended outputs and unnecessary waste of raw materials, e.g., producing tables of wood, residuals consist of pieces of wood of different sizes from rejects and down to chips and sawdust. The ways of improving the use of raw materials and thereby reducing the amount of residuals are more or less of the same nature as factors explaining substitution possibilities between material and service inputs in Subsection 5.1. However, inefficient use of service inputs (labour and capital) should not be confused with substitution between labour and raw materials on a frontier isoquant for intended output as shown in Fig. 3.

There is another type of problem within the efficiency strand of research not often mentioned concerning the behaviour of (or the management of) firms. It is difficult to assume, as in standard production theory using frontier functions only, that inefficient firms can optimise in the usual sense of obtaining maximal profit or minimising costs. It is very seldom that production functions are formulated for inefficient firms in non-parametric analyses. Introducing behaviour in non-parametric DEA models for a unit it is necessary to assume that frontier technology is used if there are no known obstacles for being efficient. If firms do know the frontier, why do they end up being inefficient? To appeal to randomness only is not so satisfying.

However, in the real world all firms, also inefficient ones, have to react to e.g. environmental regulation. When efficiency is estimated the observations are usually taken as given and no

behavioural action on the part of the units is assumed to take place. It is the analyst that creates an optimisation problem when calculating efficiency measures. This may be a reason for the lack of pursuing policy instruments in the literature addressing efficiency when both desirable and undesirable outputs are produced. In the environmental economics literature not addressing efficiency issues, the design of policy instruments, playing on giving firms incentives to change behaviour as to emitting pollutants, is of paramount interest. However, the assumptions in the inefficiency literature based on Shephard (1970) in Subsection 4.1 are made for measuring efficiency, and may not be suitable for developing policy instruments applied to all units in an industry. We saw this in Färe et al (1986) making introduction of regulation of emissions change the form of the production possibility set for all units, and not addressing the reactions of each individual unit to the regulation. If economic behaviour is applied in the efficiency literature, then the unit in question typically operates on the frontier.

7.2 Efficiency and productivity measures

Efficiency measures

Concerning the estimation of the unknown frontiers a non-parametric DEA model, build up as a polyhedral set, assuming standard axioms such as compactness, convexity and monotonicity, can be applied to estimate the efficiency measures based on the estimate of the best practice frontier that the data at hand can give us. The technical jointness characteristic of producing simultaneously intended and unintended outputs has been satisfied by splitting the production function into two separate frontier functions as in (7). However, the DEA models look “normal”; we cannot see that the separate technologies satisfies technical jointness but for the imposed curvature of the residuals function. We use the observations of inputs and outputs to estimate two radial output-oriented efficiency measures. Therefore, the observations of each unit have a unit sub-index as in the standard DEA model. This does not mean that the total use of inputs is a summation over units.

It is standard to estimate the border of the intended output set and find the projection points for inefficient units, calculating an output-oriented efficiency measure. As we see from Fig. 3 and Fig.

5(a) the production possibility sets for the intended output are assumed to be convex and the borders concave.

The efficiency measure for the desirable output is $E_y = (y^{obs} / y^*) \in (0,1]$. The index "*obs*" indicates the observation of the intended output, and "*" indicate the maximal output for the given inputs in the frontier function in (7); $y^* = f(x_M^{obs}, x_S^{obs})$. The inputs are observations. In the non-parametric case, the following LP optimising problem is set up for finding the efficiency score for unit *i* belonging to a set of *N* units:

$$
\begin{aligned}
& 1/E_{y_i} = Max_{\lambda,\theta_i}\ \theta_i \\
& s.t. \\
& \sum_{j=1}^{N} \lambda_j y_j \geq \theta_i y_i \\
& \sum_{j=1}^{N} \lambda_j x_{kj} \leq x_{ki},\ k \in M, S \\
& \sum_{j=1}^{N} \lambda_j = 1, \lambda_j \geq 0, \theta_i \geq 1
\end{aligned}
\qquad (12)
$$

Variable returns to scale (VRS) is specified. Solving (12) for λ_j and the score θ_i we find y_i* as $y_i^* = \theta_i y_i^{obs}$.

The efficiency measure for the undesirable output is $E_z = (z^*/z^{obs}) \in (0,1]$, $z^* = g(x_M)$. Here z^* is the minimal amount of pollutants for given input x_M. To call such a measure for 'environmental efficiency' or "eco-efficiency" as done in the literature occurs to me as somewhat misplaced; within our production model we do not know anything about what happens in the environment when emitting residuals, and we do not know the consumers' evaluation of the degradation of environmental qualities.[43] However, what we know is the amount of residuals discharged to the environment. In the literature based on the Shephard approach of a single function using DEA to calculate efficiency measures, both using the hyperbolic measure and using the directional distance

[43] The term environmental performance measure or index is used within business economics based on sustainability concerns for firms' production. The win-win theme of Porter is investigated correlating environmental performance indices and profit. For construction of the indices, see e.g. Dragomir (2018) for a review of 172 papers on environmental performance, and Esty and Cornelius (2002) going through a long list of measures for World Economic Forum.

function, the measure of technical efficiency and a measure termed environmental performance are linked together through a common parameter.[44]

The optimisation problem for unit *i* is:

$$
\begin{aligned}
&E_{z_i} = Min_{\lambda',\varphi_i}\ \varphi_i \\
&s.t. \\
&\sum_{j=1}^{N} \lambda'_j z_j \le \varphi_i z_i \\
&\sum_{j=1}^{N} \lambda'_j x_{Mj} \ge x_{Mi} \\
&\sum_{j=1}^{N} \lambda'_j = 1, \lambda'_j \ge 0, \varphi_i \ge 0
\end{aligned}
\tag{13}
$$

Solving for λ_i' and φ_i we find $z_i^* = \varphi_i z_i^{obs}$.

The border of the production possibility set for the unintended output in Fig. 5(b) is convex. The inequalities in (13) are then opposite to the similar inequalities in (12). The efficient points spanning the frontier faces all have the same material input quantities in the two problems, but the frontier for the unintended output does not have service inputs.

The materials balance is valid for all observations including inefficient ones, but projection points are not observations. The question is how to check if these points satisfy the materials balance. We have to combine y_i* and z_i* obtained from solving different programming problems, and the weights λ_j in (12) and λ_j' in (13) may be different.

Let us assume that we have an inefficient unit *i* with a projection point being on an efficient facet. In the case of the production variables being single, the materials balance of the projection point should be:

$$ax_{Mi} = b\theta_i y_i + c\varphi_i z_i \tag{14}$$

[44] Using the two-equation models in Section 5 ensures independent assessment of technical efficiency for the intended output and an "environmental" assessment for the unintended output.

The expansion of y_i ($\theta_i \geq 1$) must be counteracted by the reduction in z_i ($0 \leq \varphi_i \leq 1$) for the materials balance to hold. However, there is no guarantee that the materials balance is obeyed. The projection point on the frontier segment for an inefficient unit is the weighted average of the two frontier units spanning the segment using their intensity weights λ'. In view of the materials balance holding at a very micro level, and that it has to be complete as to substances, I drop this route. Other considerations are that the true production function will typically be more efficient than the frontier estimated using DEA, and probably not have the piecewise linear form of a DEA frontier. Therefore, the true materials balance may not be the same as the balance based on the estimates of faces of the frontier function. An output- oriented projection point for the intended output is a weighted point with the λ-weights and output levels of the relevant efficient points, and the same is the case for the unintended output using λ'-weights. However, the observed material input is the same in both problems so any bias in the efficiency scores may not be so problematic. The service input does not appear in (13). The question is if this can affect estimates of efficiency scores. However, as shown in Figs. 4 and 5 a choice of factors considering the production of the intended output gives specific values of the material input. A simple test is to check the discrepancies between the left- and right-hand sides of (14).

Productivity measures

It is not only of interest to estimate efficiency measures for the two types of outputs, but also to measure the productivity change of them. Suppressing the unit index for convenience the efficiency measures can straightforwardly be converted to separate standard Malmquist productivity change indexes for each output using discrete time periods t:

$$M_y^{t,t+1} = \frac{E_y^{t+1}}{E_y^t} > 1 \text{ progress, } < 1 \text{ decline}$$
$$M_z^{t,t+1} = \frac{E_z^t}{E_z^{t+1}} > 1 \text{ progress, } < 1 \text{ decline} \qquad (15)$$

With standard, I mean that the efficiency scores are calculated relative to a benchmark frontier based on constant returns to scale, making output orientation equal to input orientation of efficiency scores (see Førsund 2016). Calculating the productivity change index for the unintended output z the time indices for the efficiency scores are simply switched. A decrease in the

unintended output is regarded as productivity progress. Regarding policy use of efficiency and productivity results separate measures for intended and unintended outputs seem to yield the most interesting information.

8. Conclusions

Models that we make for calculating efficiency measures for production activities are quite aggregated compared to the engineering level of real life. In a study of efficiency in metal machining industry, Kurz and Manne (1963) identify 129 separate production functions for basic activities. It is of paramount importance that the much simpler models we make capture essential features of production activities we analyse (Frisch 2010). However, the Shephard-inspired efficiency models involving intended and unintended outputs are too simple, based on a single equation in the form of a distance function as a function of all output and input variables. The main problem is that the importance of the type of joint production faced by having intended outputs produced at the same time as generating unintended ones, is not sufficiently understood. The type of joint production must be such that unintended outputs are impossible to avoid producing. The Frisch (1965) categories of technical jointness and extreme jointness of outputs imply that both types of outputs are generated by the same inputs simultaneously in the same activity. In this paper, the technical jointness is assumed (less strict than extreme jointness including the Leontief type of models), opening up for change in the input mix and levels to generate different mix of the type of outputs. However, in output space, this implies that there is no trade-off isoquant between intended and unintended output for given inputs; there are just points in the output space generated by a different mix and different levels of inputs.

An important assumption is that one or more of the inputs must be material. The materials balance (Ayres and Kneese 1969) tells us that matter contained in inputs cannot disappear, but will be contained in the intended outputs or discharged to the environment as waste or residuals. These residuals are pollutants if causing environmental problems and that there is a willingness to pay to reduce the amounts. The two thermodynamic laws ensure that intended outputs cannot utilise all mass; some positive amount of unintended waste will always occur. If we assume that intended and unintended outputs compete for the material inputs, then efficient production of the intended

outputs imply that there is a minimum of mass ending up in the unintended outputs. Furthermore, this minimum amount implies that there cannot be, with inputs given, any trade-off isoquant between intended and unintended outputs when production of the intended output is efficient. The null jointness assumption of the intended output and the unintended one results in positive slopes of output isoquants for given inputs. However, this goes against the material balance.

Shephard-inspired models have become very popular judged by the citations. However, the type of model I have developed in Subsection 5.1 takes explicitly the type of joint production into consideration and has no problem obeying theoretically the materials balance.

The model used in Subsection 5.1 has been the simplest one with two types of outputs and two types of inputs. Introducing several variables of both types should be explored. Extending the model (7) by entering more single equations for both types following the scheme of factorially determined multi-output functions is one possibility. Another development may be a combination of factorially determined multi-output functions satisfying technical jointness and assorted production. However, assuming a trade-off between residuals may be due to inefficiencies not existing for the efficient frontiers.

After revealing the problems with the Shephard-inspired approaches, I hope the tide may turn and a rise in the use of the alternative models will come.

References

Ambec S, Coheny MA, Elgiez S, and Lanoie P (2013) The Porter hypothesis at 20: can environmental regulation enhance innovation and competitiveness? *Review of Environmental Economics and Policy* 7(1): 2-22

Arjomandia A, Dakpo HK, and Seufert JH (2018) Have Asian airlines caught up with European airlines? A by-production efficiency analysis. *Transportation Research Part A* 116(October): 389-403

Ayres RU and Kneese AV (1969) Production, consumption and externalities. *American Economic Review* 59(7): 282-297

Baumgärtner S and de Swaan Arons J (2003) Necessity and inefficiency in the generation of waste: a thermodynamic analysis. *Journal of Industrial Ecology* 7(2): 113-123

Baumol WJ and Oates W (1988) *The theory of environmental policy*. Cambridge University Press, Cambridge, second edition. (First edition 1975. Prentice Hall Inc., New Jersey)

Bostian M, Färe R, Shawna Grosskopf S and Lundgren T (2016) Environmental investment and firm performance: a network approach. *Energy Economics* 57(June): 243-255

Brännlund R and Lundgren T (2009) Environmental policy without cost? A review of the Porter hypothesis. *International Review of Environmental and Resource Economics* 3(1): 75-117

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

Coelli TJ, Rao DSP, O'Donnell JO and Battese GE (2005) *An introduction to efficiency and productivity analysis. Second edition*. Springer Science+Business Media, Inc., New York

Cooper WW, Seiford LM and Tone K (2007) *Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software*. *Second edition*. Springer Science+Business Media, Inc., New York

Dakpo KH, Jeanneaux P, and Latruffe L (2016) Modelling pollution-generating technologies in performance benchmarking: recent developments, limits and future. Prospects in the nonparametric framework. *European Journal of Operational Research* 250(2): 347-359. http://dx.doi.org/10.1016/j.ejor.2015.07.024.

Dakpo KH, Jeanneaux P, and Latruffe L (2017) Greenhouse gas emissions and efficiency in French sheep meat farming: a non-parametric framework of pollution adjusted technologies. *European Review of Agricultural Economics* 44(1): 33-65. http://dx.doi.org/ 10.1093/erae/ jbw013

Dragomir VD (2018) How do we measure corporate environmental performance? A critical review. *Journal of Cleaner Production* 196(20 September), 1124-1157

Ebert U and Welsch H (2007) Environmental emissions and production economics: implications of the materials balance. *American Journal of Agriculture Economics* 89(2): 287-293

Esty D and Cornelius PK (eds) (2002) *Environmental Performance Measurement: The Global Report 2001-2002*. The World Forum, Oxford University Press, Oxford

Färe R and Grosskopf S (1983) Measuring output efficiency. *European Journal of Operational Research* 13(2): 173-179

Färe R, Grosskopf S, and Pasurka C (1986) Effects on relative efficiency in electric power generation due to environmental controls. *Resources and Energy* 8(2): 167-184

Färe R, Grosskopf S and Pasurka C (2013) Joint production of good and bad outputs with a network application. In: Shogren J (ed) *Encyclopedia of energy, natural resources and environmental economics*. Vol 2, pp. 109-118. Elsevier, Amsterdam

Färe R, Grosskopf S, Lovell CAK and Pasurka C (1989). Multilateral productivity comparisons when some outputs are undesirable: a nonparametric approach. *Review of Economics and Statistics* 71(1): 90-98

Frisch R (1935) The principle of substitution. An example of its application in the chocolate industry. *Nordisk Tidskrift for Teknisk Økonomi* 1: 12-27

Frisch R (1965). *Theory of production.* D. Reidel, Dordrecht

Frisch R (2010) *A dynamic approach to economic theory. The Yale lectures by Ragnar Frisch, 1930*. In: Bjerkholt O and Qin D (eds). Routledge Studies in the History of Economics. Routledge, London and New York

Førsund FR (1972) Allocation in space and environmental pollution. *Swedish Journal of Economics* 74(1): 19-34

Førsund FR (1973) Externalities, environmental pollution and allocation in space: a general equilibrium approach. *Regional and Urban Economics* 3(1): 3-32

Førsund FR (1985) Input-output models, national economic models, and the environment. In: Kneese AV and Sweeney JL (eds). *Handbook of natural resource and energy economics, vol.* I, Chapter 8, pp. 325-341. Elsevier Science Publishers BV, Amsterdam

Førsund FR (1998) Pollution modelling and multiple-output production theory. *Discussion Paper* # D-37/1998, Department of Economics and Social sciences, Agricultural University of Norway (also republished as *Memorandum* No 10/2016 Department of Economics University of Oslo found on the internet)

Førsund FR (1999) On the contribution of Ragnar Frisch to production theory. *Rivista Internazionale di Scienze Economiche e Commerciali (International Review of Economics and Business)* 46(1): 1-34

Førsund FR (2009) Good modelling of bad outputs: pollution and multiple-output production. *International Review of Environmental and Resource Economics* 3(1): 1-38

Førsund FR (2010) Dynamic efficiency measurement. *Indian Economic Review* 45(2), 125-159. Also published as Chapter 4 (pp. 187-219) in Ray SC, Kumbhakar SC, Dua P (eds) (2015) *Benchmarking for performance evaluation. A frontier production approach*. DOI 10.1007/978-81-322-2253-8_4. Springer (India), New Delhi- Heidelberg-New York-Dordrecht-London

Førsund FR (2016) Productivity interpretations of the Farrell efficiency measures and the Malmquist index and its decomposition. In: *Advances in efficiency and productivity*, Aparicio J, Lovell CAK and Pastor JT (eds). Chapter 6, 121-147. Springer International Publishing AG, Cham

Førsund FR (2018a) Multi-equation modelling of desirable and undesirable outputs satisfying the materials balance. *Empirical Economics* 54(1): 67-99. DOI 10.1007/s00181-016-1219-9

Førsund FR (2018b) Pollution meets efficiency: multi-equation modelling of generation of pollution and related efficiency measures. In: Pang R, Lovell CAK, and Bai X (eds). *Energy, environment and transitional green growth in China*, Chapter 3, pp. 37-79. Springer Nature Pte Ltd, Singapore

Førsund FR (2018c) Productivity measurement and the environment. In: Grifell-Tatjé E, Lovell CAK, Sickles R (eds) *The Oxford handbook of productivity analysis*. Chapter 8, pp. 37-79. Oxford University Press, Oxford

Førsund FR and Strøm S (1974) Industrial structure, growth and residuals flows. In: J. Rothenberg and I. G. Heggie (eds). *The management of water quality and the environment*, Chapter 2, pp. 21-69. MacMillan, London

Førsund FR and Strøm S (1976) The generation of residuals flows in Norway: an input - output approach. *Journal of Environmental Economics and Management* 3: 129-141

Jevons WS (1883) *The theory of political economy* (first published 1871, third edition available on internet in The Online Library of Liberty). Macmillan, London

Jaraite-Kazukauske J, Kazukauskas A and Lundgren T (2014) Determinants of environmental expenditure and investment: evidence from Sweden. *Journal of Environmental Economics and Policy* 3(2): 148-166. DOI: 10.1080/21606544.2013.875948

Johansen L (1960) *A multi-sectoral study of economic growth*. North-Holland Publishing Company, Amsterdam

Kurz HD (1986). Classical and early neoclassical economists on joint production, *Metroeconomica* 38(1): 1-37

Kurz M and Manne A (1963) Engineering estimates of capital-labor substitution in metal machining. *American Economic Review* 53(4): 662-681

Leontief W (1970) Environmental repercussions and the economic structure: an input-output approach. *The Review of Economics and Statistics* 52(3): 262-271

Lanoie P, Laurent-Lucchetti J, Johnstone N, and Ambec S (2011) Environmental policy, innovation and performance: new insights on the Porter hypothesis. *Journal of Economics and Management Strategy* 20(3): 803-842

Leontief W and Ford D (1972) Air pollution and the economic structure: empirical results of input – output computations. In: Brody A and Carter A (eds). *Input – output techniques*, pp. 9-30. North-Holland, Amsterdam-London

Murty S (2015) On the properties of an emission-generating technology and its parametric representation. Economic Theory 60(2): 243-282

Murty S and Russell RR (2002) On modeling pollution-generating technologies. *Mimeo*, Department of Economics, University of California, Riverside (revised version July 2005).

Murty S and Russell RR (2017) Bad outputs. In: Ray SC, Chambers R, and Kumbhakar S (eds). *Handbook of Production Economics, Vol. 1 (Theory)*, Chapter 10, version May 2017. Forthcoming, first edition in 2021. Springer Nature, Heidelberg

Murty S and Russell RR (2018) Modeling emission-generating technologies: reconciliation of axiomatic and by-production approaches. *Empirical Economics* 54(1): 7-30. DOI 10.1007/s00181-016-1183-4

Murty S and Nagpal R (2019) Measuring output-based technical efficiency of Indian coal-based thermal power plants: a by-production approach. *Indian Growth and Development Review*. Vol. ahead-of-print, DOI: 10.1108/IGDR-05-2018-0058

Murty S, Russell RR and Levkoff SB (2012) On modeling pollution-generating technologies. *Journal of Environmental Economics and Management* 64(1): 117-135

Pethig R (2006) Non-linear production, abatement, pollution and materials balance reconsidered. *Journal of Environmental Economics and Management* 51(2): 185-204

Porter ME and van der Linde C (1995) Toward a new conception of the environment-competitiveness relationship. *Journal of Economic Perspectives* 9(4): 97-118

Russell RR (1998) Distance functions in consumer and producer theory. In: Färe R, Grosskopf S and Russell RR (eds). *Index numbers: essays in the honour of Sten Malmquist*. Chapter 1, pp.7-90. Kluwer Academic Publishers, Boston/London/Dordrecht

Shephard RW (1970) *Theory of Cost and Production Functions*. Princeton University Press, Princeton NJ

Shephard RW and Färe R (1974) The law of diminishing returns. *Zeitschrift für Nationalökonomie* 34(1-2): 69-90

Walheer B (2018) Output, input, and undesirable output interconnections in data envelopment analysis: convexity and returns-to-scale. *Annals of Operations Research*, First Online: 17 September 2018