MEMORANDUM

No 09/2017
September 2017

Pollution Meets Efficiency: Multi-equation modelling of generation of pollution and related efficiency measures*

Finn 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 08/2017	John K. Dagsvik Invariance Axioms and Functional Form Restrictions in Structural Models
No 07/2017	Eva Kløve and Halvor Mehlum Positive illusions and the temptation to borrow
No 06/17	Eva Kløve and Halvor Mehlum The Firm and the self-enforcing dynamics of crime and protection
No 05/17	Halvor Mehlum A polar confidence curve applied to Fieller's ratios
No 04/17	Erik Biørn Revisiting, from a Frischian point of view, the relationship between elasticities of intratemporal and intertemporal substitution
No 03/17	Jon Vislie Resource Extraction and Uncertain Tipping Points
No 02/17	Wiji Arulampalam, Michael P. Devereux and Federica Liberini Taxes and the Location of Targets
No 01/17	Erik Biørn Identification and Method of Moments Estimation in Polynomial Measurement Error Models
No 19/16	Erik Biørn *Panel data estimators and aggregation*
No 18/16	Olav Bjerkholt *Wassily Leontief and the discovery of the input-output approach*

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http://www.sv.uio.no/econ/english/research/unpublished-works/working-papers/

Pollution Meets Efficiency:
Multi-equation modelling of generation of pollution and related efficiency measures*

by

Finn R Førsund

Department of Economics, University of Oslo

Abstract The generation of unintended residuals in the production of 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. The materials balance tells us that residuals stem from the use of material inputs. The modelling of joint production must therefore reflect this. A multi-equation model building on the factorially determined multi-output model of classical production theory can theoretically satisfy the materials balance. Potentially complex technical relationships are simplified to express each of the intended outputs and the unintended residuals as functions of the same set of inputs. End-of-pipe abatement activity is introduced for a production unit. Introducing direct environmental regulation of the amount of pollutants generated an optimal private solution based on profit maximisation is derived. Serious problems with the single-equation models that have dominated the literature studying efficiency of production of intended and unintended outputs the last decades are revealed. An important result is that a functional trade-off between desirable and undesirable outputs for given resources, as exhibited by single-equation models, is not compatible with the materials balance and efficiency requirements on production relations. Multi-equation models without this functional trade-off should therefore replace single equation models. Extending the chosen multi-equation model to allow for inefficiency, three efficiency measures are introduced: desirable output efficiency, residuals efficiency, and abatement efficiency. All measures can be estimated separately using the non-parametric DEA model.

Keywords Materials balance; Joint production; Residuals generation; Single-equation and multi-equation models; End-of-pipe abatement; Efficiency measures; Data Envelopment Analysis (DEA)

JEL Classification C51, D24, D62, Q50

* The paper is based on a presentation at the 2016 Asia-Pacific Productivity Conference, Nankai University, Tianjin, China, 7-10 July, and building largely on the work in progress in Førsund (2017a,b). I am indebted to Kenneth Løvold Rødseth, Victor V. Podinovski and an anonymous reviewer for challenging and constructive comments improving the paper. The paper is forthcoming as Chapter 3 in the book *Energy, environment and transitional growth in China*, Pang R, Lovell CAK, and Bai X (eds.), Springer.

1 Introduction

Pointing out the importance of the materials balance principle Ayres and Kneese (1969); Kneese et al (1970) signalled the start of a new more realistic way of modelling the interaction between human activities of consumption and production and the discharge of residuals to the environment that can be polluting. The concept of (negative) externalities had been used before in the literature to analyse pollution. However, somewhat innocent examples like vibrations from a confectionary's machines disturbing a doctor having a consulting room next door (Coase 1959, p. 26), sparks from a locomotive causing forest fire, and smoke from a factory chimney dirtying washing hanging out to dry (Pigou 1920), were used. The materials balance principle underlined the pervasiveness of generation of residuals caused by using material resources and the unavoidability of their generation, invoking the thermodynamic laws. The same principle holds for energy inputs. Energy residuals are heat and noise. As to energy production like charcoal or electricity the second law of thermodynamics tells us that all energy contained in primary inputs cannot be fully utilised in the energy outputs due to the entropy created (Baumgärtner and de Swaan Arons 2003).

We now face threats of global warming due to emission of greenhouse cases, and increasing urban health problems mainly due to emissions from the transport sector, and problems due to increased acidity of lakes and oceans from burning fuel like coal and oil for thermal electricity production, and residential heating and cooking. The capacity of Nature to absorb emissions from human activities have long since been exhausted, and the exponential accumulation of some substances in the environment may result in the necessity of a drastic future cut in carbon-based energy use if global disasters are to be avoided. A necessary international cooperation to reduce the emission of global pollutants started with the Kyoto Protocol in 1997, and the Paris Agreement in 2016 is the last effort of the United Nations. To achieve results reliable modelling is needed on all levels of aggregation, also on the micro level studied in this paper.

The purpose of the paper is to develop a way to model the generation of residuals in production (or consumption) activities when producing intended outputs that complies with the materials balance. A distinction is made between an efficient production of desirable outputs for given resources and inefficient operations facilitating measuring both efficiency in producing desirable outputs and efficiency in generating residuals. The dominating single equation model in empirical studies comprising resources and two types of output; desirable and undesirable, is shown not to comply with the materials balance and efficiency properties of the production relations, both in the case of strong (free) disposability of outputs and inputs and weak disposability for desirable- and undesirable outputs together. It is demonstrated in the paper that separating production relations for desirable outputs and undesirable ones is in theoretical compliance with both the materials balance and efficiency of production relations.

The paper is organised as follows. Section 2 states the general model blocks of environmental economics limited to a static analysis, the definition of the materials balance, and provides a brief non-technical overview of recent developments concerning the joint generation of desirable and undesirable outputs in the case of inefficiency. Section 3 discusses the concept of joint production and the Frisch classification scheme. It is demonstrated that a single functional representation of frontier technology relying on a trade-off between desirable and undesirable outputs for given resources does not satisfy the materials balance. In Section 4 the multi-equation model based on a special case of multiple output production set out in Frisch (1965) satisfying the materials balance, is introduced and discussed. End-of-pipe abatement is introduced in Section 5, and the impact of regulating the emission of pollutants studied. Inefficiency is discussed in Section 6. The assumption of weak disposability that has dominated efficiency studies of joint desirable and undesirable outputs is scrutinised and found to violate the materials balance principle and efficiency assumptions of production relations. Section 7 introduces efficiency measures that can be estimated for a non-parametric multi-equation production model. Section 8 concludes.

2 Environmental economics

Concern about the environment has old roots in economics, as indicated in Section 1. We will focus on the modelling of relationships after the introduction of the materials balance principle in Ayres and Kneese (1969).

2.1 Environmental economics post externality models

The need for sound modelling of the interaction between human activities and Nature is obvious for the understanding of how to deal with the problems in a way that is most effective in utilising the trade-offs between man-made goods and the environmental qualities. Within the strand of research of environmental economics the main model elements to capture are (see Førsund 1985; 2011; Førsund and Strøm 1988; Perman et al 2011 (first edition 1996)) :

(a) The generation of residuals in production and consumption and discharge to receptors.

(b) The natural processes taking place in the environment as reactions to discharge of residuals, like transformation of residuals by diluting, decaying, decomposing, and transportation between and among receptors.

(c) Defining the environmental services "produced" by the environmental medium and establishing the impact on these of ambient concentrations of residuals.

(d) Evaluating the preferences attached to changes in environmental services, including the time perspective (of the "present generation").

The materials balance, based on the first and second thermodynamic laws, tells us that production activities using material inputs and energy will also generate material or energy residuals. Therefore production activities represent joint production; at least one desirable output is produced and at least one residual is generated simultaneously.

The receiving bodies of Nature, the environmental receptors, play a decisive role in the economic analysis of pollution. The view common in environmental economics is that the receptors provide man with two types of services: residual disposal services and environmental services. The former type relates to the inherent generation of residuals by the materials-processing economy of an industrialised society, and the last type is an omnivorous category of recreation activities like sport fishing, boating, skiing, etc., amenity services, aesthetic values, including the intrinsic value of Nature, and the provision of extraction possibilities from mineral deposits, water, air, etc.

A residual is defined as a pollutant if the corresponding disposal service of receptors negatively affects, quantitatively or qualitatively, the raw materials and recreation services "produced" by the receptors (points (c) and (d) above). The discharge of residuals does not of necessity generate pollution. The natural environment has an assimilative capacity. Owing to dilution, decay, decomposition, chemical transformation, etc. occurring in nature, there are certain threshold values of ambient residual concentrations that must be exceeded before harmful effects appear.

A general equilibrium analysis must show the trade-offs open to rational decisions. However, this paper will only focus on the first point (a) above. (Dynamic problem caused by accumulating residuals in the environment will thus not be covered.)

Significant sources for change as regards point (a) are

(i) The scale of the activities and the output mix among activities
(ii) The input mix in an activity
(iii) Process techniques of production and consumption
(iv) The product characteristics, including durability
(v) Modification[1] of primary residuals ("end-of-pipe" treatment)
(vi) Recycling of residuals
(vii) The location of activities

We will assume that changes in process techniques (option (iii)) are rather modest and short-term measures (done within a year), that the products remain the same (option (iv)) and that recycling of waste materials (option (vi)) is internal only. The last option (vii) is not useful for global pollutants, but for local or regional pollutants like e.g. acid rain or pollutants emitted to air causing localised health effects.

2.2 The materials balance[2]

The materials balance concerns the first step (a) above in Subsection 2.1 in environmental economics modelling. We will simplify and use production activity to cover economic activity. It is the mass of material inputs that appears in the materials balance relation, and it is therefore convenient to operate with two classes of inputs; material inputs (tangible raw materials) x_M

[1] As observed in Ayres and Kneese (1969, p. 283) abatement does not "destroy residuals but only alter their form". Following Russell and Spofford (1972), the concept of "modification" should be used instead of waste treatment or purification to underline the conservation of mass. The mass of residuals does not physically disappear by waste treatment or purification.

[2] The materials balance is quite seldom mentioned in papers published in operational research journals or papers written by researchers from that field. In a recent survey article (Sueyoshi et al 2017) based on 693 papers using data envelopment analysis within energy and environment materials balance is never mentioned once.

and non-material inputs x_S that we will call service inputs (Ayres and Kneese 1969, p. 289). These inputs are not "used up" or transformed in the production process. The materials balance tells us that mass contained in material inputs x_M cannot disappear, but must be contained in either the products y or end up as residuals z. All three types of variables are vectors. The residuals are discharged to the natural environment (point (b) in Subsection 2.1). The variables must be expressed in the same unit of measurement in the materials balance relation. Weight of mass is a natural unit of measurement. The weight of the different inputs can then be summed over the number of material inputs and the same can be done with outputs and residuals:

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

There are n_M inputs containing mass (there are n_S service inputs and $n_M + n_S = n$ inputs), m outputs y and K residuals z. The weights a_{jk}, b_{ik}, c_k convert the unit of measurements commonly used for the variables (piece, area, length, etc.) into weight. (The parameters a_{jk} are also called emission coefficients.) 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), while the second line shows the total mass balance for the production unit. One issue is the creation of residuals during the production process also containing materials provided free by nature; like oxygen for combustion processes and oxygen used to decompose organic waste discharged to water (biological oxygen demand, BOD), 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.

For each production unit we have an *accounting identity* for the use of materials contained in the input x_M. It follows from Equations (1) that the residuals cannot exceed the material content of inputs measured in the same unit; $c_k z_k \le \sum_{j=1}^{n_M} a_{jk} x_{Mj} \, (k=1,...,K)$. The materials can be part of the intended goods y or contained in the residuals z. 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 a production function, but serves as a restriction on specifications of these (more on this later in Subsection 4.3).

The materials balance is valid at a real-life micro level. If production relations are specified at a sufficiently detailed level, we do not have to worry about the materials balance being fulfilled. However, as expressed in Frisch (1965, p. 14): “If we go into details we shall find that the number of circumstances which in one way or another can influence a production result is endless.” He mentions both gravity and molecular forces”, and continues: “No analysis, however completely it is carried out, can include all these things at once. In undertaking a production analysis we must therefore *select* certain factors whose effect we wish to consider more closely.” It is unavoidable to simplify, but this must be based on a good engineering understanding of the activity in question, and following the principle of Ockham’s razor. The specification may then not satisfy the materials balance accurately, but we should be satisfied if our specification is “accurate enough”, and especially avoid specifying relations that cannot in principle conform to the materials balance principle.

2.3 Literature on modelling production of goods and generation of waste

This subsection is an overview of main modelling issues occurring after the seminal paper Ayres and Kneese (1969) was published that will be brief and not show the formal models. However, the key models and issues will be treated in detail in later sections.

The formal model in Ayres and Kneese (1969) is basically an input-output model covering the complete flow of materials between production and consumption and discharge to the natural environment, formulated as a static general equilibrium exercise in the spirit of Walras – Cassel. The use of linear relationships with fixed coefficients served their purpose of demonstrating the pervasiveness of residuals generation, but lacked flexibility regarding technology.

More conventional input-output models including pollutants were formulated by Leontief (1970); Leontief and Ford (1972). An abatement sector dealing with pollutants was introduced. The fixed input-output coefficients were extended to include fixed emission coefficients for various pollutants calculated as emissions per unit of output. Recognising the role of material inputs, fixed coefficients related to outputs were assumed, and also that there were fixed coefficients in production in general, as there are in the standard input-output model. Førsund and Strøm (1974) introduced extensive input-output emission coefficients for Norway in a multi-sector model of economic growth to predict the time paths of discharge of a large number of pollutants, following the economic growth of sectors. Based on data for Norway, the costs of obtaining a “greener” mix of final deliveries for a given amount of primary inputs were shown in Førsund and Strøm (1976); Førsund (1985), the last paper also providing a survey of input-output models including residuals.

A more flexible modelling of production was formulated in Førsund (1972) based on a special formulation of joint production in Frisch (1965) termed factorially determined multi-output production.[3] This model is the main model of this paper and will be extensively treated in Section 4. Suffice it to say that the main idea is that inputs generate simultaneously both intended outputs and unintended ones in the form of waste or residuals, in accordance with the materials balance principle. It is assumed that each output has its own production function in the same set of inputs.[4] This model was extended in Førsund (1973) to include end-of-pipe abatement of residuals.

Baumol and Oates (1988) (first published in 1975) introduced a transformation function in desirable and undesirable outputs and inputs.[5] However, the possibility of allocating given resources to produce a different mix of outputs that was a consequence of the formulation was not discussed. To overcome the inherent problem of allocating zero resources to produce undesirable outputs, residuals were treated as inputs without any discussion. It will be pointed out in Subsections 3.3 and 4.3 that this procedure is counter to the materials balance. The use of a Baumol and Oates type of transformation function is widely adopted in the environmental economics literature. In the well-reputed textbook of Perman et al (2011, p. 25) the assumption of using residuals as inputs is defended, based on their production function (2.3) for firm i: $Q_i = f_i(L_i, K_i, M_i)$ where Q_i is the desirable output, L_i labour, K_i capital and M_i residuals:

> Equation (2.3) may appear strange at first sight as it treats waste flows as an input into production. However, this is a reasonable way of proceeding given that reduction in wastes will mean reductions in output for given levels of the other inputs, as other inputs have to be diverted to the task of reducing wastes.

[3] This type of model but without reference to Frisch (1965) was used in Mäler (1974). His book was written when the author was a visiting scholar at Resources for the Future (RFF) invited for a year by Allen V. Kneese.

[4] This model applied with explicit reference to Frisch (1965) to production of both desirable and undesirable outputs, was, to the best of our knowledge, first used in Førsund (1972); (1973) and developed further in Førsund (1998); (2009); (2017a).

[5] A production possibility set was also introduced using the transformation relation such that the value of the transformation function is zero for efficient utilisation of resources and less than zero for inefficient operations, but no inefficiency issues were discussed. The solutions to the optimisation problems were based on the production of desirable outputs being efficient.

Cropper and Oates (1992, p. 678) are adopting the same type of arguments. However, such a diversion of inputs may be relevant in a macro setting, but not in a setting of a single firm. A marginal productivity is calculated for an input keeping all other inputs constant. However, if production is efficient then output cannot increase by increasing a residual because the other inputs are constant, so the materials balance rules out both an increase in output and an increase in residuals. The residuals are generated by increasing material inputs, and thus cannot itself be treated as an input.

The Ecological Economics journal started publishing in 1989. Joint production is regarded as a fundamental part of ecological economics. As Baumgärtner et al (2001, p. 365) state, "the concept of joint production should be considered as one of the conceptual foundations of ecological economics". Joint production will be discussed in Section 3.

Pethig (2003); (2006) follow up the general equilibrium approach of Ayres and Kneese (1969). However, the materials balance is used as a part of the production relations. This usage is criticised in Subsection 4.3.

So far the reviewed papers did not discuss inefficiency of operations. The production relations were based on efficient utilisation of inputs. Within the axiomatic approach to measuring inefficiency, Färe et al (1986); (1989) were the first empirical papers to introduce generation of residuals, or bads as these outputs were called, together with desirable outputs, or goods. Then eco-efficiency could be measured. Especially the 1989 paper spawned a large number of papers (556 citations in SCI per 01.05, 2017). The 1986 paper was somewhat peculiar assuming a technology with strong disposability of the bads before the introduction of regulation of residuals, and then assuming weak disposability of the bads after the introduction of the policy. (Shephard (1970) introduced the concept weak disposability that will be discussed in Section 6.) It is rather questionable if imposing a regulation can change the nature of technology in such a way (see Section 5 where abatement is introduced without any change in the production technology). The opportunity cost of regulation is measured as the relative loss of outputs based on the two sets of different hypothetical *frontier* values given the inputs of the observations.

Färe et al (1989) introduced a hyperbolic efficiency measure expanding the goods with a common scalar and contracting the bads with the inverse of the scalar to project an inefficient observation to a reference point on the frontier. This was done in order to credit producers for "their provision of desirable outputs and penalize them for their provision of undesirable outputs" (Färe et al 1989, p. 90). A problem with this procedure is the arbitrariness of using a single scalar only, there is no obvious rational for this and no argument is offered. The problem is that the common scalar implies an arbitrary trade-off between goods and bads confounding the efficiency analysis as such[6]. The assumptions that goods and bads are jointly weakly disposable, but that goods alone are strongly disposable, are also made without any explanation. (The criticism of the single equation model and weak disposability is presented in Subsections 3.3, 4.3 and Section 6.)[7]

The use of a directional distance function instead of a radial one to discriminate between goods and bads were introduced in Chung et al (1997), and the approach has become popular.[8] However, it is based on a single-equation model and assuming weak disposability. An expansion factor for outputs that enters additively for goods but is subtracted for the bads when identifying frontier points (footnote 6 also applies here) is estimated. "Rewarding" the production of the good and "punishing" the production of the bad with the same factor is just an implicit relative evaluation of these outputs that is quite arbitrary. In addition the choice of direction to the frontier will influence the measures.

Consequences on efficiency of introducing environmental regulation was put forward in Porter (1991); Porter and van der Linde (1995), and called The Porter hypothesis in the literature.[9] It is based on the existence of inefficiency, but the approach is different from the axiom-based measures of efficiency, being purely empirically based. The hypothesis is that substantive environmental regulation with flexibility of firms' choice of abatement techniques may induce firms to innovate to such a degree that profit increases. Such regulation represents a win-win situation. It is stated that the pessimistic view stems from considering a static situation only, but that the pressure of environmental regulation induces a dynamic process of change representing retooling, process improvement and technical change, which more than offsets the abatement costs. However, Porter and van der Linde do not present any formal mechanism supporting the cost-offset hypothesis, but refer to a few examples of successful adaptation and technical change. The Porter hypothesis and attempts to model the positive dynamics, empirical

[6] In Førsund (2017b), it is argued that desirable and undesirable outputs must be measured in the same unit in order to perform a trade-off, cf. point (d) in Subsection 2.1. This can be done by introducing a damage function, as also used in Førsund (2017a,b).
[7] Førsund (1998) was the first to criticise both the assumption of bads as inputs and the weak disposability assumption. However, the first submission to a journal of the paper was rejected, and an improved version Førsund (2009) published first 11 years later.
[8] However, problems with translation properties are pointed out in Aparicio et al (2016), and infeasibility problems pointed out in Arabi et al (2015).
[9] This section builds on Førsund (2017b), Chapter 8.5.

studies and critique of the hypothesis (e.g. Palmer et al 1995), are extensively reviewed in Brännlund and Lundgren (2009); Lanoie et al (2011); Ambec et al (2013). The latter three references provide long lists of references to the literature on the Porter hypothesis.

Porter and van der Linde suggest two different dynamic effects. A neat illustration of the story told in Porter and van den Linde (1995) of the increased efficiency effect and the shift in technology effect, is presented in Brännlund and Lundgren (2009), connecting the Porter hypothesis to the efficiency literature. First, assuming that there is inefficiency in utilisation of resources before the introduction of environmental regulation, this inefficiency is reduced or even removed after regulation has been introduced. Second, the regulation induces new technology to be developed, shifting the production function outwards. This is set out in Fig. 1. In the space of the desirable output (q in the original figure) and emissions z the pre-regulation

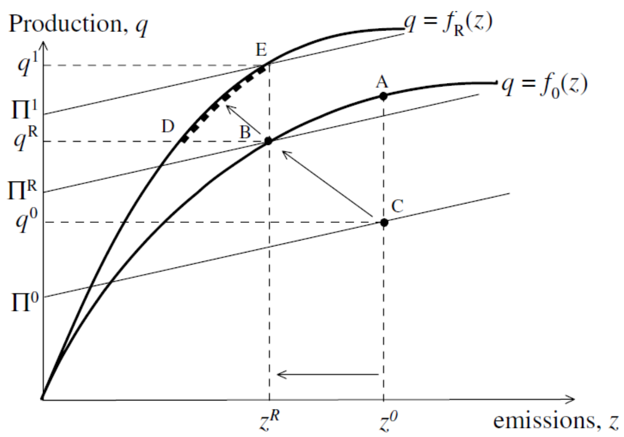

Figure 1. The Porter hypothesis
Source: Brännlund and Lundgren (2009), p.83

position of the firm is at the inefficient point C below the initial frontier production function $f_0(z)$. The efficient point A on the frontier shows the production the firm could have had corresponding to emission z^0. After introducing regulation, the firm improves its efficiency and reduces the emissions down to the regulated amount, z^R, and increases output from q^0 to q^R at point B on the initial frontier. Then there is a shift of the frontier due to innovation after introducing regulation to $f_R(z)$ where the point E is the efficient point for the level z^R of the reduced emission. The firm continues to reduce emissions and increase output q, and profit Π moving towards the new frontier.

In a series of eight more or less overlapping papers published in the period 2010-2013 (see Førsund (2017a) for the list of these references and evaluations), Sueyoshi with co-authors actually employed the Frisch (1965) factorially determined multi-output model with both good and bad outputs being a function of the same inputs (see e.g. models 4 and 5 in Sueyoshi and Goto 2010, p. 5905). This is as formulated for frontier models in Førsund (1973); (1972); (2009), but without these references, using non-parametric DEA models for empirical applications mainly to the energy sector. No explanation is given for the choice of this type of model. Sueyoshi et al developed separate efficiency measures for desirable and undesirable outputs, but emphasis was put on unified measures by solving for the combined production possibility sets.

The multi-equation model of Førsund (2009) is followed up in Murty et al (2012) using a model called the by-production approach.[10] In the theoretical model with abatement, the first relation is a transformation relation between a desirable output, an abatement output and two types of resources; one pollution generating and the other not. The second relation has the pollution (or residual) as a function of the polluting input (positive impact) and the abatement output (negative impact). The production possibility set is formed by the intersection of the two sets based on these relations. Efficiency measures for the two types of output separately and a form of aggregated measure were developed for non-parametric DEA models.[11] Førsund (2017a) argues against the usefulness of aggregate measures for policy purposes.

The abatement activity is only indirectly treated in a non-transparent way in Murty et al (2012), maybe due to reflecting process changes. In Førsund (2017a) it is pointed out that it is difficult to get data for process changes and resources consumed in such activities. Instead an end-of-pipe abatement facility is explicitly modelled there, separate efficiency measures for desirable outputs and residuals efficiency are developed, and a measure for abatement efficiency.

Murty and Russell (2016)[12] show that the multi-equation by-product model in Murty et al (2012) (with abatement) has an axiomatic foundation supporting such models, reconciling the abstract axiomatic characterization of an emission-generating technology in Murty (2015) with the empirically oriented by-production technology formulated by Murty et al. (2012).

[10] The name is meant to point to the production of both desirable goods and residuals. However, the name is not according to the classical economist, calling by-products commercial outputs, but with less value than other goods. Their word for residuals was waste, see next Subsection 3.1.

[11] Dakpo et al (2016) review weak disposability models and the by-production model. Hampf (2017) reviews single equation models only, and has several critical remarks to the typical Färe et al models of desirable and undesirable outputs.

[12] Førsund (2017a), Murty and Russell (2016) and Hampf (2017) are forthcoming in a special issue 'Good modelling of bad outputs' of *Empirical Economics*, see Kumbhakar and Malik (2017).

3 Joint production

The materials balance forcefully establishes that any production involving material inputs results in two types of outputs; desirable and undesirable. Therefore, a joint output model must be used in order to model such type of production.

3.1 The historical background

Most of the current textbooks, at least on a lower level, dealing with production theory assume a single output being produced by two or more inputs. However, this choice of modelling is not based on any empirical evidence that this is the dominating form of production. On the contrary, joint production seems to be the general rule in practice. As pointed out in Kurz (1986, pp. 1-2):

> The view "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 exception", has been advocated already one century ago by W.S. Jevons.

Kurz (1986) reviews how a number of classical and early neoclassical economists, among them Adam Smith, Karl Marx, von Thünen, Longfield, Mill, von Mangoldt, Jevons, and Marshall, treat joint production. The examples used by these economists were mainly drawn from agriculture; like raising sheep yielding wool and mutton, animal rearing yields meat and hides, growing wheat yields grain and straw, and forestry yields timber and firewood, etc.

A standard textbook way to represent a multiple-output multiple input production relation is to use a single implicit functional representation:

$$F(y,x) = 0\,, F'_y > 0, F'_x < 0 \tag{2}$$

$F(.)$ is commonly called the transformation function. y and x are vectors of outputs and inputs, respectively. The signing of partial derivatives identify outputs and inputs. We assume that $F(.)$ is continuous, but not necessarily differentiable at all points.

There is a clear distinction between inefficient and efficient operation. The production possibility set corresponding to (2) can be written

$$F(y,x) \le 0\,, \tag{3}$$

and contains in principle all feasible production plans. An engineer will probably not waste his time mapping inefficient ways of producing; a blueprint of technology represents efficient operations. The production function concept (2) is attached to efficient operations that are on the border of the set. In the efficiency literature the efficient way of producing is termed the frontier function (or best practice as used in Farrell 1957) due to the factual observations used to estimate the frontier function). Inefficient observations are in the interior of the set.

The concept of disposability is expressed by assuming that for a feasible point, an increase of the *x*-vector for constant outputs leaves the new point inside the production set, and a decrease of the output vector *y* for constant inputs leaves the new point inside the set. On the frontier, we see from (2) that an (infinitesimal) increase in an input leads to the interior of the set, as will a decrease in an output.[13] The monotonicity expressed by the partial derivatives in (2) corresponds to the free (strong) disposability of outputs and inputs of the set (3). The production possibility set allows observations to be located in the interior of the set, so such a set is therefore a natural starting point for analysing inefficiency. We will return to this point later in Section 6.

According to Kurz (1986, p. 16) Karl Marx researched production technologies extensively and in addition to agriculture examples had many other examples of joint production, like mining, forestry, paper manufacturing, the chemical industries, the textile industries, mechanical engineering, etc. Marx divided products into a main desired product, and one or several by-products that may or may not be useful, and that, at any rate, are of secondary economic interest. He was especially concerned by waste and stated (according to Kurz 1986, p. 16): “The so-called *waste* plays an important role in almost every industry.” True to form Marx called waste *excretions of production*. Furthermore, Marx stated that the “excretions of production” should be reduced to a minimum, and the immediate utilisation should be increased to a maximum of all raw- and auxiliary materials required in production. This sounds very modern!

Jevons introduced the distinction between commodities and discommodities, the last category could cause inconvenience or harm (Jevons 1965, p. 58), and he pointed out that discommodities could have negative value. He used as an example of discommodity waste from a chemical plant fouling the water downstream (Jevons 1965, p. 202).

[13] To say that inputs and outputs can be disposed of by throwing them away (Shephard, 1970, p. 14) is not in accordance with economic use of inputs and outputs.

The classical and neoclassical economists focused much of the discussion of joint production on the problem of unique determination of the output prices. There is a problem of determining the share of costs due to joint outputs. It was often assumed that market forces would lead to as many equations as outputs, and that a unique set of prices could be determined.

There is one more recent definition of joint production in the literature (Pasinetti 1980) that has some following based on Sraffa (1960) that should be mentioned. Considering time as periods, capital is entered as an input at the start of a period, and defined as an output at the end of the period. This type of joint production is not the type of joint production that we are concerned with in this paper and will be disregarded in our classification.

3.2 Frisch on joint production

The materials balance tells us that desirable and undesirable outputs are produced jointly. Therefore the modelling of joint production is essential within environmental economics. Joint production takes place when the production unit in question produces more than one output. According to Frisch (1965), that has a comprehensive discussion and classification of joint production, joint production implies that there is a technical connection between products; because there are certain inputs either which *can* be used or on technical grounds *must* be used jointly, or because there are inputs that can be used alternatively for one product or the other.[14] In The New Palgrave, producing outputs by separate production processes is also classified as joint production; there is a choice how to allocate a given amount of inputs to outputs. This is a typical situation in international trade when countries are considered as production units. In Chambers (1988); Kohli (1983); Nadiri (1987, p. 1028) this is called non-joint production.[15] In the two first references, a main example is how production is modelled in the international trade literature.

[14] According to Kurz (1986, p.25), Mangoldt's definition is about the same as the one of Frisch: "pure joint production (or joint production in the technical sense) and what may be called competing, alternative or rival (Edgeworth) production which derives from the fact that a firm's (given) productive equipment may be used for several purposes."

[15] Using non-jointness when defining joint production seems a little awkward; sounding almost like a contradiction. Nadiri (1987, p. 1028) claims that absence of non-jointness is a crucial test of joint production, in spite of including non-jointness as part of the definition of joint production: "Joint production includes two cases: (1) when there are multiple products, each produced under separate production processes - i.e. the production function is non-joint […]". He uses the term "intrinsic jointness" when there is jointness in a technical sense.

We will use the Frisch (1965) classification below.[16] Two main forms of joint production are suggested:

(i) Inputs can be used to produce different outputs within the same general production technology. Examples in Frisch (1965) are that a piece of agricultural land can be used alternatively to grow different crops, and that a wood cutting machine tool can be used to produce different types of wood articles. The producer has a freedom of choice as to the mix of products he wants to produce. Frisch calls this *assorted production*.

(ii) The technical process is such that it is impossible to produce one product without at the same time producing one or more other products; using coal as input gas, coke, and tar are produced, and raising sheep results in wool and mutton.

The connection between outputs demands, according to Frisch (1965, p. 269), that production laws cannot be studied separately for each product, but must be considered simultaneously for all connected products. In order to catch the engineering complexities of multioutput production Frisch (1965) generalised various possibilities by introducing a system of μ equations between m outputs y and n inputs x[17]:

$$F^i(y_1,...,y_m,x_1,...,x_n)=0,\ i=1,...,\mu \tag{4}$$

These relations are frontier ones. Corresponding production possibility sets will be

$$F^i(y_1,...,y_m,x_1,...,x_n)\le 0,\ i=1,...,\mu \tag{5}$$

The two classes (i) and (ii) above are special cases of (4). The production possibility set for the system of equation (4) will be the intersection of the production possibility sets (5) for each equation (see Chambers 1988, p 290).

Frisch (1965) introduced the concept of degree of assortment α that tells us the limits for reallocating inputs on outputs, $\alpha = m-\mu$. If we have only a single relation in (4), i.e. μ = 1 - as in (3) - then the degree of assortment is maximal; $\alpha^{\max} = m-1$. If there is no assortment, i.e., there is no choice of output mix given the inputs, then $\alpha^{\min} = m-\mu = m-m = 0$. There are as many equations μ as there is products m.

[16] There are unfortunately few references to Frisch (1965) about joint production, neither Chambers (1988); Kohli (1983); nor Nadiri (1987) refer to Frisch. It seems appropriate to make his take on joint production better known.
[17] In a very readable book, Whitcomb (1972) discusses the connection between externalities and joint production. He refers both to Frisch (1965) and specifies a variant of the system (4), and of Ayres and Kneese (1969), but does not use the materials balance explicitly in his analysis.

An important case is $m \geq \mu$; the degree of assortment is non-negative as assumed above. However, in the system (4) there may be more relations than products so the case $m < \mu$ cannot be excluded. If this is the case then there are one or more pure product bands independent of factors. Frisch (1965) calls the number of such equations for the degree of coupling κ. This is not determined by m, n, μ, but is expressed by the greatest number of equations in (4) that do not contain any of the inputs when transforming the equations in such a way that as many of them as possible are free from inputs (Frisch, 1965, pp. 278-279). The band (or coupling) between outputs is expressed by:

$$F^c(y_1,..,y_m) = 0\,,\; c \in C \tag{6}$$

where C is the set of relations between outputs only of the μ equations in (4). In the classical literature on joint production, it was often assumed a fixed relation between outputs, e.g. the quantity of wool bears a fixed relation to the quantity of mutton (Frisch 1965, p 271).

There may also be pure factor bands between inputs, i.e. relations between inputs independent of outputs:

$$F^b(x_1,..,x_n) = 0, b \in B \tag{7}$$

where B is the set of relations between inputs only of the μ equations in (4), e.g. a chemical process where inputs must be applied in fixed proportions.

The efficient border of the production possibility set (3) is specified as a single functional relationship in (2). This is commonly done, but we see that the system of equations (4) is much more general. However, it may be problematic to impose convexity assumptions on the general specification of the intersection of μ technology sets (5).

An important special case of the system (4) is that the equations can be solved with respect to the m products (cf. the concern of the classical economists mentioned in Subsection 3.1), and where the m ensuing production functions are single valued. This is the case of *Factorially determined multi-output production* (Frisch 1965, p. 270):[18,19]

---

[18] Kohli (1983) introduces this case in his Definition 4 on p. 213 and calls it “non-joint in input prices”. This seems a little peculiar name since there are no input prices appearing in the definition (however, duality results use shadow prices). He has no reference to Frisch (1965) that introduced this type of relation decades before.

[19] Chambers (1988) calls the factorially determined functions for generalised fixed coefficients technologies.

$$y_i = f_i(x_1,...,x_n),\ i = 1,...,m \tag{8}$$

The same set of inputs appears in all separate production functions. Both the degree of coupling and the degree of assortment are zero. The products are separable, but the ratios between outputs are not fixed, but changes with input mix. The mix of wool and mutton depends on the breed of sheep and maybe feeding, and the mix of eggs and poultry meat depends on the feeding.

Within this case, there are important sub-cases. Frisch (1965, p. 275) claims that necessary and sufficient condition for coupled (joint) products in the case of factorially determined multi-output production functions is that there exist a functional relationship $F(y_1,...,y_m) = 0$, independent of inputs. One way of obtaining a fixed ratio between outputs is:

$$y_i = c_i f(x_1,...,x_n) \Rightarrow \frac{y_i}{y_j} = \frac{c_i}{c_j},\ i, j = 1,...,m,\ i \neq j \tag{9}$$

where the c's are constants. The technology is the same for all outputs except for a scaling constant c_i implying a fixed ratio between the outputs. An example of coupled products is refining of crude oil and the distillates emerging from the same process.

There may be more complex couplings than (9). The ratios between products may be a function of the quantities of outputs, but the degree of assortment is still zero. A complete coupling occurs when isoquants in the input space coincide, and substitution regions are identical (see Frisch 1965, p. 273 for an illustration). The relation between outputs for the same isoquant is independent of input quantities.

3.3 Restrictions on production models

Going back to classical or neoclassical economists concerned with joint production in Subsection 3.1, the three categories of outputs specified were main products, by-products and waste products. By-products (the alternative spelling 'bi-products' is used in Frisch 1965, p. 11) are commercial products of more minor economic importance than the main products. Waste products without economic value to the producer are termed residuals in this paper.

In the influential textbook by Baumol and Oates (1988, first edition 1975) the essence of their model can be captured by specifying a single transformation relation as the border of the set and the production possibility set as follows (more based on externalities modelling than referring to the materials balance)[20]:

$$F(y,z,x)=0\ ,F'y,F'z>0,F'x<0$$
$$F(y,z,x)\leq 0 \tag{10}$$

Notice that with the sign conventions for the partial derivatives all variables exhibit strong (free) disposability. However, the question is if this relation can function as the efficient border of the production possibility set as relation (2) does for the set (3). As the first relation in (10) stands it has a maximal degree of assortment according to the scheme of Frisch (1965), meaning that all the inputs can be reallocated to produce the desirable products *y* and no resources used to produce the undesirable products *z*, unless more conditions are specified. However, this goes against the fact that the residual *z* is not a result of choice as is the case with the desirable outputs, but is physically linked to the material inputs used in the production of desirable outputs. Baumol and Oates may have been aware of this problem, because without telling the reader they assume that the *z* variables function *as if* they are inputs. The formal Pareto-optimal results when they maximise the utility of one consumer for a given input vector *x*, under the condition that all other consumers' utilities shall not be lower than given levels, then apparently seems to make sense. However, this cannot be done because as we see from the materials balance (1), the material content is distributed on products *y* and residuals *z*. Residuals cannot be reduced in (10) for given *x* because the transformation function is by definition efficient, in the sense that it is constructed by maximising outputs *y* for given inputs *x*, neglecting residuals *z* because they are undesirable outputs. The maximal possible amount of raw materials is already extracted from *x* to produce desirable outputs *y*, and this amount cannot then be increased for a given *x* vector.

The conjecture is that a single transformation relation cannot work without specifying restrictions on the degree of assortment. However, even more serious is the combination of the materials balance and the efficiency assumption that the transformation function is based on. The combination of these two factors implies that we cannot operate with a functional trade-off between a desirable output and a residual. The option to reallocate inputs between desirable goods and residuals is simply not available. The residuals are generated simultaneously with desirable outputs by using material inputs. Some sort of separation between modelling of

[20] The assortment property is not discussed in Baumol and Oates (1988, Chapter 4). The materials balance principle is not mentioned in the book.

production relations for the desirable and undesirable output is needed. This point will be developed further in Section 6.

4 Multi-equation models for desirable and undesirable outputs

To make a useful model is an art. As quoted in Subsection 2.1 Ragnar Frisch was fully aware of the need for simplification. He introduced the term 'model world' in Frisch (2010, pp. 31-32):

> The observational world itself, taken as a whole in its infinite complexity and with its infinite mass of detail, is impossible to grasp. […] In order to create points where the mind can get a grip, we make an intellectual trick: in our mind we create a little *model world* of our own, a model world that is not too complicated to be overlooked, and which is equipped with points where the mind can get a grip, so that we can find our way without getting confused. And then we analyse this little model world instead of the real world. […] It shall picture those indefinable things in the real world which we may call 'essentials' [...]

Part of the 'essentials' is that the model should satisfy the materials balance and efficiency properties of the production relations. A solution to the problems is to employ the Frisch (1965) scheme of factorially determined multi-output production in the previous Subsection 3.2 as in (8), introducing residuals as outputs in the same way as desirable outputs. The adoption of a multi-equation model instead of a single-equation one is crucial for satisfying the materials balance and efficiency conditions. More specifically, as stated in the previous Subsection, there cannot be any functional trade-off between desirable and undesirable outputs for given resources.

4.1 The Frisch multi-equation model

As stated previously, residuals are generated simultaneously with the desirable products 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 activity and generation of pollutants. The model from the production theory of Frisch (1965) presented in Subsection 3.2 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 undesirable residual are added to the single-output functions for desirable goods:[21]

[21] Leontief type models with fixed coefficients are not considered.

$$y = f(x_M, x_S),\ f'_{x_M}, f'_{x_S} > 0\ ,\ f''_{x_M}, f''_{x_S} < 0$$
$$z = g(x_M, x_S), g'_{x_M} > 0, g'_{x_S} < 0, g''_{x_M}, g''_{x_S} > 0 \tag{11}$$

To keep the model as simple as possible we consider a single desirable output *y* (or the good for short) that is the purpose of the production activity, and a single residual or undesirable output *z* (a pollutant or a bad for short). Two types of inputs only are also specified following Ayres and Kneese (1969); material inputs x_M and non-material inputs, or service inputs x_S. Generalising to multi-output and multi- pollutants can be done just by adding more equations, one for each variable, keeping the same inputs (their number can easily be expanded too) as arguments in all relations (see Førsund 2009).

In the previous Subsection 3.3, it was stated that the model must have a certain property of separability. The model (11) satisfies this property because the production of desirable outputs is not influenced by undesirable outputs, and *vice versa* for the production of undesirable outputs.[22]

It should be stressed that the two relations in (11) do not represent physically separate technologies. It is the analyst that simplifies a complex technology of simultaneous transformations to the two relations. Changes in inputs generate simultaneously both the intended and unintended outputs. Generation of residuals cannot be controlled independently, but follows from the use of the inputs needed for production of the intended outputs.

The material inputs are *essential* in the sense that we will have no production neither of material goods nor bads if $x_M = 0$: [23,24]

$$y = f(0, x_S) = 0, z = g(0, x_S) = 0 \tag{12}$$

The function *f*(.) is defined by maximising *y* for given inputs. The partial productivities in the good output production have the standard properties of positive but decreasing values. The signing of partial derivatives of the residuals function may be more unconventional. It seems reasonable to assume positive but increasing marginal productivity of the material input, and

[22] The Frisch model of factorially determined multi-output equations is not the only model having the sufficient separability properties.

[23] One or more service inputs may also be essential, but the point is that residuals are in general an unavoidable feature using material inputs in production. Although $y = f(x_M, 0) = 0$ we may have $z = g(x_M, 0) > 0$; e.g. as in a fully automated thermal electricity-generating plant running in a spinning mode (the energy stored by spinning is then not considered an output).

[24] In Murty and Russell (2016, Section 5) x_M is called jointly essential with *z*, it is rather obvious that *z* cannot be zero for $x_M > 0$, however, Rødseth (2017a) covers the possibilities with the concepts output- and input essentiality.

negative but decreasing marginal productivity of the service input. The positive partial productivity of service inputs x_S in the desirable output production function and the negative sign in the residuals generation function can be explained by the fact that more of a service input improve the utilisation of the given raw materials through better process control, fewer rejects and increased internal recycling of waste materials.[25] The negative partial derivative of service inputs in the residuals function mirrors the positive sign in the output function. The function $g(.)$ is defined by minimising z for given inputs. The residuals generation function may degenerate to a fixed relation between residuals and raw materials similar to Leontief technologies, but then we will have a Leontief relation for the good y also.

4.2 Substitution possibilities

There will in general be substitution possibilities between material and service inputs. The rate of substitution evaluated at a point on an isoquant for y in (11) is $(-f'_{x_M} / f'_{x_S}) < 0$ in the interior of the substitution region (this is a Frisch concept for the economic region; i.e. all marginal productivities in goods production are positive). This is the amount of the service input that has to be increased if the material input is reduced with one unit, keeping output y constant. Considering several material inputs there may be substitution possibilities between them also, e.g. between coal and natural gas, that will keep the output constant, but decrease the generation of bads if the marginal contribution of gas to creation of bads is smaller than the marginal contribution of coal.

There is also substitution between the two types of inputs in the residuals-generating function. The marginal rate of substitution is positive, $(-g'_{x_M} / g'_{x_S}) > 0$ in the interior of the substitution region for bads due to the marginal productivity of service inputs being negative in this case. The necessary increase in the service input to keep a constant level of the residual when the material input increases with one unit, is increasing following the signing of the partial derivatives in (11). This implies a special form of isoquants in the factor space and the direction of increasing residual level compared with a standard isoquant map for the output, as seen in Fig. 2. (The substitution regions, the borders of which have zero marginal productivities, are not shown.) The isoquants for the two outputs can be shown in the same diagram because the

[25] Cf. the famous chocolate production example in Frisch (1935), discussed in Førsund (1999), of substitution between labour and cocoa fat due to more intensive recycling of rejects not filling the forms the more labour and less cocoa fat that are employed producing the same amount of chocolate.

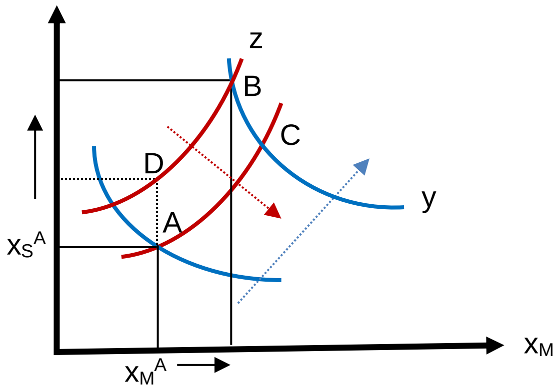

Figure 2. *Isoquants for the production of y and z*

arguments in the functions are the same. The level of the residual z is increasing moving South-East (red isoquants) in the direction of the broken red arrow, while the level of the intended (desirable) good y is increasing moving North-East (blue isoquants) in the direction of the broken blue arrow. Going from point A to point B in input space, increasing both the material and service inputs, but changing the mix markedly towards the service input, we see that the production of the residual z has decreased while the production of output y has increased. Reducing the service input but increasing the material input going from point B to point C, keeping the same level of the desirable output, the level of the undesirable output increases. All points of the type (x_S^A, x_M^A) in input space generating points (y^A, z^A) in output space are frontier points.

There are obviously limits to substitution between material and service input keeping the same desirable output.[26] Moving along the y isoquant from point A in a North-West direction there is a limit to the amount of raw materials that can be extracted from the material input and keeping the output constant, i.e. there is a lower limit on how much the residual generation can be reduced. The lower limit of the residual is reached at the border of the substitution region for the good output isoquant in question. (This is not illustrated in the figure.) Another angle on this lower limit is keeping the material input constant at the level x_M^A at point A, and then see how much residuals can be reduced increasing the service input from x_S^A. Let us say point D will be the point with the minimum generation of residuals z, but then the good output has also increased. Point D is then on the border of the substitution region (zero marginal productivity for service input) for the isoquant for the new level of good output. The minimum level of

[26] Continuing the chocolate example in footnote 25: when so much labour is employed so that all the defect chocolates re-circulated to the production cannot be increased any more employing more labour, based on a given mass of raw material with a minimum of cocoa fat for the required taste.

residual generation depends on the level of both types of inputs, as does the level of maximal good output.

Obviously, there must also be upper bounds on emission generation for given amounts of emission-generating inputs; residuals cannot exceed the material inputs measured in mass units (see the discussion of (1)). Since minimum levels are the crucial variables in the analysis upper bounds will not be specified explicitly for convenience (following Murty and Russell 2016).

In addition to the two ways of reducing generation of residuals by input substitutions there is the obvious way of reducing the production of desirable products by scaling down the use of both inputs. However, this is often the most expensive way to reduce residuals generation (Rødseth 2013).

4.3 The materials balance and the multi-equation model

Model (11) is a theoretical one, our model world, and as such is compatible with the materials balance. A theoretical model that is not compatible with the materials balance is obviously inferior to a model that does comply. Notice that the materials balance (1) is a physical law and should not be regarded as a separate part of the production relations (11). The observations of y and z generated by inputs x_M and x_S through the $f(.)$ and $g(.)$ functions must satisfy the materials balance. Thus, this identity constrains what kind of production relations to specify, but does not give any specific information as to the nature of the technology. It should be born in mind that the system (11) of production functions is a long way from describing physical engineering relations in real life details. As is standard in economics, the relations are extreme simplifications, but containing the essential features necessary for the analyses we want to do in our model world. As stated in Subsection 2.2 the materials balance is functioning on a much more detailed level of aggregation, especially when representing the residuals discharged to the environment and the part of residuals that are due to physical/chemical processes of combustion. It will be difficult to get data on the level necessary to control the materials balance numerically.[27]

It may be the case that the materials balance principle is taken a little too literally or philosophically in ecological economics doing practical modelling (Baumgärtner et al 2001;

[27] A practical use of the materials balance is the estimation of emission coefficients, e.g. when coal is used in thermal electricity generation, assuming a specific physical composition of coal and optimal running of the process. Then, because the complete contents of coal end up as residuals, knowledge of the combustion process allows the emission coefficients concerning the substances actually discharged to the external environment to be calculated.

Lauwers 2009). It should be born in mind that the materials balance, as an identity for all kinds of processes using material inputs, cannot give any information about a specific technology at hand, but only give some restrictions on what kind of relations to specify. A restriction mentioned in Pethig (2003) is that the Cobb–Douglas function cannot be used because of the extreme substitution possibility between inputs. A problem in Pethig (2003); (2006) is the use of the materials balance in specifying the residuals generation function by just inserting for z in the materials balance identity (1). It is rather difficult to believe that such a relation can properly represent any specific technology.

How can we then know that the relations (11) comply with the materials balance principle? The short answer is that we cannot know this until we have accurate observations, but due to the requirement of details, this will be quite difficult to carry out. However, what we do know from the results of Section 3 is that there cannot be what we can call a direct functional basis for a trade-off between goods y and bads z.[28] In Fig. 2 we have no trade-off between y and z for given x; i.e. at point A the output levels y and z are given for the input levels x_M^A and x_S^A at A. To change the mix between y and z always requires changing input mixes and levels.

5 End-of-pipe abatement and regulation

5.1 End-of-pipe

We will add an independent abatement process to the multi-equation model (11). End-of-pipe abatement often consists of a facility separated from the production activity. Other abatement options in the short run is to retool the processes and do small-scale changes. These options are alternatives to integrated technological process solutions. However, it is often rather difficult to identify such activities distinct from the general process activity and to identify the inputs involved. It is easier to do this with a stand-alone abatement facility in terms of inputs used and outputs produced. Add-on abatement requires that we make a clear distinction between primary pollutants z from the production process and pollutants z^D actually discharged to the environment. Primary pollutants can then be regarded as an input to the abatement process. In

[28] This trade-off should not be confused with a correlation between y and z depending on indirect effects. Increasing x_M in (11) will lead to increases in both y and z, thus we have a positive correlation, increasing x_S will increase y but decrease z, and thus we have a negative correlation.

addition to inputs like labour, capital, and energy, other inputs like absorbing substances, chemicals and specialised capital, may have to be used in order to convert part of the primary pollutants z into abated pollutants z^a as outputs creating less harm (usually no harm at all is assumed in applications) than the primary ones (Førsund 2009)[29]. In the long run there may be a choice between end-of-pipe abatement and large-scale investment in new technology integrating production processes and abatement. The time horizon for environmental improvement, uncertainty about what can be achieved by new technology, and uncertainty about the future regulatory regime may determine the choice between these two options.

Expressing the abated residuals as outputs we formulate the following abatement production function (see also Førsund (1973); Pethig (2006); Färe et al (2013); Hampf (2014); Førsund (2009), the last paper provides a generalisation to more than one primary residual, and the introduction of new types of abatement outputs with detrimental environmental effects):[30]

$$\begin{aligned} &\delta z^a = zA(x_M^a, x_S^a),\ \ A'_{x_M}, A'_{x_S} > 0, \frac{\delta z^a}{z} \in [0,1] \\ &z^D = z - \delta z^a \geq 0 \end{aligned} \qquad (13)$$

The abatement activity receives the primary residual z appearing in (11) and uses resources x_M^a, x_S^a to modify z into another form z^a that by assumption (for convenience) can be disposed of without social or private costs. In order to express the residual variables in the same unit, we can convert abatement residuals z^a, typically given another form than the primary residual, into units of primary residual applying a conversion coefficient δ. The theoretical feasible range of modification is from zero to one. The partial productivities in the abatement production function are assumed positive. Increases in the abatement inputs contribute to an increase in the relative share of abated amount and an absolute increase for a given amount of primary residual.

To make sense of the abatement function (13) it is assumed that the amount of abatement inputs determines the capacity to treat the primary residual generated by the production system (11). It may be more realistic that capital equipment determines a physical capacity to treat the primary residual. However, we do not want to introduce an analysis of investing in abatement

[29] Modification and recycling of residuals using factorially determined multioutput production functions was introduced already in Førsund (1973).

[30] Hampf (2014) has a similar specification of the abatement function with primary residuals as input together with a stage-specific amount of non-polluting abatement inputs (same as similar inputs in the production of the good output in stage 1), a shared input of the two stages, and part of the output from stage 1 as an input. The abated amount is the single output, and the secondary residual emitted to the environment is residually calculated as in Førsund (2009); (2017a).

capacity. We let the amount of current inputs in (13) determine the abated amount and assume that the maximal relative abatement level that realistically exists will not be reached for economic reasons (due to sufficiently decreasing marginal productivities of the inputs).

The second equation defines the amount of residual z^D that is actually emitted to the environment. It is often called the secondary residual in the environmental economics literature, but also controlled emission is used. One may think of the secondary pollutant as an output, but it is more to the point (and analytically more convenient) to regard the secondary pollutant to be determined residually.[31] It is assumed that the secondary residual has the same form as the primary residual, e.g. measured in CO_2, or SO_2, or in the form determined by the combustion process or production process in general. It is typically the case that at least all gaseous residuals cannot be dealt with completely and modified to harmless substances, so $z > \delta z^a \Rightarrow z^D > 0$. A limit around 95 % is often mentioned in practice for the ratio of e.g. flue-gas desulphurisation. The partial productivities in the abatement production function are assumed positive. Increases in the abatement inputs contribute to an increase in the relative share of abated amount and an absolute increase for a given amount of primary residual. Given the amount of the primary residual from the production stage, and knowing the rate of abatement *A*, both the absolute amounts of the two abatement outputs can be calculated: $\delta z^a = Az,\ z^D = (1-A)z$.

The multiplicative decomposition of primary pollutants and the relative abatement part facilitates focussing on the latter as the endogenous variable of the end-of-pipe abatement activity. It may be assumed that the function $A(.)$ is concave.

Usually abatement is represented by a cost function in the environmental economics textbooks (Førsund and Strøm 1988; Perman et al 2011, see also Rødseth and Romstad (2014, p.119) for a non-parametric application to US electricity generation regulating sulphur emissions). The main advantage is simplification (see Førsund 2017b), but the details of a physical abatement production function (13) are then hidden. Here it is chosen to focus on the relative amount of primary residual that is modified to other forms, e.g. from gas to solid waste. We can also say that there are two outputs generated by the abatement activity; the harmless abatement residual z^a and the remaining amount of the primary residual in its original form.

[31] The abatement stage in Färe et al (2013, p. 112) does not, somewhat awkwardly, show the use of the abated amount explicitly, neither in the definition of their production possibility set (16) nor in their model equations (18), (19) and (20). In Pethig (2006, p. 189) the primary residual seems to be the output of abatement.

Applying the materials balance principle to (13) the abatement activity will add to the total mass of residuals if material inputs are used; the material factors and primary pollutants are now inputs to a production process and the mass is distributed on the output z^a and the secondary pollutant z^D. The total mass of residuals has increased, but the point is that abatement means less mass of the harmful residual; $z^D < z$.

In the environmental efficiency literature, the resources of a firm are often regarded as given, and then increased abatement will imply fewer resources to produce the intended output and thereby decreasing the generation of primary pollutants (see e.g. Martin 1986; Murty et al 2012; Färe et al 2013; Murty and Russell 2016). To do this requires a restriction to be imposed on the availability of inputs (done only in Färe et al 2013). However, this problem is created by the analyst and does not necessarily reflect decisions of a firm having access to markets for inputs to given prices. If it is assumed that abatement is a separate identifiable activity, as e.g. end-of-pipe, and inputs are sourced in markets, there is no reason to assume that abatement resources are taken from the production inputs of a firm. Thus, abatement does not influence the output directly, but increases the cost of production and may then indirectly reduce output and production inputs. It is closer to reality at the micro level not to consider a common resource pool for the production unit, but to regard the activities (11) and (13) as separate “profit centres”.

We recommend to follow this approach and thus avoid constructed trade-offs not embedded in technology. The abatement inputs therefore have a super-index “*a*” to indicate abatement inputs. It may also be the case that there are specific types of abatement inputs, e.g. chemicals and capital equipment, not used in the production process itself. In the case of thermal electricity generation, it is quite usual that abatement activities require electricity as an input. Carbon capture and storage may draw as much as 20 % of the gross production of electricity. However, this electricity can be formally regarded as a bought input so (13) may still be used.

5.2 Imposing a constraint on emission

Environmental regulatory agencies typically prefer direct regulation and not indirect economic instrument. The most common type of direct regulation is to impose an upper limit on discharge of harmful residuals on firms. In order to predict how a firm reacts to direct regulation it is necessary that the firm acts rationally, commonly interpreted as meaning in a private economic sense. It is then standard to assume that the firm starts out being technically efficient and not to be inefficient as was the case discussing the Porter hypothesis in Subsection 2.3.

For simplicity, we consider a single undesirable output only. An environment agency may impose an upper limit z_R^D on the amount emitted from a firm during a specific time period; $z^D \leq z_R^D$. The firm's optimisation problem, cast as a profit maximisation problem, becomes

$$\begin{aligned}
&\text{Max } py - \sum_{j=M,S} q_j x_j - \sum_{j=M,S} q_j^a x_j^a \\
&s.t. \\
&y = f(x_M, x_S) \\
&z = g(x_M, x_S) \\
&\delta z^a = zA(x_M^a, x_S^a) \\
&z^D = z - \delta z^a \\
&z^D \leq z_R^D
\end{aligned} \tag{14}$$

The optimisation problem may be written more compactly as

$$\begin{aligned}
&\text{Max } pf(x_M, x_S) - \sum_{j=M,S} q_j x_j - \sum_{j=M,S} q_j^a x_j^a \\
&s.t. \\
&g(x_M, x_S)(1 - A(x_M^a, x_S^a)) \leq z_R^D
\end{aligned} \tag{15}$$

The necessary first-order conditions are:

$$\begin{aligned}
&pf'_{x_j} - q_j - \lambda g'_{x_j}(1-A) = 0,\ j = M, S \\
&-q_j^a + \lambda z A'_{x_j} = 0,\ j = M, S
\end{aligned} \tag{16}$$

Here λ is the shadow price on the emission constraint. Assuming that the constraint is binding the shadow price shows the gain in profit of marginally relaxing the constraint.

Without the regulation on discharge of residuals, the standard first-order condition is $pf'_{x_j} = q_j,\ j = M, S$; the value of the marginal productivity of a factor is equal to the factor price. With regulation binding the unit factor cost will increase for the material input but decrease for the service input, thus leading to a substitution between the factors. However, costs will go up leading to reduced output. If abatement is used this means that abatement is cheaper than reducing discharge of residuals by only reducing production of the good, and reduction of output will then not be so great as without abatement.

6 Allowing for inefficient operations

6.1 Defining inefficiency

In view of the importance of the materials balance for how to specify a technology based on using material inputs, it might be of interest to expand on the meaning of inefficiency. Inefficiency arises in general when the potential engineering or blue-print technology, the frontier for short, is not achieved when transforming inputs into outputs, assuming that this is feasible.[32] For given desirable outputs too much resource 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 Subsection 2.2). It is the amount of residuals and outputs that have potentials for change, while the a, b, c coefficients and the inputs in Eq. (1) remain the same. The combustion process may be less efficient in converting the raw material into heat, and a different mix of combustion substances may be produced than at efficient operation, e.g., for thermal electricity production based on coal, the mix of substances CO_2, CO, particles, NO_x and ash may differ between inefficient and efficient operations.

Another source of inefficiency is the occurrence of rejects and unintended waste of raw materials, e.g., producing tables of wood, residuals consists 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 4.2.

There is another type of problem in 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

[32] 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 most modern technology (Førsund 2010).

the usual sense of obtaining maximal profit or minimising costs, as modelled in the previous Subsection 5.2. There is no production function formulated for inefficient firms in non-parametric analyses. Introducing behaviour in non-parametric DEA models for a unit it is assumed that frontier technology is used. However, in the real world all firms, also inefficient ones, have to react to e.g. environmental regulation. If firms do know the frontier, how come they end up being inefficient? To appeal to randomness only is not so satisfying. (See e.g. Førsund (2010) for a review of reasons for inefficiency.) 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, is of paramount interest, as exemplified in Subsection 5.2. However, the assumptions in the inefficiency literature reviewed in Subsection 2.3 are made for measuring efficiency, and are not 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 assumed in the efficiency literature, then the unit in question operates on the frontier.

6.2 The production possibility set

The general production possibility set allowing for inefficiency including both desirable and undesirable outputs is:

$$T = \{\, (y,z,x) \,|\, y \geq 0 \text{ and } z \geq 0 \text{ can be produced by } x \geq 0\} \tag{17}$$

Such a definition covers the possibility of both efficient and inefficient operations. The border of the production possibility set is commonly referred to as the frontier and expresses efficient operation. This frontier corresponds to the transformation relation (2) in neoclassical production theory used in Section 3.

The technology set (17) can equivalently be represented by the output set

$$P(x) = \{ (y,z) | x \geq 0 \text{ can produce } y \geq 0 \text{ and } z \geq 0\} \tag{18}$$

In the case of desirable outputs it is obvious that efficient use of resources implies that maximal amount of these outputs are produced for given resources. Concerning undesirable outputs these are automatically kept at a minimum given the maximisation of desirable outputs.

6.3 Weak disposability

In order to operate the single equation model (10) with undesirable outputs avoiding the zero solution for residuals pointed out in Subsection 3.3, restrictions must be placed on the production possibility set.[33] This has typically been done in the axiomatic efficiency literature by imposing weak disposability, a mathematical concept introduced by Shephard (1970), defined as

$$\text{If } (y,z) \in P(x), \text{then } (\theta y, \theta z) \in P(x) \text{ for } 0 \leq \theta \leq 1 \tag{19}$$

This means that along the frontier desirable and undesirable outputs must change with the same (segment-specific in the case of a non-parametric frontier) proportionality factor. No economic or engineering reasoning for this restriction is given in Shephard (1970), but it may resemble the assumption of fixed input-output coefficients in input-output models including pollution, as in the fixed coefficient model of Ayres and Kneese (1969) reviewed in Subsection 2.3 that is backed up by economic reasoning and empirical findings.

Illustrations of weak disposability for output sets, taken from the first illustration of weak disposability of desirable and undesirable outputs in Shephard (1970), are presented in Fig. 3. The desirable output is u_2 and the undesirable is u_1. The trade-off contours for two levels of inputs are shown together with the Leontief (1970) case of a fixed relationship between the two outputs as indicated by the ray $0\bar{u}$.[34] The contour curves starting from the origin (thinner lines that are not part of the efficient frontier according to Shephard) secure the condition of inevitability of positive undesirables when desirable output is positive, termed the null-jointness

[33] In Färe et al (2013, p. 110) it is stated: "which [without a restriction] as pointed out by Førsund [2009] would give us a […] nonsensical result that zero bads can be achieved at no costs […]".

[34] Notice that using input-output type of models does not support the assumption of weak disposability, as is made clear in Fig. 3; the input-output assumption means that there is only a *single* ratio between the good and the bad, not many as illustrated by the two other trade-off curves. However, notice that the Leontief assumption is valid for the point $\bar{u}$ only. Furthermore, weak disposability is not a case of Frisch (1965) output couplings as in Eq. (6).

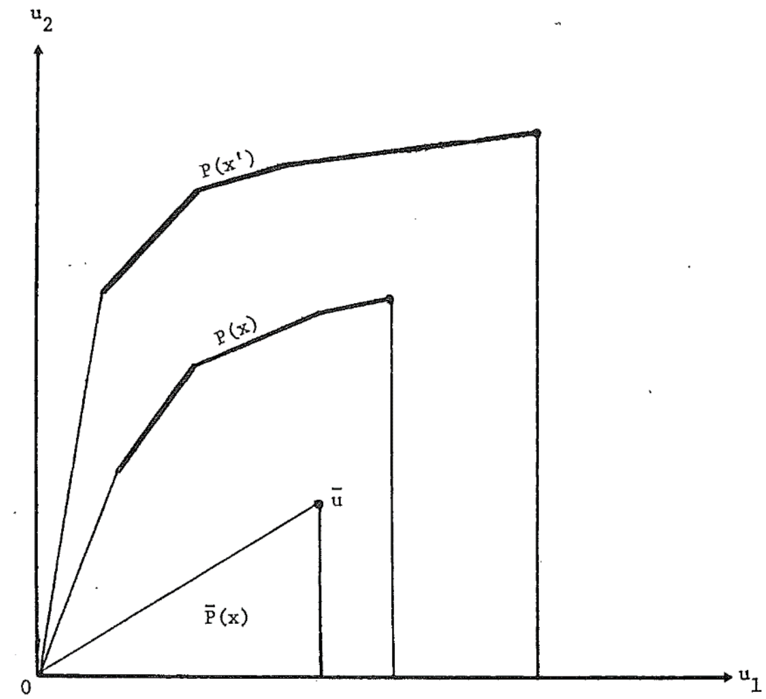

FIGURE 32 (a): OUTPUT SETS FOR A PRODUCTION CORRESPONDENCE WITH WEAK DISPOSAL (u_1 NOT DESIRABLE), $x' \geqq x$

Figure 3. *Illustration of weak disposability*
Source: Shephard (1970, p. 188)

condition in Shephard and Färe (1974) [35].

An explanation of the simultaneous reduction of desirable and undesirable outputs along a trade-off curve often used is that inputs are reallocated to abatement of pollutants[36]. However, it seems rather difficult to both have constant inputs along the curve and to take some inputs away to be used in another activity. If abatement is to take place it must be introduced explicitly, and show the connection between input use and abatement.

A problem with the approach of Shephard to overcome the problem of strong disposability of the residual is the coupling between desirable and undesirable outputs. The situation is that the couplings are between raw materials and the outputs taken place simultaneously. Specifically, the single-equation model using distance functions cannot capture this fact. The popular use of the directional distance function (Chung et al 1997):

[35] Note that Shephard (1970, p. 187) was aware of the fact that production relations need not be of a single-equation type: “It is useful to reiterate at this point that the foregoing assumptions for the production correspondence do not exclude the technology being composed of several processes (or sub-technologies) which are to be jointly planned, as well as situations where joint outputs are inherently involved.”

[36] Färe et al (2008, p. 561) state: [...] “disposal of bad outputs is costly – at the margin, it requires diversion of inputs to ‘clean up’ bad outputs” […]

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

where $P(x)$ is the output production set (18), has the problem that assuming differentiability, as is often done (Färe et al 2013), 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), and the trade-off curve is illustrated in numerous papers by Färe et al (2013) and authors of similar models. However, such a trade-off is not compatible with the material balance.[37]

6.4 Recent attempts to improve the single-equation model

In Rødseth (2017a) there are interesting attempts to reconcile the type of efficiency model used in Chung et al (1997) based on directional distance functions with the materials balance, extending the model with abatement and also some new axioms. (The new model is applied in Rødseth 2016.) However, the model remains a single-equation one. Such a model is based on a trade-off between desirable and undesirable outputs. As shown previously a single-equation model is not compatible with the materials balance, and thus the Rødseth (2017a) model cannot save the weak disposability model of Färe et al. The problem is that both the good and the bad output are arguments in the single directional distance function. Another problem is that the materials balance, that is an accounting identity, is used as a production technology. As explained in Subsection 4.3 (see also Førsund (2017a, Subsection 4.4), the materials balance expressed by Eq. (1) is an accounting identity and cannot explain how residuals are created within a production process. However, introducing axioms of jointness of inputs and outputs are improvements over the assumption of null jointness of the desirable and undesirable outputs.

The new model in Rødseth (2017a) is implemented empirically in Hampf (2017) and compared with Färe et al models applied to the same data. However, choosing the best model based on empirical applications is not the approach recommended in Section 4. Theory should come first.

Abatement is introduced in Rødseth (2014); (2016); (2017b). However, abatement as a production activity is not modelled explicitly. In Färe et al (2013) explicit end-of-pipe

[37] A peculiarity with the trade-off in Färe et al (2013) is that the trade-off occurs with the output for final consumption and the secondary pollutants from the abatement stage, and not between the total output of the good (electricity) and the generation of pollutants in the production stage. However, it is the last trade-off that is the functional trade-off that goes against the materials balance principle in the single-equation model of the production stage.

abatement is added to the production of desirable and undesirable outputs. Inputs to abatement come from a given resource pool by reallocation, and in addition abatement receives part of desirable output as input together with primary pollutants. However, the two distinct production activities are lumped together using a directional distance function with final delivery of desirable outputs, secondary undesirable outputs, and total "source" resources.[38]

A similar two-stage approach is also developed in Hampf (2014). The distinct production activities are as the first stage producing desirable and undesirable outputs, and intermediate desirable output used as input in the abatement production, and at the second stage producing abatement outputs using undesirable output from the first stage as input together with non-polluting inputs and a commonly shared input. A restriction in the form of a material balance is introduced in the first stage, so no production relation proper is used for the undesirable output. The modelling of the production activities of desirable and undesirable outputs remain a single equation that does not satisfy the materials balance.

It should be emphasised that the arguments as formulated in Subsection 3.3 are not only concerning weak disposability, but also strong disposability. It was demonstrated how also strong disposability fails. The point is that single-equation models when material inputs are involved cannot fulfil the materials balance and efficiency conditions for the frontier relations. There must be a clear disentanglement between the modelling of the production of desirable and undesirable outputs.[39] It is the single-equation approach that is at fault, not specifically the imposition of weak disposability. The crucial feature of the Frisch-inspired two-equation model Eq. (11) in Section 4 is just the separate frontier functions for goods and bads.

The single equation model has apparently been successfully applied in the numerous empirical studies found in the literature. The data have seemingly allowed the model to be estimated. However, the ease of obtaining estimates of efficiency does not guarantee that the results are correct. Unfortunately, at the level of abstraction of such models the risk is that a 'false frontier'[40] is estimated, i.e., the data fit a model that goes against the physical law of materials

---

[38] A material balance restriction is mentioned, but not implemented in the empirical model. Weak disposability is assumed.

[39] This is also the message in Murty and Russell (2016, Section 6) stating: […] "the complex real-world trade-offs among inputs and outputs in these technologies cannot be captured by a single functional relation. For example, it is impossible for a single function to capture, simultaneously, the positive relations between emissions and emission-causing inputs and the positive relations between emissions and intended outputs."

[40] This apt expression is due to Barnum et al (2016).

balance principle, and against a fundamental efficiency requirement of a frontier production function.[41]

7 Efficiency measures and their estimation in the multi-equation model

7.1 The production possibility sets of the factorially determined multi-equation model

The multi-equation frontier model (11) with add-on abatement (13) can be straightforwardly extended to include inefficient operations. It remains to show how such a model can be implement empirically.The multi-equation model with abatement allowing inefficiency can be set up using inequalities (with the partial derivatives of the functions as given in (11) and (13)):

$$\begin{aligned} & y \le f(x_M, x_S) \\ & z \ge g(x_M, x_S) \\ & \delta z^a / z \le A(x_M^a, x_S^a) \end{aligned} \tag{20}$$

Following Murty et al (2012) the production possibility sets can formally be written:

$$\begin{aligned} & T_1 = \left\{ (x_M, x_S, y) \middle| y \le f(x_M, x_S) \text{ and } y \ge 0, x_M \ge 0, x_S \ge 0 \right\} \\ & T_2 = \left\{ (x_M, x_S, z) \middle| z \ge g(x_M, x_S) \text{and } z \ge 0, x_M \ge 0, x_S \ge 0 \right\} \\ & T_3 = \left\{ (x_M^a, x_S^a, z, z^a) \middle| \delta z^a \le zA(x_M^a, x_S^a), z^a \ge 0, z \ge 0, x_M^a \ge 0, x_S^a \ge 0 \right\} \end{aligned} \tag{21}$$

The functions *f*(.), *g*(.) and *A*(.) represent the frontier technologies. For given inputs the realised amount of the desirable output may be less than the potential, the primary pollutant may be greater than the potential, and the relative share of abated primary residuals may be less than the potential at each frontier technology, respectively.

7.2 The multi-equation by-product model

The by-product model in Murty et al (2012, p. 122) with abatement (also used in Murty and Russell 2016) has two frontier relations:

[41] Dakpo et al (2016, p. 356) argue that all the different models introduced should be estimated for comparison. As mentioned previously this is also the approach in Hampf (2017). However, in light of the risk of estimation a 'false model', one cannot identify the "best" model in such a way. The only way is to choose the theoretically best model.

$$f(x_M, x_S, y, y^a) = 0$$
$$z^D = g(x_M, y^a) \tag{22}$$

(The notation in Model (11) is used.) The variable y^a is called abatement output, but its functional role is unclear. The partial derivative of the goods in the first relation is assumed positive and the partial derivatives for the inputs are assumed negative. In the second residual-generating equation the partial derivative of the polluting input is assumed positive and the partial derivative of the abatement output negative. The undesirable output z^D is the secondary residual, i.e. the residuals actually emitted to the environment, see Eq. (13). We notice that the residual z^D does not appear in the first relation, and that the desirable good does not appear in the second relation, thus the generation of emissions is independent of intended-output production and usage of non-emission-causing inputs. (This is in accordance of the definition of the emission-generating technology of Murty et al (2012) as shown in Murty and Russell 2016, Section 6, Theorem 1.)

The two production possibility sets can be written:

$$T_1 = \{(x_M, x_S, y, y^a, z^D) \in R_+^5 \,\big|\, f(x_M, x_S, y, y^a) \le 0\}$$
$$T_2 = \{(x_M, x_S, y, y^a, z^D,) \in R_+^5 \,\big|\, z^D \ge g(x_M, y^a)\} \tag{23}$$

The technology set T for the total activity is the intersection of the two subsets; $T = T_1 \cap T_2$. Murty and Russell (2016) combine Murty et al (2012) and Murty (2015) to make the case for an axiomatic foundation for the multi-equation model.

Comparing the frontier models (21) and (23) we see that the Murty et al (2012) model does not conform to the factorially determined multi-output format regarding the residuals-generating relation by specifying the secondary residual as output and materials inputs and abatement output as inputs.[42] How abatement takes place is then rather hidden. End-of-pipe abatement is ruled out, so there must then be some internal adjustment of technology or recycling of raw materials (cf. the chocolate production example in footnote 25). A problem excluding the non-polluting input in the residuals-generating function is that reducing the residual by input substitution, as explained in Subsection 4.2 (see Figure 2), is not reflected in the specification of the residuals relation. However, more seriously, as explained in Subsection 4.1, positive

[42] The multi-equation model in Serra et al (2014) is based on the development in Førsund (2008) (an improved version of this working paper is Førsund 2009) and Murty et al (2012). Both polluting and non-polluting inputs are specified to produce residuals emitted to the environment (see their Eq. (3)), i.e. no abatement is taking place.

marginal productivity as assumed in the first relation in (22), i.e. increasing x_S partially increasing y is usually obtained by utilising raw material better. This then implies less residuals for constant x_M, but the second relation states that only change in x_M can influence the generation of residuals and not changes in x_S. This seems a drawback and goes against knowledge about substitution (cf. footnote 25).

The Frisch scheme of joint production separating outputs and having the same set of inputs as arguments in all production functions is a well-argued scheme, and is especially so in our case of simultaneous production of both goods and bads, because it is just the inputs that are used producing a desirable output that also generates the nondesirable outputs.

In accordance with theorems in Murty and Russell (2016) a strategy for efficiency measures is to introduce separate measures for each of the different activities. Then the Farrell (1957) technical measures of efficiency may be used (these are equivalent to distance functions), giving us three types of measures based on relative distance from best-practice frontiers: desirable output efficiency E_y, primary residual efficiency E_z, and abatement efficiency $E_{A,}$ all three measures restricted to be between zero and one. Efficiency measures can in general be either input oriented or output-oriented. In our setting output orientation seems to be a natural choice.

7.3 The 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. However, forming the residual production possibility set is not quite standard due to the negative sign of the derivative of the service input.

In the three DEA optimisation problems below for unit i among N units in total, variable returns to scale functions are specified (for simplicity a single output and two inputs are specified):

$$
\begin{aligned}
& 1/E_{y_i} = Max_{\lambda,\theta}\,\theta \\
& s.t. \\
& \sum_{j=1}^{N} \lambda_j y_j \geq \theta y_i, i = 1,...,N \\
& \sum_{j=1}^{N} \lambda_j x_{kj} \leq x_{ki},\ i = 1,...N,\ k \in M, S \\
& \sum_{j=1}^{N} \lambda_j = 1, \lambda_j \geq 0, \theta \geq 0
\end{aligned}
\tag{24}
$$

The optimal solution of the weighted sum of observed outputs and inputs of the efficient units spanning the frontier are the output and input values at the frontier segment for the radial projection of observations (y_i, x_i), (z_i,x_i).

Remember that we have assumed that the function $g(.)$ is convex when formulating the primary residuals efficiency measure:

$$
\begin{aligned}
& E_{z_i} = Min_{\lambda',\varphi}\,\varphi \\
& s.t. \\
& \sum_{j=1}^{N} \lambda'_j z_j \leq \varphi z_i, i = 1,...,N \\
& \sum_{j=1}^{N} \lambda'_j x_{kj} \geq x_{ki},\ i = 1,...N,\ k \in M, S \\
& \sum_{j=1}^{N} \lambda'_j = 1, \lambda_j \geq 0, \varphi \geq 0
\end{aligned}
\tag{25}
$$

For the two first frontier production relations in theoretical models in (20), a unit that is on the frontier for the intended output, will also be residual-efficient because of the combined effect of the materials balance and the efficiency assumptions of the functions. (All points on isoquants illustrated in Fig. 2 are by definition efficient.) However, the estimation of the border of a polyhedral set implies typically a negative bias of the frontier technology compared with the unknown theoretical model. It may then be the case that best practice points spanning the set may not be efficient within the true unknown technologies (20). A best practice unit in the problem (24) in desirable output production may not be efficient in undesirable output production in the problem (25), and *vice versa*.

The materials balance identity is not specified for the efficiency problems above. It holds for the two problems together, not (24) and (25) separately, but only if the polyhedral model is the true theoretical model. The concern with the materials balance estimating a non-parametric frontier using DEA is then that projections to the frontier in problems in (24) and (25) of inefficient points may not satisfy the relevant materials balance conditions. The projection points for inefficient observations within the *N* units are:

$$\begin{aligned}&\sum_{j=1}^{N}\lambda_j y_j, \sum_{j=1}^{N}\lambda_j x_{kj}, k\in M,S\\&\sum_{j=1}^{N}\lambda'_j z_j, \sum_{j=1}^{N}\lambda'_j x_{kj}, k\in M,S\end{aligned} \tag{26}$$

These points are not observations, but constructs of the analyst. Assuming projection points being on efficient faces, i.e. all the inequalities in (24) and (25) hold as equalities, it may be tempting to say that the materials balance restriction for the frontier projection of unit *i* is

$$a\sum_{j=1}^{N}\lambda_j x_{Mj} \equiv b\sum_{j=1}^{N}\lambda_j y_j + c\sum_{j=1}^{N}\lambda'_j z_j \Rightarrow ax_{Mi} = b\theta y_i + c\varphi z_i \ , i=1,...,N \tag{27}$$

However, this is only correct if the border of the estimated polyhedral set is the true frontier. The materials balance condition in (27) can be checked by inserting the optimal solution for the projected residuals point solving (25) and the solution for the desirable output solving (24) into (27), thus exposing difference between the left-hand and right-hand of (27). (Notice that we must have $\sum_{j=1}^{N}\lambda_j x_{Mj} = \sum_{j=1}^{N}\lambda'_j x_{Mj}$ by definition.)

The expansion of y_i ($\theta \geq 1$) must be counteracted by the reduction in z_i ($0 \leq \varphi \leq 1$). However, without imposing this restriction on projection points on the frontier there may be no guarantee that this is fulfilled. It may be a problem that the frontier output projection points come from two different models, while the inputs are the same.[43] Regarding weakly efficient faces there will be slacks on constraints yielding zero shadow prices. However, the set of these units may be different between the models. Material inputs with zero shadow prices not impacting the efficiency scores must also be counted in the materials balance.

Imposing a materials balance constraint on projection points as in Rødseth (2017a) in the single-equation model is not straightforward in the multi-equation model. However, given the possibility of biased estimation using DEA it may not be desirable to force the materials balance condition upon synthetic projection points possibly changing the estimates of efficiency scores.

[43] In Dakpo et al (2017) the problem with connecting the two problems is suggested to be solved by imposing equality between the frontier inputs using a model minimising the amount of the undesirable output.

In the non-parametric estimation model for abatement efficiency the observed amount of primary residual for unit *i* is now given from the production stage and not appearing in the model determining the frontier relative degree of abatement due to the assumption of multiplicative decomposition of the abatement function in the first relation in (13):[44]

$$
\begin{aligned}
&1/E_{A_i} = Max_{\lambda'',\phi}\,\phi \\
&s.t. \\
&\sum_{j=1}^{N}\lambda_j'' A_j \ge \phi A_i,\, i=1,\dots,N \\
&\sum_{j=1}^{N}\lambda_j'' x_{kj}^{a} \le x_i^{a},\, i=1,\dots,N,\ k\in M,S \\
&\sum_{j=1}^{N}\lambda_j'' = 1, \lambda_j'' \ge 0, \phi \ge 0
\end{aligned}
\tag{28}
$$

Once we have the solution for the relative abatement the absolute amounts of abatement residuals and secondary residuals for a projection of an inefficient unit to the frontier can be calculated.

For the materials balance to hold in the models in (22) the relations must be a “good” representation of the production relations (see Subsection 2.2). A problem is that it is quite difficult to verify the goodness. One may doubt that the piecewise linear frontiers, or the faceted structure of the borders of the production possibility sets, meet a goodness criterion. There is also the problem of the variables with zero shadow prices generating faces not of full dimension regarding forming projection points of inefficient observations on the frontier. However, forming the materials balance all variables containing mass must be counted, also for units with zero shadow prices.

The term environmental efficiency or eco-efficiency is used somewhat differently in the literature and is not used in the efficiency measures introduced above. One reason for this is that one would expect that environmental efficiency has something to do with what happens within the environment in terms of degradation of environmental qualities, cf. points (c) and (d) in Subsection 2.1. However, the most common notion of environmental efficiency is showing the potential relative reduction in emission of residuals. The so called unified approach in Sueyoshi and Goto (2010) and the average measure over activities used in Murty et al (2012) have the drawbacks that they combine different types of measures without realising that such

[44] Hampf (2014) also solves separate optimisation problems for the production stage and the abatement stage, but this is done by minimising the weighted emissions in the two stages.

aggregation depends on being able to compare the measures in the same unit; aggregating percentages is not the solution.

However, for policy purposes the individual measures above provide most valuable information for designing specific direct regulations or indirect economic instruments.

8 Conclusions

The introduction of the materials balance in the environmental economics literature (Ayres and Kneese 1969) heralded a new approach to modelling the interactions between the production of desirable outputs and the natural environment. The materials balance tells us that mass (and energy) in an economic activity cannot disappear, but only takes on different forms. Surveying the use of the materials balance 30 years after Ayres and Kneese (1969) pioneered the concept within environmental economics, Pethig (2003), from a standpoint of ecological economics, complains that the materials balance has not been used to the extent it warrants.

However, the position in this paper (supported by Murty et al 2012; Murty and Russell 2016) is that the materials balance is important when picking the model to use. The materials balance is an accounting identity and cannot give information about specific technologies explaining the transformation of resources to desirable and undesirable outputs, so an active use of the materials balance condition may not be necessary if the right model is picked. In addition, at the aggregation level the models are usually formulated on, it may be difficult to represent all the physical quantities involved. Data accuracy is also a question.

In production activities involving material inputs, the simultaneous generation of desirable outputs and residuals as undesirable outputs, the latter turning up as pollutant in the natural environment, must be captured in a sufficiently realistic way. Classical and neoclassical economists were concerned with production of waste and have many interesting observations that should be utilised. In the efficiency literature the last decades, the most popular approach to empirical efficiency studies of simultaneous production of ‘goods’ and ‘bads’ has been to apply a single-equation model. To assume a mathematical property of weak disposability of the production possibility set allowing for inefficient observations, has then been seen necessary. This property blocks the maximal assortment case of using all resources on desirable outputs resulting in zero emission of residuals.

However, a main result of the paper is that a functional trade-off between desirable and undesirable outputs, as implied by the weak disposability model, is not theoretically compatible with the materials balance and efficiency in resource utilisation. Notice that it was shown in Subsection 3.4 that also strong disposability of outputs is not compatible with this trade-off. But more importantly, this implies further that it is the format of a single equation model to tackle efficiency measurement when producing both desirable and undesirable outputs using material inputs, which is at fault, not weak disposability as such. The main message of the paper is that the single-equation model, which has been almost exclusively used in the literature about inefficiency when dealing with material-based bads, is not able to conform to the materials balance and efficiency requirements on frontier relations. A multi-equation model is required separating production relations of desirable and undesirable outputs.

A multi-equation model, based on 'classical' joint production theory, that theoretically satisfies the materials balance and frontier efficiency requirements, is developed in the paper, and shown to function well both in an efficient and in an inefficient world. It is also straightforward to understand the mechanisms of the model without mathematical knowledge necessary to relate to rather complex axiomatic approaches.

The model proposed in the paper can straightforwardly extended to cover abatement efforts of the end-of-pipe type.

The single-equation models based on weak disposability have had a good run for decades. However, as happens with technologies when experiencing technical progress in an economy also happens to models: they become outmoded and should then substituted with better ones; the multi-equation models. As Ragnar Frisch expresses it: […] "we disregard a model world as soon as we get upon the idea of another model world which 'smells' better" (Frisch 2010, p. 33).

It was conjectures that single-equation models cannot comply with the materials balance, and furthermore that a specific type of a multi-equation model can obey the materials balance. Further research will be focussed on substantiating more formally this conjecture.

Other research tasks are implementing empirically the type of multi-equation models including abatement proposed in this paper. More challenging are introducing dynamics not only involving embodied technologies, but also dynamic analyses of how inefficiencies are reduced due to pressure of environmental regulation, i.e. tackling the Porter hypothesis in a dynamic framework.

As underlined in the paper generation of residuals occurs when material inputs are used. Typical industries studied in the environmental efficiency literature are thermal generation of electricity and pulp and paper. In addition, we have material throughput industries such as oil refineries, other chemicals, steel and iron, aluminium, and other energy-intensive industries, as well as food processing and cement. A common feature for all these industries is that much of the key technologies are embodied in the capital equipment.

The pace of technical progress depends on investments in new technology. A consequence is that care must be exercised when having observation for several vintages of plants when using DEA to estimate the best practice frontiers. The risk is great for estimating a 'false frontier', in the sense that there may be a mix of plants of different vintages spanning out the frontier. An efficiency measure may then give a false picture of obtainable improvement (Førsund 2010; Belu 2015, point to some related problems). Developing more appropriate models for tackling vintage structures when studying environmental efficiency is a challenge for future research.[45]

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

Arabi B, Munisamy S and Emrouznejad A (2015) A new slacks-based measure of Malmquist–Luenberger index in the presence of undesirable outputs. *Omega* 51 (March), 29-37, http://dx.doi.org/10.1016/j.omega.2014.08.006

Aparicio J, Pastor JT and Vidal F (2016) The directional distance function and the translation invariance property. *Omega* 58 (January), 1-3, http://dx.doi.org/10.1016/j.omega.2015.04.012

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

Barnum D, Coupet J, Gleason J, McWilliams A and Parhankangas A (2017) Impact of input substitution and output transformation on data envelopment analysis decisions. *Applied Economics* 49(15), 1543-1556. Published online 24 August 2016. http://dx.doi.org/10.1080/00036846.2016.1221042

[45] Hampf and Rødseth (2015) find that most of the efficiency differences in U.S. power plants measured by electricity generation using coal can be explained by the age of plants.

Baumgärtner S, Dyckhoff H, Faber M, Proops J, and Schiller J (2001) The concept of joint production and ecological economics. *Ecological Economics* 36(3), 365-372

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* (second edition). Cambridge: Cambridge University Press (first edition 1975, Washington: Prentice Hall)

Belu C (2015) Are distance measures effective at measuring efficiency? DEA meets the vintage model. *Journal of Productivity Analysis* 43(3), 237-248

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

Chambers RG (1988) *Applied production analysis*. Cambridge: Cambridge University Press

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

Coase R (1959) The federal communications commission. *Journal of Law and Economics* 2 (October), 1-40

Cropper ML and Oates WE (1992). Environmental economics: a survey. *Journal of Economic Literature* 30(2), 675-740

Dakpo KH, Jeanneaux P, 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

Dakpo KH, Jeanneaux P, 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. https://doi.org/10.1093/erae/jbw013

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. Amsterdam: Elsevier

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

Farrell MJ (1957) The measurement of productive efficiency of production. *Journal of the Royal Statistical Society, Series A*, 120(III), 253–281

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

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

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

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: *Handbook of natural resource and energy economics, vol. I*. Kneese AV and Sweeney JL (eds.). Chapter 8, pp. 325-341. Amsterdam: Elsevier Science Publishers BV

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 (Re-printed as Memorandum No 10/2016 Department of Economics University of Oslo, available 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. New Delhi- Heidelberg-New York-Dordrecht-London: Springer

Førsund FR (2011) Industrial ecology: reflections of an environmental economist. In Batabyal AA and Nijkamp P (eds.). *Research tools in natural resource and environmental economics*. Chapter 14, pp. 423-455. Singapore: World Scientific

Førsund FR (2017a) Multi-equation modelling of desirable and undesirable outputs satisfying the materials balance. *Empirical Economics*. Published online 12 January 2017, DOI 10.1007/s00181-016-1219-9

Førsund FR (2017b) Productivity measurement and the environment. In: *The Oxford handbook of productivity analysi*s, Grifell-Tatje E, Lovell ACK, Sickles R (eds.), Chapter 8. Forthcoming. Oxford: Oxford University Press

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

Førsund FR and Strom S (1976) The generation of residual flows in Norway: an input-output Approach. *Journal of Environmental Economics and Management* 3(2), 129-141.

Førsund FR and Strøm S (1988) *Environmental economics and management: pollution and natural resources*. London: Croom Helm. (Also re-published in the series Routledge Revivals, 2011. Abingdon-New York: Routledge.)

Hampf B (2014) Separating environmental efficiency into production and abatement efficiency: A nonparametric model with application to US power plants. *Journal of Productivity Analysis* 41(3), 457-473

Hampf B (2017) Measuring inefficiency in the presence of bad outputs: does the disposability assumption matter? *Empirical Economics*. Published online 07 January 2017, doi:10.1007/s00181-016-1204-3

Hampf B and Rødseth KL (2015) Carbon dioxide emission standards for U.S. power plants: an efficiency analysis perspective. *Energy Economics* 50(1), 140-153

Jevons WS (1965) *The theory of political economy* (first published 1871, reprinted 1965). New York: Augustus M. Kelley

Kneese AV, Ayres RU and d'Arge RC (1970) *Economics and the environment. A materials balance approach*. Resources for the Future, Washington. Baltimore: Johns Hopkins Press

Kohli U (1983) Non-joint technologies. *The Review of Economic Studies* 50(1), 209-219

Kumbhakar SC and Malikov E (2017) Good modeling of bad outputs: editors' introduction. *Empirical Economics*. First Online: 27 March 2017. DOI 10.1007/s00181-017-1231-8

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

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

Lauwers L (2009) Justifying the incorporation of the materials balance principle into frontier-based eco-efficiency models. *Ecological Economics* 68(8), 1605–1614

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

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. Amsterdam-London: North-Holland

Mäler K-G (1974) *Environmental Economics: A Theoretical Inquiry*. Baltimore: The Johns Hopkins Press

Martin RE (1986) Externality regulation and the monopoly firm. *Journal of Public Economics* 29(3), 347–362

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 (2016) Modeling emission-generating technologies: reconciliation of axiomatic and by-production approaches. *Empirical Economics*. Published online 28 September 2016, DOI 10.1007/s00181-016-1183-4

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

Nadiri I (1987) joint production. In: *The new Palgrave. A dictionary of economics*. Eatwell J, Milgate M, Newman P (eds.). Vol 2 (E to J), pp.1028-1030. London, New York, Tokyo: Macmillan

Palmer K, Oates WE, and Portney PR (1995) Tightening environmental standards: the benefit-cost or the no-cost paradigm? *Journal of Economic Perspectives* 9(4), 119-132

Pasinetti LL (1980) Introductory note: joint production. In: *Essays on the theory of joint production*, Pasinetti LL (ed.). London and Basingstoke: Macmillan.

Perman R, Ma Y, Common M, Maddison D and McGilvray J (2011) *Natural resources and environmental economics* (4th edition). Harlow: Pearson Education Limited (First edition 1996, Longman Group Limited)

Pethig R (2003) The 'materials balance' approach to pollution: its origin, implications and acceptance. University of Siegen, *Economics Discussion Paper* No. 105-03, 2003

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

Pigou AC (1920) *The economics of welfare*. London: Macmillan

Porter ME (1991) America's green strategy. *Scientific American* April, 168

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 CS and Spofford WO Jr (1972) A quantitative framework for residuals management decisions. In: Kneese AV and Bower BT (eds.). *Environmental quality analysis: theory and method in the social sciences*, pp. 115-179. Baltimore-London: The Johns Hopkins Press

Rødseth KL (2013) Capturing the least costly way of reducing pollution: A shadow price approach. *Ecological Economics* 92(August), 16-24

Rødseth KL (2014) Efficiency measurement when producers control pollutants: a non-parametric approach. *Journal of Productivity Analysis* 42(2), 211–223 (DOI 10.1007/s11123-014-0382-2)

Rødseth KL (2016) Environmental efficiency measurement and the materials balance condition reconsidered. *European Journal of Operational Research* 250(1), 342-346, http://dx.doi.org/10.1016/j.ejor.2015.10.061

Rødseth KL (2017a) Axioms of a polluting technology: a materials balance approach. *Environment and Resource Economics*. 67(1), 1-22. DOI 10.1007/s10640-015-9974-1

Rødseth KL (2017b) Environmental regulations and allocative efficiency: application to coal-to-gas substitution in the U.S. electricity sector. *Journal of Productivity Analysis* 47(2), 129-142, DOI 10.1007/s11123-017-0495-5

Rødseth KL and Romstad E (2014) Environmental regulations, producer responses, and secondary benefits: carbon dioxide reductions under the acid rain program. *Environmental Resource Economics* 59(1), p. 111-135, DOI 10.1007/s10640-013-9720-5

Serra T, Chambers RG and Oude Lansink A (2014) Measuring technical and environmental efficiency in a state-contingent technology. *European Journal of Operational Research* 236(3), 706-717

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

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

Sraffa P (1960) *Production of commodities by means of commodities*. Cambridge: Cambridge University Press

Sueyoshi T and Goto M (2010) Should the US clean air act include CO_2 emission control? Examination by data envelopment analysis. *Energy Policy* 38(10), 5902–5911

Sueyoshi T, Yuana Y and Goto M (2017) A literature study for DEA applied to energy and environment. *Energy Economics* 62(February), 104-124

Whitcomb DK (1972) *Externalities and welfare*. New York and London: Columbia University Press