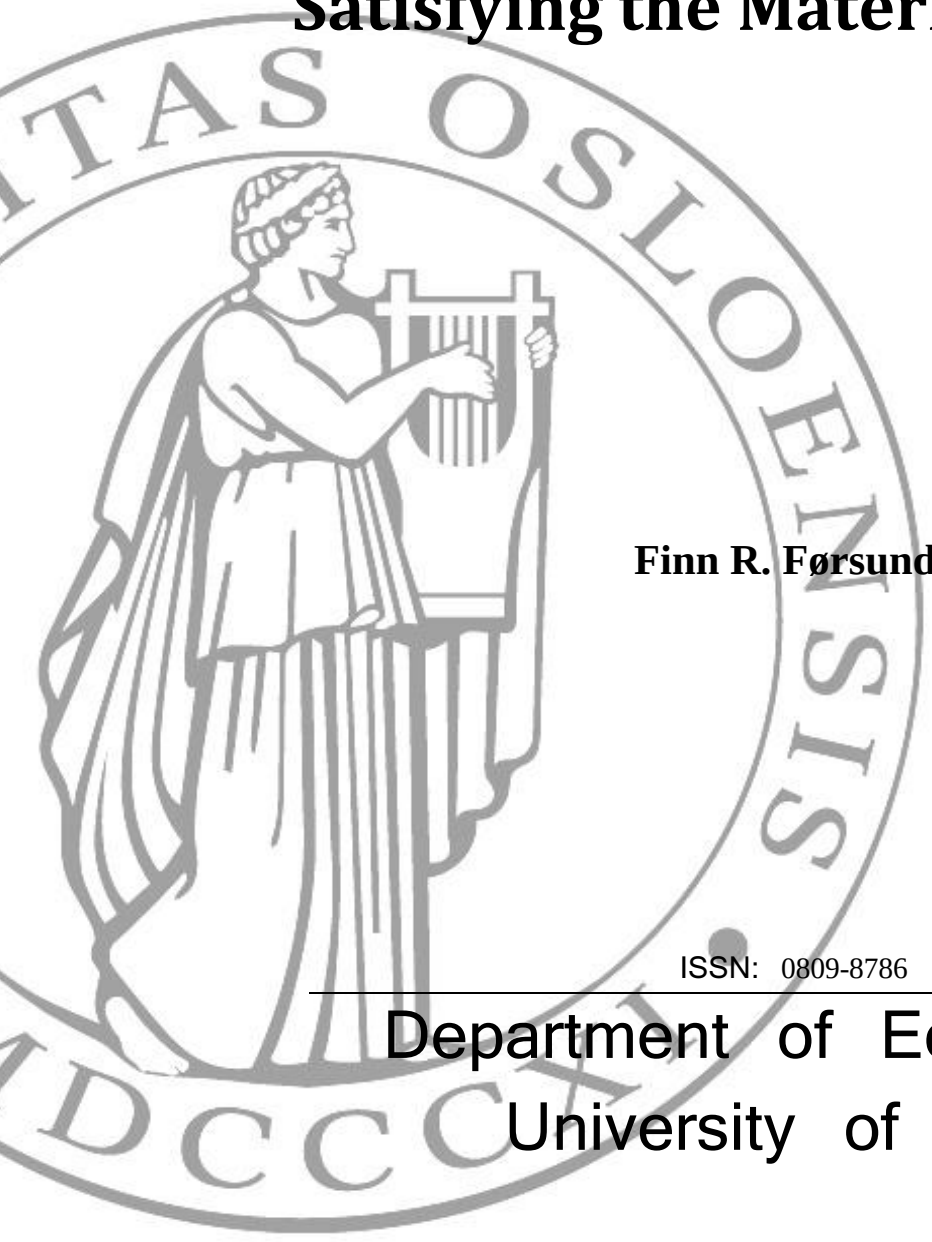
MEMORANDUM

No 03/2016

Multi-equation modelling of Desirable and Undesirable Outputs Satisfying the Material Balance



Finn R. Førsund

ISSN: 0809-8786

Department of Economics
University of Oslo

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

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

In co-operation with
The Frisch Centre for Economic Research

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

Last 10 Memoranda

No 02/16	Ingrid Hjort *Potential Climate Risks in Financial Markets: Report from a workshop, January 20, 2016*
No 01/16	Ingrid Hjort *Potential Climate Risks in Financial Markets: A Literature Overview*
No 22/15	Geir B. Asheim and Frikk Nesje *Destructive intergenerational altruism*
No 21/15	Rolf Golombek, Alfonso A. Irarrazabal, Lin Ma *OPEC's market power: An Empirical Dominant Firm Model for the Oil Market*
No 20/15	Moritz A. Drupp, Mark C. Freeman, Ben Groom and Frikk Nesje *Discounting Disentangled*
No 19/15	Simen Markussen and Marte Strøm *The Effects of Motherhood*
No 18/15	Marit Linnea Gjelsvik, Ragnard Nymoen, and Victoria Sparrman *Have Inflation Targeting and EU Labour Immigration Changed the System of Wage Formation in Norway*
No 17/15	Geir B. Asheim and Ivar Ekeland *Resource Conservation across Generations in a Ramsey-Chichilnisky Model*
No 16/15	Nina Drange, Tarjei Havnes and Astrid M. J. Sandsør *Kindergarten for All: Long-run Effects of a Universal Intervention*
No 15/15	Paolo G. Piacquadio *The Ethics of Intergenerational Risk*

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http://www.sv.uio.no/econ/english/research/unpublished-works/working-papers/

Multi-equation modelling of Desirable and Undesirable Outputs Satisfying the Material Balance

by

Finn R. Førsund*

Department of Economics, University of Oslo

Abstract: The modelling of the interaction between production activities and the natural environment requires formulating a multioutput production function for intended and unintended outputs. The key feature when modelling joint production of intended outputs and unintended residuals is that the latter stem from the use of material inputs. A multi-equation model building on the factorially determined multi-output model of classical production theory satisfies the material balance that tells us that the mass contained in inputs cannot disappear, but must turn up in the desirable outputs or end up as residuals. Each of the intended outputs and the residuals are functions of the same set of inputs. Some problems with the single equation models most often found in the literature are demonstrated. Abatement activity in the form of end-of-pipe is added and an optimal planning solution is derived using the concept of an environmental damage function for the impact of discharge of residuals into the natural environment. It is shown that the traditional environmental policy instruments, like direct regulation restricting the amount of undesirable residuals discharged to the environment, a Pigou tax on pollutants, and cap and trade all function well. Extending the multi-equation model to allow for inefficiency, three efficiency measures are introduced; desirable output efficiency, residuals efficiency and abatement efficiency. It is demonstrated that these measures can be estimated independently using the DEA model.

JEL classification: D62, Q50

Keywords: Desirable and undesirable outputs; Materials balance; Factorially determined multioutput production; Abatement; Efficiency measures; DEA

* I am indebted to Rolf Färe, Benjamin Hampf and Kenneth Løvold Rødseth for comments improving the paper.

1. Introduction

On a backdrop of a long tradition within economics of treating environmental problems as a case of externalities[1], giving the modelling a somewhat innocent or non-urgent flair, the publishing of the seminal paper in economics of Ayres and Kneese (1969), coining the phrase *materials balance*, heralded a new view within economics of the pervasiveness and seriousness of environmental pollution (for a book-length exposition of the approach see Kneese et al., 1970). The first law of thermodynamics tells us that matter (and energy) cannot disappear. If all the material inputs into an activity are not embedded in the products the activity is set up to deliver, then the difference must be contained in residuals discharged to the environment. If we weigh the inputs employed in an activity, including non-paid factors like oxygen from the air, and weigh the products that are the purpose of activities, the difference is the residuals that may turn out to be polluting the natural environment. The concept of materials balance underlines the inevitability of residuals generation when employing material resources.

The problem of pollution as a by-product of economic activity is a major topic in contemporary environmental economics, ranging from global warming due to generation of greenhouse gases to local air-water- and land quality deteriorations due to emission of different polluting substances. When modelling environmental – economic interactions it is important to capture the main features of such interactions obeying fundamental physical laws. However, according to Pethig (2003) investigating the use of the insights in Ayres and Kneese (1969), the materials balance approach has been step motherly treated in the literature.

The possibility of inefficient operations has not been stressed in general environmental economics. The generation of residuals occurs typically within technically efficient activities of production (and consumption). The foundation of the efficiency literature is based on the assumption of the existence of inefficiencies of economic activities. The research strand has developed from the seminal paper by Farrell (1957) on definitions of efficiency measures and the use of the concept of an efficient frontier production function and seminal papers on estimating parametric functions (Aigner et al 1977)) and non-parametric frontiers using linear

[1] See Mishan (1971) for a review of the earlier externalities literature and Fisher and Peterson (1976), Cropper and Oates (1992) for reviews of the literature covering the 70-ies and 80-ies decades.

programming (Charnes et al 1978). The axiomatic approach to specifying properties of non-parametric production technologies started with Shephard (1953; 1970).

Recognising pollutants as unavoidable by-products of economic activity Färe et al. (1986); (1989) were (to our knowledge) the first to introduce undesirable outputs in an empirical model covering inefficiency by proposing to impose the property of *weak disposability*, that was introduced in Shephard (1970), on the production possibility set for intended outputs and by-products. These two papers[2] have spawned a strand of research, applying the same assumption, followed up also by the originators (see Färe et al., 1996; Färe et al., 2001; Färe et al., 2004; Färe et al., 2005; Färe et al., 2014), extending the approach to directional distance functions (Chung et al., 1997). Papers applying weak disposability have been published in a wide range of journals like Ecological Economics, Energy Economics, Journal of Econometrics, Journal of Environmental Management, Journal of Productivity Analysis, Journal of Regional Science, Resources and Energy, Resource and Energy Economics, Review of Economics and Statistics, among others.

However, a characteristic of the inefficiency literature dealing with both desirable and undesirable outputs has been that there were hardly any traces of insights from environmental economics on how to formulate the production model, the materials balance being especially neglected. But this has changed in some of the recent papers (see e.g. Coelli et al 2007; Färe et al 2013; Hampf 2014; Rødseth 2015).[3]

The purpose of this paper is to present a most simple model (building on Førsund 1972; 1973; 1998; 2009) satisfying the essentials of environment – economic interactions and obeying the material balance. Furthermore, this model is extended to covering inefficient operations and thus efficiency measures involving environmental aspects can be explored. The approach is an alternative to using inefficiency models extended to also including undesirable outputs based on the assumption of weak disposability of desirable and undesirable outputs.

The alternative model specifies the generation of residuals simultaneously with producing desirable outputs using two types of equations; one type relating each of the desirable outputs to the same set of inputs but allowing for different production functions (this is the factorially determined multi-output system introduced in Frisch 1965). Each residual, or undesirable output, is also generated by the same set of inputs as the desirable outputs to extend the Frisch

[2] The papers have 59 and 440 citations, respectively, according to the Web of Science, as of 09.12.2015.
[3] See also the extensive review in Dakpo et al (2016) of different approaches to modelling.

scheme. This model obeys the material balance and abatement of the end-of-pipe type can be added straightforwardly.

The focus will be on production activities within firms only. However, consumer activities modelled as household production functions can also be studied using the same type of model. Activities such as heating/cooling of homes, preparing food, generating wastes such as solids and liquids, etc. are processes dealing with materials an energy and thus must also obey material and energy balances.

The paper is organised as follows. Section 2 introduces the materials balance equation, and Section 3 points to serious problems of the single equation model most commonly used in the literature. Section 4 introduces the multiple equation model of Frisch (1965) and specifies the factorially determined multi-output production model where each residual or undesirable output is generated by the same set of inputs as the desirable outputs. This model satisfies the materials balance. The impact and nature of technical change in such a model is discussed. In Section 5 abatement of the end-of-pipe type is introduced as an extension to the model. Direct and indirect policy instruments that can realise the optimal social solution are studied in Section 6. Introducing inefficiency, some problems with the weak disposability efficiency model are taken up in Section 7. The factorially determined multi-output production model is extended to cover inefficient operations in Section 8, and schemes to estimate all relations involved and efficiency measures are also developed. Section 9 concludes.

2. The material balance equation

Assuming that one or more of the inputs x to a production process of a unit consists of physical mass, this mass will not disappear during the production process, but must either be contained in the products y being the intended outputs, or become residuals z emitted to the external environment. Thus, a material balance exists for each observation (generated by a specific technology), and a, b, c are coefficients converting units of inputs x and desirable outputs y and residuals z to a common mass unit:

$$ax \equiv by + cz > 0 \qquad (1)$$

For each production unit we have an accounting identity for the use of materials contained in the input *x*; the material can be part of the goods *y* or contained in the bads *z*. The relation holding as an identity means that it must hold for any accurately measured observation, being efficient or inefficient. The relation is not a production function, but serves as a restriction on specifications of these.

We simplify for convenience by operating formally with only a single input, output and residual respectively, but generalisation to multiple inputs, outputs and residuals is a straightforward summation done over the type of each variable using weight as the common unit. Therefore we will often use the plural form for the variables in the text. There may be several types of residuals generated by the same raw material, and there may be several types of both raw materials and desirable outputs. Using the total weight of raw materials, outputs and residuals the coefficients and *a*, *b*, *c* are not necessary, but if we want to specify the mass balance for each type of residual, like carbon or sulphur, unit coefficients for each of the three types of variables are convenient. A residual emitted to air from a combustion process may contain materials supplied by the air and not contained in the material input, like SO_2, and in that case the coefficient *c* is the unit of sulphur contained in SO_2. Thermal electricity generation using coal *x* containing *ax* amount of carbon, but zero amount of carbon is contained in electricity *y* i.e. $b = 0$, and *cz* of carbon is contained in the residual *z* that is CO_2 (if all carbon in the coal appears in the CO_2 then $ax = cz$). Assuming that inputs and outputs, respectively, are homogenous across *n* units the coefficients *a* and *b* must be equal by definition for each type of input and output; we must have $a_j = a \forall j, b_j = b \forall j,\ j = 1,...,n$. The mass coefficient *c* is also the same across units *j*, assuming an efficient operation, if the material residuals created by combustion processes and determined by physical/chemical laws appear in other combinations than in the materials[4]. The need for coefficients is not general. Considering the production of wooden furniture, say a table, using electricity as the only energy source, the residuals also consists of wood and all three coefficients are unity.

[4] However, in the case of thermal electricity production based on coal the formation of gases in the combustion may be influenced by the temperature, lower temperature reducing the *c*-coefficient for CO_2, and increasing the carbon in the ash, but also reducing the *c* coefficient for NOx and thus being one way of reducing emission of both types of gaseous residuals for given input of coal. However, electricity production is then also reduced, so reducing combustion temperature is one way of abating harmful emissions at the cost of output. The total amount of emission is the same equal to *a*x summing over all types of substances contained in coal.

3. The transformation relation

The standard way of representing a multi-output, multi input production activity involving residuals is to use a transformation relation in the vectors *y*, *z*, *x* representing desirable outputs, undesirable outputs and inputs, respectively

$$F(y,z,x)=0, \frac{\partial F}{\partial y}>0, \frac{\partial F}{\partial z}>0, \frac{\partial F}{\partial x}<0 \qquad (2)$$

The function $F(y,z,x)$ is the transformation function and the conventional signing of the derivatives defines *y* and *z* as outputs and *x* as inputs. The conditions on the derivatives imposing monotonicity on the functions imply that all variables are strongly disposable.

A relation $F(y,x)=0$ with desirable goods only is in general defined such that each element of *y* is maximised for fixed values of the other outputs and a given input vector *x*, and some standard properties of shape are assumed for the function *F*(.). When including undesirable variables *z* a question is how the transformation relation should be defined. There are four possibilities:

i) Treat the undesirable outputs in the same way as desirable outputs, i.e. $F(y,z,x)=0$ is such that each element of *y* is maximised for fixed values of the other two types of outputs, and each element of *z* is also maximised for fixed values of the other outputs and a given input vector *x*.

ii) Keep the undesirable outputs at constant levels and maximise each element of *y* for given values of the other *y* variables and a given input vector *x*.

iii) Maximise each element of the *y* vector for fixed values of the other desirable outputs, given values of undesirable outputs and given values of the inputs, and minimise for each element of the undesirable vector keeping the other undesirable values constant, as well as all desirables outputs and all inputs.

iv) Neglect the undesirable outputs and focus on the maximal value of each element of the vector *y* of desirable outputs for a given vector of inputs.

The production activity in question is set up in order to produce desirable outputs using inputs. The undesirable outputs are byproducts of the production process and assumed to be discharged to one of the three receptors of Nature; air, water and land. We assume that the

byproducts are generated due to one or more inputs being material (i.e. we neglect non-material external effects). It is then the case that the undesirable byproducts will consist of materials that restrict the possible volume of desirable outputs. Therefore in setting up an efficient transformation of inputs to desirable outputs as little as possible of the material inputs should be wasted on the undesirable outputs. If the transformation can be modelled using a single transformation relation this relation has in general a maximal degree of assortment (Frisch 1965) (details are set out in Section 4), implying that there should be no production of undesirable outputs given that the objective is to produce a maximal amount that is technically feasible. This supports point *iv)* as the relevant way of defining the property of the transformation function.

Accepting (2) as it stands, by differentiating a relationship can be established between a pair of outputs for given inputs (and the level of the other outputs). Illustrating such a relationship a cocave transformation curve is usually drawn up in a two-dimensional output space showing the rate of transformation between the two outputs (y_1 and y_2) for given values of the other outputs (including z) and inputs:

$$F'_{y1}dy_1 + F'_{y2}dy_2 = 0 \Rightarrow -\frac{dy_2}{dy_1} = \frac{F'_{y1}}{F'_{y2}} \tag{3}$$

Likewise a relationship between a pair of inputs given the level of outputs (and the other inputs) can be established. Illustrating this relationship in a two-dimensional input space using a concave curve shows the rate of substitution between two inputs.

A crucial point is now whether the conventional trade-off relation between two desirable outputs also holds for a trade-off between a desirable and an undesirable output. It turns out that such a single equation frontier transformation relation will be in conflict with the material balance. Consider the standard textbook relationship between the two outputs y and z for a given x (simplifying to single variables of each type of variable) for unit j and that the material balance $ax_j^0 \equiv c_j z_j^0 + by_j^0$ is fulfilled for $F(y_j^0, z_j^0, x_j^0) = 0$. Differentiation of the transformation relation for given x_j^0 at this point yields

$$F'_y dy_j^0 + F'_z dz_j^0 = 0 \tag{4}$$

Apparently this seems to yield the standard marginal rate of substitution with the correct sign since the partial derivatives are both positive.[5] However, we can only increase (decrease) *y* by decreasing (increasing) *z* for a constant x^0 and this is not possible on the frontier because by definition frontier points are efficient. If it was possible to decrease (increase) *z* to provide more (less) mass to the production of *y* of the given mass ax_j^0 then the point (y_j^0, z_j^0, x_j^0) would not be efficient. This reasoning can be repeated for any point on the frontier, thus demonstrating that a transformation relation between desirable and undesirable outputs goes against the materials balance and fundamental efficiency principles of utilising resources.

One can question if differentiation of the material balance equation (1) w.r.t. *y* and *z* has any meaning at all; z cannot go down for given *x*, and *y* cannot go up for given *x*. One may say, hypothetically, that for *y* to go up given *x*, *z* has to go down, but if the unit is on the efficient frontier this is impossible due to the definition of *F*(.) as giving maximal *y* for given *x*.

To avoid the problem above one possibility used in the literature is to treat residuals generation *as if* they are inputs. This option is followed without any comment or explanation in the influential textbook by Baumol and Oates (1975, Table 4.1, p. 39). When a defence of the procedure is offered it is argued that good outputs increase when residuals generation increases because this means that less resources are used on pollution abatement, and these freed resources are then transferred to output production (see e.g. Cropper and Oates (1992).[6]

Another argument used is that generators of residuals need services from Nature to take care of these residuals, and that such services functioning as inputs can be measured by the volume of residuals (Considine and Larson, 2006, p. 649). However, this argument cannot satisfy the need at the micro level to explicit model the generation of residuals. A partial increase in a residual as input cannot technically explain that a good output increases by reasoning that inputs are reallocated from abatement activity to the production of goods. By definition the inputs that are explicitly specified in this relation must be kept constant. Having sort of additional inputs behind the scene is not a very satisfactorily way of modelling.

[5] Pethig (20003; 2006) uses the material balance (1) to confirm the signing. However, as argued above it can be questioned if this is a sound procedure.

[6] "…waste emissions are treated simply as another factor of production; this seems reasonable since attempts, for example, to cut back on waste discharges will involve the diversion of other inputs to abatement activities – thereby reducing the availability of these other inputs for the production of goods" (Cropper and Oates 1992, p. 678).

4. Multi-equation production functions

The simultaneous production (within the time period considered) of multiple outputs can be of several types (see Frisch (1965) for an overview). Inputs may be employed alternatively to produce different outputs, e.g., a piece of agricultural land may be used to produce potatoes or wheat, a wood-cutting tool may be used to produce different types of furniture, etc. There is freedom of choice in what outputs to produce. At the other end of the scale we may have multiple outputs due to technical jointness in production; sheep yield mutton as well as wool, cattle yields beef and hide, we get both wheat and straw, and coal can be converted to coke and gas, to use classical examples from Edgeworth and Marshall. As an extreme form of jointness we have that outputs are produced in fixed proportions, as the distillates of crude oil in a refinery

Frisch multioutput modelling

Frisch generalised various possibilities by introducing a system of μ equations between m outputs y and n inputs x:

$$F^i(y,x)=0,\ i=1,...,\mu \tag{5}$$

Some of the relations may be between outputs only ('Output couplings') and other relations may be between inputs only ('Factor bands'; $m - \mu < 0$). As a standard case Frisch (1965) introduced 'Assorted production' in order to capture that for given inputs you can have an assortment of outputs. The core production function apparatus of Frisch (1965, Part four) is based on this concept. Degree of assortment is defined as the difference between number of outputs and equations: $\alpha = m - \mu$. This key Frisch concept implies that when having just one equation in the output and input variables, $F(y, x) = 0$, 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 $\alpha^{min} = m - \mu = m - m = 0$. A special case of this situation with $m = \mu$ is 'Product separation' of each output being a function of the same set of inputs; this is the case of 'Factorially determined multi-output production'.

The factorially-determined multiple-output model

Pollution is generically a problem with joint outputs in economic activities of production and consumption. As pointed out in Sections 1 and 2, there is a material balance that accounts for

where the mass contained in material inputs end up; in the desirable output or in the natural environment. If all the material inputs into an activity are not embedded in the products the activity is set up to deliver, then the difference must be contained in residuals discharged to the environment. It seems important to capture these physical realities from use of material inputs in any sound modelling of the interaction economic activity and generation of pollutants. It will then be clarifying to distinguish between input factors with material content (raw materials) being affected physically by the production process and factors unchanged (not used up) by the production process, the main groups of the latter being labour, capital and external services.[7] A model from production theory, the *factorially determined multioutput model* (Frisch 1965), seems tailor-made for capturing the physical process of generation of residuals[8]:

$$\begin{aligned} &y = f(x_M, x_S),\ f'_{x_M},\ f'_{x_S} > 0 \\ &z = g(x_M, x_S), g'_{x_M} > 0, g'_{x_S} \le 0 \\ &ax_M \equiv by + cz,\ a > 0, b \ge 0, c > 0 \end{aligned} \tag{6}$$

The function $f(.)$ is defined as maximising y for given inputs. The material inputs are denoted x_M, and x_S are service inputs not consumed (remaining physically intact) and providing services. (Electricity as an input is immaterial without weight and is not usually being classified as a service input, but may be regarded as one in our setting.) The positive partial productivity of service inputs in the desirable output production function and the negative sign in the residuals generation function can be explained by 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.[9] The negative partial derivative of service inputs in the residuals function mirrors the positive sign in the output function. 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.

To keep the model as simple as possible we consider a single output y that is the purpose of the production activity and is the desirable output (or the good output for short), and a single residual or undesirable output z (a pollutant or a bad for short). Generalising to multi output and multi pollutants can be done just by adding more equations, one for each variable,

[7] This grouping of inputs was introduced in Ayres and Kneese (1969, p. 289).
[8] This model was applied consciously to generation of residuals for the first time, as far as I know, in Førsund (1972); (1973), and developed further in Førsund (1998); (2009).
[9] 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 the more labour and less cocoa that are employed.

keeping the same inputs as arguments in all relations, see Førsund (2009).

Thermal generation of electricity is special in the sense that no material input is contained in the good output. For a given amount of fossil primary fuel more electricity can be generated by a better control of the combustion process using more service inputs. But the material content of the fuels is the same. However, looking at the energy balance more of the potential heat in the fuels will be used for electricity production, so the energy balance supports the substitution between the two types of inputs, although the partial derivative for service inputs may then be zero in the residuals function.

Scale properties are found by the proportional changes β_y and β_z in outputs generated by a proportional change α in inputs. Within our additive structure we simply get by differentiating w.r.t. α evaluated without loss of generality at points where α, β_y, β_z are equal to 1;

$$\begin{aligned}
\frac{\partial \beta_y}{\partial \alpha} y &= \frac{\partial f(\alpha x_M, \alpha x_S)}{\partial \alpha} = \sum_{M,S} (f'_{x_M} + f'_{x_S})_{|\alpha,\beta_y=1} \Rightarrow \frac{\partial \beta_y}{\partial \alpha} = \varepsilon_y = \frac{\sum_{M,S} (f'_{x_M} + f'_{x_S})}{y} \\
\frac{\partial \beta_z}{\partial \alpha} z &= \frac{\partial g(\alpha x_M, \alpha x_S)}{\partial \alpha} = \sum_{M,S} (g'_{x_M} + g'_{x_S})_{|\alpha,\beta_z=1} \Rightarrow \frac{\partial \beta_z}{\partial \alpha} = \varepsilon_z = \frac{\sum_{M,S} (g'_{x_M} + g'_{x_S})}{z}
\end{aligned} \tag{7}$$

The variables ε_y and ε_z are the scale elasticities. In classical production theory a standard assumption is that the positive marginal productivities of inputs are decreasing. Frisch (1965) introduced the *Regular Ultra Passum Law* of the development of the scale elasticity along a non-decreasing ray in input space assuming that the scale elasticity declines monotonically from value greater than 1 to values less than 1, thus implying a unique ray optimal scale of $\varepsilon_y = 1$. Adopting this assumption implies the *opposite* development of the scale elasticity for the residuals function starting with values smaller than 1, passing through 1 and increasing with expanding inputs. This means that the marginal productivities of the inputs in the residuals function are increasing, and that there is a unique point along a ray where the scale elasticity $\varepsilon_z = 1$. (It is assumed that $\varepsilon_z > 0$.) But the two scale elasticities may not necessarily be equal to 1 for the same value of inputs. The scale properties are unique for each relation although the same change in inputs generates the response in the respective outputs. It does not have any good meaning to look for a common scale property for the system as a whole. The signing of first and second-order derivatives above is consistent with the function $f(.)$ being concave and the function $g(.)$ being convex.

The materials balance, the third equation in (6), tells us that good outputs *y* cannot be produced without residuals *z* if the complete amount of the material content of the inputs is not contained in the outputs. Thermal electricity generation is an extreme case were all material content of the material input ends up as waste products. The material balance is fulfilled for any combination of inputs in the model (6) because of the unique correspondence between input use and outputs generated by these inputs. Therefore, in our type of model, the material balance is an accounting identity since it holds at any point in input-output space including frontier points. We also see that there is no transformation relation between the outputs for given levels of inputs (this will be illustrated in Fig. 1 below).

The material inputs are *essential* in the sense that we will have no production neither of goods nor bads if $x_M = 0$:[10,11]

$$y = f(0, x_S) = 0, z = g(0, x_S) = 0 \tag{8}$$

There will in general be substitution possibilities between material and service inputs, the rate of substitution evaluated at a point on an isoquant is $(-f'_{x_M} / f'_{x_S})$. This is the amount of material input that is reduced if the service input is increased 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, $(-g'_{x_M} / g'_{x_S}) > 0$ due to the marginal productivity of service inputs being negative. 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. 1. The isoquants for the two outputs can be shown in the same diagram because the arguments in the functions are the same. The level of the residual *z* is increasing moving South-East (red isoquants and broken arrow) direction while the level of the intended (desirable) good *y* is increasing moving North-East (black isoquants and broken 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

[10] Essentiality of factors plays a similar role as null-jointness of desirable and undesirable outputs using weak disposability.

[11] Service inputs are also essential, but the point with (8) is to underline the inevitability of generating residuals when using material inputs.

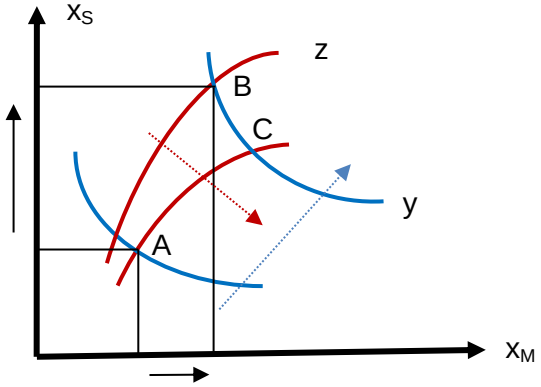


Figure 1. *Isoquants for the production of y and z*

residual *z* has decreased while the production of output *y* has increased. Going from point B to point C keeping the same level of the desirable output, reducing the service input but increasing the material input we see that the level of the undesirable output increases.

The isoquant in Fig. 1 are drawn without any limits for material inputs. However, there are certainly limits in practice regarding the extent of substitution possibilities that should be considered in any empirical analysis. The material balance tells us that the theoretical lower limit for substitution along an isoquant for the desirable output is when there is zero residual generated, i.e. all material content of the raw materials goes into the product. (In the case of electricity generation a theoretical limit is that all the heat energy is converted to electric energy, but the feasibility of such a limit goes against the second law of thermodynamics.)

The theoretical lower limit may be represented by the isoquant showing the zero level for the undesirable product in Fig. 1. This is the upper limit for applying service inputs securing a location of the isoquant maps to be within the economic region, i.e. the substitution region of Frisch (1965) (the uneconomic region has also been called the region with congestion in the efficiency literature). There will also be upper limits for applying material inputs securing a location of the isoquant maps to be within the economic region. However, these limits are not constrained by the material balance.

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 and thereby scale down the use of material inputs. However, this is often the most expensive way to reduce residuals generation (Rødseth 2013).

In a dynamic setting the most promising way is to change the production technology so as to create less pollution for a constant output of goods. Technology improvements that may be small-scale and introduced within not so long time periods (e.g. a year) may be considered variable factors,[12] but technology improvement may also need large capital investments, changing the main production processes into technologies that use less raw materials, or processes them in such a way that less waste of material inputs occur. Such changes will be more of a long-term character based on real capital with a long technical lifetime. Technology change means a simultaneous change of the functional forms *f*(.) and *g*(.) over time:

$$\begin{aligned} &f^{t_2}(x_M, x_S) > f^{t_1}(x_M, x_S) \\ &g^{t_2}(x_M, x_S) < g^{t_1}(x_M, x_S),\ t_2 > t_1 \end{aligned} \tag{9}$$

Technical change in the two production functions in (6) may be illustrated in Fig. 1 by just changing the level labelling of the two sets of isoquants. Positive technical change of desirable output production means producing more for given inputs, while positive 'green' technical change in the residuals production function means generating less residual for the same input levels.

In the relative short run another possibility is to install a separate facility using the residuals from (6) as inputs and processing them in such a way that less harmful pollutants result, e.g. capturing particles using electrostatic filters on smoke stacks, converting an air pollution into a solid waste problem, or using wet scrubbers to convert an air pollutant into liquid waste. Such facilities are called end-of-pipe treatment in environmental economics and will be addressed in the next section.

5. End-of-pipe abatement

In the weak disposability literature abatement has typically been mentioned, but not explicitly modelled.[13] We will add a specific abatement process to the multi-equation model (6). End-of-pipe abatement often consists of a facility separated from the production activity. Another abatement option in the short run is to retool the processes and do small-scale changes. This

[12] Waste heat may be recaptured by applying more capital in the form of heat exchangers and reduce the amount of residuals for constant primary energy, and thus increase production (Martin, 1986).

[13] In Färe et al. (2001, p. 387) it is stated. "…, abatement uses resources that otherwise could have been used to expand production of the good output", and 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…" But note the recent literature introducing abatement in efficiency models mentioned in Section 1.

option is an alternative 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 residuals can then be regarded as an input to the abatement process. In addition other inputs - like labour, capital, and chemicals - absorbing substances and energy, 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 is assumed in applications) than the primary ones (Førsund, 2009)[14]. 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.

As observed in Ayres and Kneese (1969, p. 283) abatement does not "destroy residuals but only alter their form". Expressing the abated residuals as outputs we formulate the following abatement production function (see also Førsund (1973); Pethig (2006); Hampf (2014); Førsund (2009), the latter provides a generalisation to more than one primary residual and the introduction of new types of abatement outputs):

$$\begin{aligned} &\frac{\delta z^a}{z} = A(x_M^a, x_S^a),\ A'_{x_M}, A'_{x_S} > 0, \frac{\delta z^a}{z} \in [0,1] \\ &z^D = z - \delta z^a \geq 0 \\ &a' x_M^a + cz \equiv b' z^a + cz^D \end{aligned} \tag{10}$$

The abatement activity receives the primary residual z defined by (6) 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. It is assumed that the function $A(.)$ is concave. Usually abatement is represented by a cost function in the economics literature. 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. 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. The latter amount z^D is the secondary residual as it is called in the environmental economics

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

literature. 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. 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 and then do a simple subtraction shown in the second equation in (10). It is assumed that the abated residuals do not create any environmental damage. The feasible range of modification is from zero to 1. 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 for 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 knowing the rate of abatement A both the absolute amounts of the two abatement outputs can be calculated: $\delta z^a = Az,\ z^D = (1-A)z$. Applying the materials balance principle in the last equation of (10) the abatement activity will add to the total mass of residuals if material inputs are used, but the point is that abatement means less mass of the harmful residual; $z^D < z$. The conversion coefficient for the material input is now *a'* that in general is different from *a*, and likewise *b'* is in general different from *b*. The *c* coefficient is the same as in (6). The conversion coefficients measure the common substance in all variables in the same unit, e.g. weight (not accounting the substances in the residuals added from the air during the combustion). The mass of the primary residual on the left-hand side is now functioning as an input. The right-hand side shows where the mass ends up.

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). To do this requires a restriction to be imposed on the availability of inputs. 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. It is closer to reality not to consider a common

resource pool for the production unit, but to regard the activities (6) and (10) 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. But this electricity can be formally regarded as a bought input so (10) may still be used.[15]

6. Optimal solutions for the multi-equation model

The social planner solution

As a reference for studying optimality of policy instruments for environmental regulation social value considerations will be introduced. The standard social planning problem is to maximise consumer plus producer surplus, using demand functions for the desirable outputs, and given (positive) input prices q_j, to calculate input cost. The pollutants are evaluated through a monetised damage function:

$$D = D(z_1^D, .., z_k^D),\ \frac{\partial D}{\partial z_s^D} \geq 0, s = 1, ..., k \tag{11}$$

The damage function is a typical relationship that is used in environmental economics to capture the willingness to pay of consumers for environmental qualities. We are now looking at a model consisting of Eq. (6), the multi-output technology and (10) as the end-of-pipe abatement and (11) as social evaluation of damage caused by residuals. There may be several types k of secondary residuals z^D that are emitted to the environment and causing damage. For simplicity we consider a single undesirable output only. Using demand functions $p(y)$ on price form and assuming that given input prices q_j are used for social evaluation of inputs the social planning problem is:

[15] There is thus no need to use the so-called network model because a part of the gross output is used as an input in the abatement process as done in Färe et al (2013); Hampf (2014).

$$\text{Max} \int_{w=0}^{y} p(w)dw - \sum_{j=M,S} q_j x_j - \sum_{j=M,S} q_j^a x_j^a - D(z^D)$$

s.t.

$$\begin{aligned} & y = f(x_M, x_S) \\ & z = g(x_M, x_S) \\ & \delta z^a / z = A(x_M^a, x_S^a) \\ & z^D = z - \delta z^a = z(1 - A(x_M^a, x_S^a)) \end{aligned} \tag{12}$$

The material balance identities are not restated in this section for convenience. We assume that the abated amount z^a (and any new residuals created by material resources used in the abatement process (see Førsund (2009) for a more elaborate specification with creation of new pollutants) are taken care of at zero social or private cost).

Inserting the production functions for the good *y*, the primary pollutant *z*, the secondary pollutant z^D and the abatement function into the objective function yields the optimisation problem:

$$\text{Max} \int_{w=0}^{f(x_M,x_S)} p(w)dw - \sum_{j=M,S} q_j x_j - \sum_{j=M,S} q_j^a x_j^a - D(g(x_M, x_S)(1 - A(x_M^a, x_S^a)) \tag{13}$$

The endogenous variables in the problem are the production process inputs and the abatement process inputs.

Assuming interior solutions for all inputs the necessary first-order conditions are:

$$\begin{aligned} & f'_{x_j} p - q_j - D' g'_{x_j} (1 - A) = 0, \; j = M, S \\ & -q_j^a + D' \text{z} A'_{x_j^a} = 0, \; j = M, S \end{aligned} \tag{14}$$

The expression $g'_{x_j}(1 - A)$ in the last term in the conditions for the production process inputs shows first the marginal increase for $j = M$ (decrease for $j = S$) of the primary pollutant of a unit increase in input *j* multiplied with a factor that shows the share of the secondary residual generated by the increase in the primary residual. The factor will typically be in between zero and one, i.e. the abatement is less than 100% but greater than zero. Positive abatement is taking place cushioning the impact on marginal damage from the full increase of the primary residual. In the case of service input being increased the marginal damage will decrease due to

a decrease of the primary residual and thus also of the secondary residual from the abatement process.

The first set of necessary conditions tell us that for each type of production-process factor the marginal revenue of increasing factor *j* and consequently increasing the good output is equal to the unit cost of the factor plus the marginal damage of increasing factor *j*. Social marginal cost is added to the unit input cost. For an increase in a service input the implied reduction in social marginal cost due to a reduction in both the primary and secondary residual is deducted from the factor price:

$$\underset{\text{Marginal revenue}}{f'_{x_j} p} = \underset{\text{Unit factor cost}}{q_j} + \underbrace{D' g'_{x_j} (1-A)}_{\text{Marginal damage of increase in a factor}}, \; j = M, S \tag{15}$$

To see the impact of the use of the two types of inputs the rate of substitution between a material input and a service input is:

$$\frac{f'_{x_M}}{f'_{x_S}} = \frac{q_M + D' g'_{x_M} (1-A)}{q_S + D' g'_{x_S} (1-A)} \tag{16}$$

The unit price on the material input is higher than the given market price, but the opposite is the case for the service factor because the impact g'_{x_S} on primary pollutants is negative. The optimal solution implies a relative reduced use of the material input compared with a solution without a damage function and abatement function. At the margin a material input generates social cost while a service input generates a saving of social cost.

The second set of conditions in (14) tells us that for each type of abatement-process inputs at the optimal level of abatement, i.e. both primary and secondary pollutants are at their optimal levels, the employment of abatement inputs should be expanded until the marginal damage, caused by the generation of secondary pollutants, is equal to the unit price of the abatement input:

$$D' z A'_{x_j^a} = q_j^a, \; j = M, S \tag{17}$$

A marginal increase in an abatement input increases the rate of abatement and consequently increases the amount of abated residuals and decreases the untreated secondary residuals, thereby lowering the marginal damage implied by this new level of secondary pollutant.

Imposing a constraint on emission

The environment agency may impose an upper limit z_R^D on the amount emitted during a specific time period; $z^D \leq z_R^D$. The firm's optimisation problem becomes

$$\begin{aligned}
&\text{Max } py - \sum_{j=M,S} q_j x_j - \sum_{j=M,S} q_j^a x_j^a \\
&s.t. \\
&y = f(x_M, x_S) \\
&z = g(x_M, x_S) \\
&\delta z^a / z = A(x_M^a, x_S^a) \\
&z^D = z - \delta z^a \\
&z^D \leq z_R^D
\end{aligned} \tag{18}$$

The optimisation problem may be written more compactly as

$$\begin{aligned}
&\text{Max } pf(x_M, x_S) - \sum_{j=M,S} q_j x_j - \sum_{j=M,S} q_j^a x_j^a \\
&s.t. \\
&g(x_M, x_S)(1 - A(x_M^a, x_S^a)) \leq z_R^D
\end{aligned} \tag{19}$$

The necessary first-order conditions are:

$$\begin{aligned}
&pf'_{x_j} - q_j - \lambda g'_{x_j}(1 - A) = 0,\ j = M, S \\
&-q_j^a + \lambda z A'_{x_j} = 0,\ j = M, S
\end{aligned} \tag{20}$$

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. Comparing (14) and (20) we see that the direct regulation can realise the optimal solution if the shadow price on the pollution constraint is equal to the marginal damage.

A Pigou tax

Let us assume that the environmental regulator uses a tax *t* per unit of secondary pollution as a regulatory instrument. Regarding the unit as a firm that maximizes profit facing competitive markets both for output and inputs, introducing a Pigou tax on secondary pollutants yields the following optimization problem:

$$\begin{aligned}
&\text{Max } py - \sum_{j=M,S} q_j x_j - \sum_{j=M,S} q_j^a x_j^a - tz^D \\
&s.t. \\
&y = f(x_M, x_S) \\
&z = g(x_M, x_S) \\
&z^a / z = A(x_M^a, x_S^a) \\
&z^D = z - \delta z^a
\end{aligned} \tag{21}$$

Using again the inputs of both the production and the abatement activities as endogenous variables the optimisation problem becomes:

$$\text{Max } pf(x_M, x_S) - \sum_{j=M,S} q_j x_j - \sum_{j=M,S} q_j^a x_j^a - tg(x_M, x_S)(1 - A(x_M^a, x_S^a)) \tag{22}$$

The necessary first-order conditions are:

$$\begin{aligned}
&f'_{x_j} p - q_j - tg'_{x_j}(1 - A) = 0,\ j = M, S \\
&-q_j^a + tzA'_{x_j^a} = 0,\ j = M, S
\end{aligned} \tag{23}$$

The optimal social solution can be implemented if the tax is set equal to marginal damage of the secondary pollutant. The rate of substitution between a material input and a service input in the production stage is

$$\frac{f'_{x_M}}{f'_{x_S}} = \frac{q_M + tg'_{x_M}(1 - A)}{q_S + tg'_{x_S}(1 - A)} \tag{24}$$

A tax on the secondary pollutant will give the firm an incentive to reduce the use of material inputs and increase the use of service inputs. However, the desirable output will decrease compared with a situation without the environmental regulation using a tax. We see comparing (20) and (23) that the tax takes the place of the shadow price on the secondary pollutant. The optimal solution can be realised if the tax is set equal to the marginal damage.

Cap and trade

Cap and trade has become popular as an indirect policy instrument starting in USA with SO_2 and introduced for CO_2 in EU and used as a limited regional experiment in China for CO_2 also and announced to be extended soon in the whole of China. In the case of several firms the regulation may be introducing tradable quotas z_{jR}^D for each firm *j* summing up to the total amount of the pollutant that the regulation will impose in the case of the localization of the

emitting firms having no site-specific environmental impact, i.e. it is the sum of discharges that creates environmental damage. If the quota price emerging from the trading is equal to the optimal marginal damage in the solution to problem (12), then the cap and trade policy instrument can also realise the social planner's solution.

7. Allowing for inefficient operations

In view of the importance of the material balance for the choice of model 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. 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 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. The material balance also holds for inefficient observations (as pointed out in Section 2). It is the amount of residuals and outputs that has a potential for change, while the a, b, c coefficients and the inputs 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 type 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 two 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 Section 4. However, in the case of inefficient operations being improved the isoquants are now shifting in the same way as for technical change as explained in Section 4.

There is another type of problem in the efficiency strand of research not often mentioned concerning the behavior of (the management of) firms. It is difficult to assume, as in standard production theory using frontier functions only, that inefficient firms can optimise in the usual sense of obtaining maximal profit or minimising costs, as modelled in the previous section. 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 observation are 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 optimization 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 Section 5.

The most common way to set up a general production possibility set allowing for inefficiency including both desirable and undesirable outputs is:

$$T = \{(y,z,x) \,|\, y \geq 0 \text{ and } z \geq 0 \text{ can be produced by } x \geq 0,\ ax \equiv by + cz\} \tag{25}$$

The materials balance is included as a condition to be satisfied. 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 can equivalently be represented by the output set

$$P(x) = \{(y,z) \,|\, x \geq 0 \text{ can produce y} \geq 0 \text{ and z} \geq 0,\ ax \equiv by + cz\} \tag{26}$$

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. But a question is if this applies also to the production of undesirable outputs, as discussed in Section 3. It seems that in the literature this is assumed without any discussion, i.e. that maximal undesirable outputs are also obtained for given inputs.

Weak disposability

In order to operate the single equation model (2) with undesirable outputs avoiding the zero solution for residuals pointed out in Section 3, restrictions must be placed on the production possibility set. This has typically been done by imposing weak disposability, a mathematical concept introduced by Shephard (1970), defined as

$$\text{If } (y,z) \in P(x), \text{then } (\theta y, \theta z) \in P(x) \text{ for } 0 < \theta \leq 1 \tag{27}$$

This means that along the frontier desirable and undesirable outputs must change with the same segment-specific proportionality factors. 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 (Leontief 1970) that is backed up by economic reasoning and empirical findings.

Illustrations of weak disposability for output sets are presented in Fig. 2 taken from the first

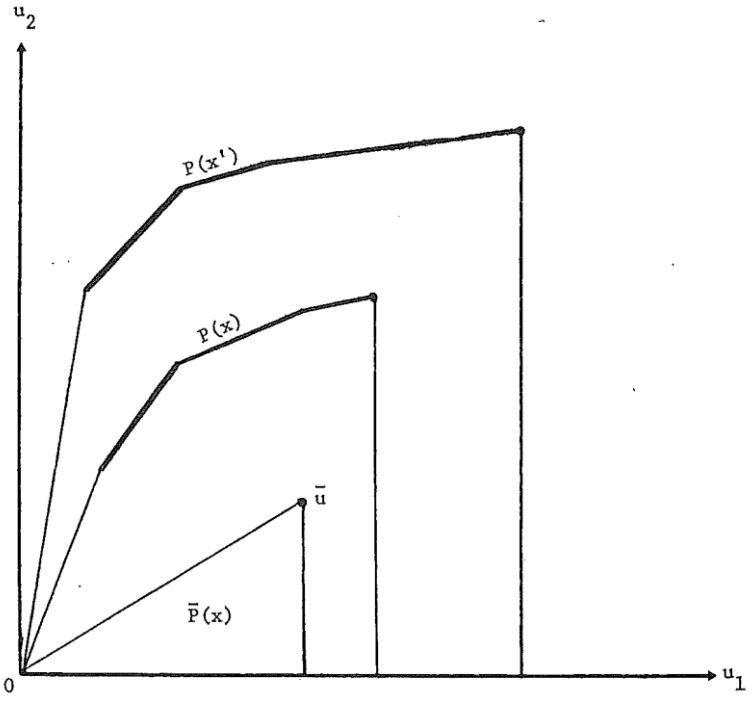


FIGURE 32 (a): OUTPUT SETS FOR A PRODUCTION CORRESPONDENCE WITH WEAK DISPOSAL (u_1 NOT DESIRABLE), $x' \geqq x$

Figure 2. *Illustration of weak disposability*
Source: Shephard (1970, p. 188)

illustration of weak disposability of desirable and undesirable outputs in Shephard (1970, p. 188). 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 case of a fixed relationship between the two outputs. The contour curves starting from the origin securing the condition of inevitability of positive undesirables when desirable output is positive, termed the null-jointness condition in Shephard and Färe (1974). 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. 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. This is done in Rødseth (2014; 2015) and used in Färe et al (2013)[16] (building on Rødseth's dissertation from 2011). In Rødseth (2015) the weak disposability assumption is found to be consistent with the material balance if abatement is introduced, but also if some special conditions (weak G-disposability) are fulfilled even without abatement. Explicit abatement within a so called network model[17] introduced in Färe et al (2013) is followed up in Hampf (2014)[18].

Maintaining the assumption of weak disposability, using actively a trade-off between desirable and undesirable outputs is, however, problematic also when abatement is explicitly introduced. Now a distinction must be made between the generation of residuals in the production stage and the actual pollutants emitted to the environment after abatement. If the trade-off is between the latter and a desirable output (as it should be) then the production of abatement cannot be kept constant when spanning the trade-off.

The weak disposability 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'[19] is estimated, i.e. the data fits a model that does not represent the true way the desirable and undesirable outputs are jointly generated.

[16] However, the abatement activity as such is not modelled explicitly as in (10).
[17] To use the term 'network' may seem to be an overkill, after all we are talking about two distinct activities only; joint production of the desirable and undesirable outputs and an end-of-pipe abatement activity.
[18] Hampf (2014) models abatement output as the difference between primary residuals and abated amount, i.e. the secondary residual in (10) (without commenting on the problem of units of measurement), but how the abatement resources, being 'non-polluting' only, are influencing the two types of abatement outputs is not quite so clear as the explicit modelling in Section 5.
[19] I owe this aptly expression to Darold Barnum.

8. Efficiency measures and their estimation in the multi-equation model

The multi-equation model (6) with add-on abatement (10) can be extended to include inefficient operations as in the single-equation model (25) with the restriction (27) of weak disposability. The multi-equation model allowing inefficiency can be set up using inequalities (with the partial derivatives of the functions as given in (6) and (10):

$$
\begin{aligned}
& y \le f(x_M, x_S) \\
& z \ge g(x_M, x_S) \\
& \delta z^a / z \le A(x_M^a, x_S^a) \\
& a x_M \equiv by + cz \\
& a' x_M^a + cz \equiv b' z^a + cz^D
\end{aligned}
\tag{28}
$$

The first of the two last identities hold for the two first production activities simultaneously, and the last holds for the abatement activity. Following Murty et al (2012) the possibility sets can be written:

$$
\begin{aligned}
T_1 &= \left\{ (x_M, x_S, y) \middle| y \le f(x_M, x_S) \text{ and } y \ge 0, x_M \ge 0, x_S \ge 0 \right\} \\
T_2 &= \left\{ (x_M, x_S, z) \middle| z \ge g(x_M, x_S) \text{and } z \ge 0, x_M \ge 0, x_S \ge 0 \right\} \\
T_3 &= \left\{ (x_M^a, x_S^a, z, z^a) \middle| \delta z^a \le zA(x_M^a, x_S^a), z^a \ge 0, z \ge 0, x_M^a \ge 0, x_S^a \ge 0, a x_M^a + cz \equiv b' z^a + cz^D \right\}
\end{aligned}
\tag{29}
$$

The materials balance condition in the set (25) must hold for the two processes of the production activity simultaneously. 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.

A possible 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, 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 between zero and one. These measures can be either input oriented or output-oriented. In our setting output orientation seems to be a natural choice.

Concerning the estimation of the unknown frontiers a non-parametric DEA model, build up as a polyhedral set, assuming standard axioms such as convexity, monotonicity and minimum

extrapolation, 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 (see e.g. Fried et al 2008). 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 tentative three DEA optimization problems below for unit *i* variable-returns-to scale functions are specified. 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 projected observation (y_i, x_i), (z_i,x_i) :

$$
\begin{aligned}
&1/E_{y_i} = Max_{\lambda,\theta}\,\theta \\
&s.t. \\
&\sum_{j=1}^{n}\lambda_j y_j \geq \theta y_i, i=1,...,n \\
&\sum_{j=1}^{n}\lambda_j x_{kj} \leq x_i, i=1,...,n, k \in M, N \\
&\sum_{j=1}^{n}\lambda_j = 1, \lambda_j \geq 0, \theta \geq 0
\end{aligned}
\tag{30}
$$

Remember that we have assumed that the function $g(.)$ is convex:

$$
\begin{aligned}
&E_{z_i} = Min_{\lambda',\varphi}\,\varphi \\
&s.t. \\
&\sum_{j=1}^{n}\lambda'_j z_j \leq \varphi z_i, i=1,...,n \\
&\sum_{j=1}^{n}\lambda'_j x_{kj} \geq x_i, i=1,...,n,\ k \in M, S \\
&\sum_{j=1}^{n}\lambda'_j = 1, \lambda_j \geq 0, \varphi \geq 0
\end{aligned}
\tag{31}
$$

As stated previously the materials balance identity is not specified here. It holds for the two problems together, not (30) and (31) separately. There is also another problem with the material balance estimating a non-parametric frontier using DEA. The problem is that projections to the frontier in problems in (30) and (31) of inefficient points must also satisfy the relevant material balance condition in (28). The projection points are

$$
\begin{aligned}
&\sum_{j=1}^{n}\lambda_j y_j, \sum_{j=1}^{n}\lambda_j x_{kj}, k \in M, S \\
&\sum_{j=1}^{n}\lambda'_j z_j, \sum_{j=1}^{n}\lambda'_j x_{kj}, k \in M, S
\end{aligned}
\tag{32}
$$

These points are not observations, but constructs of the analyst. Assuming projection points being on efficient faces, i.e. all the inequalities in (30) and (31) hold as equalities, the restriction for unit *i* is

$$a\sum_{j=1}^{n}\lambda_j x_{M\,j} \equiv b\sum_{j=1}^{n}\lambda_j y_j + c\sum_{j=1}^{n}\lambda_j' z_j \Rightarrow a x_{Mi} = b\theta y_i + c\varphi z_i \tag{33}$$

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. Regarding weakly efficient faces there will be slacks on constraints. However, these may typically be different between the models.

In the non-parametric estimation model for abatement efficiency the observed amount of primary residual for unit *i* is now given and not appearing in the model:

$$\begin{aligned}
&1/E_{A_i} = Max_{\lambda'',\phi}\,\phi \\
&s.t. \\
&\sum_{j=1}^{n}\lambda_j'' A_j \geq \phi A_i, i = 1,...,n \\
&\sum_{j=1}^{n}\lambda_j'' x_{k\,j}^{a} \leq x_i^{a}, i = 1,...,n, k \in M, N \\
&\sum_{j=1}^{n}\lambda_j'' = 1, \lambda_j'' \geq 0, \phi \geq 0
\end{aligned} \tag{34}$$

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. However, the abatement materials balance will place a restriction on these projection points that should be entered in the model:

$$\begin{aligned}
&a\sum_{j=1}^{n}\lambda_j'' x_{M\,j}^{a} + c\sum_{j=1}^{n}\lambda_j' z_j \equiv b' z_i \sum_{j=1}^{n}\lambda_j'' A_j + c\sum_{j=1}^{n}\lambda_j' z_j \sum_{j=1}^{n}\lambda_j''(1-A_j) \Rightarrow \\
&a x_{M_i}^{a} + c z_i = b'\phi A_i + c z_i (1-\phi A_i)
\end{aligned} \tag{35}$$

Without imposing this condition it may be unlikely that it will be fulfilled. Notice that problems (31) and eq. (35) are connected in the sense that the primary pollution for unit *i* in the abatement material balance is the observation for this variable in problem (31).

The term environmental efficiency is used somewhat differently in the literature and is not used in the efficiency measures introduced above. One reason for the latter 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. However, the most common notion of environmental efficiency is showing the potential relative reduction in emission of residuals. In Hampf (2014) the concept of environmental efficiency measure for units having two stages of production; production of good and bad outputs and abatement, is based on (weighted) minimal amount of emissions released to the environment to the (equally weighted) actual observed amount of emissions from a unit. This measure is further decomposed multiplicatively into a production efficiency measure and an abatement efficiency measure. However, for policy purposes it seems that the individual measures above provide most valuable information for designing specific direct regulations or indirect economic instruments.

In the literature there is an interest in presenting a single efficiency measure for models without explicit abatement. The first paper using the directional distance function for this purpose is Chung et al (1997). An additive distance function specifies the same positive additive change factor for the good output as a negative change factor for the bad projecting an inefficient observation to the frontier in a predetermined direction. A problem with this approach is that the solution for the change factor depends on the direction chosen. Another type of problem is the nature of the single efficiency measure; an inefficient unit is 'rewarded' for producing good outputs, but 'punished' for producing bads, but such a 'value comparison' is void of any real value information of the economic trade-offs between good outputs and bads.

In seven overlapping papers as to methods (Sueyushi et al 2010; Sueyushi and Goto (2010); (2011a,b); 2012a,b,c) a model similar to the first two equations in (28) is introduced (without any reference to the relevant literature) and a DEA model is used to estimate efficiency for the two activities production of desirable and undesirable outputs. As efficiency measures both the range adjusted measure and the directional distance function are applied. Abatement is not considered. What is called a unified approach joining the two activities when estimating efficiency scores is preferred. However, to have a kind of average efficiency score, in the case of range adjusted measures, based on relative distances to the frontiers does not serve this purpose. It is difficult to see that such a measure can have any useful policy purpose.

Non-parametric frontiers for desirable and undesirable outputs are illustrated in Fig. 3 (and

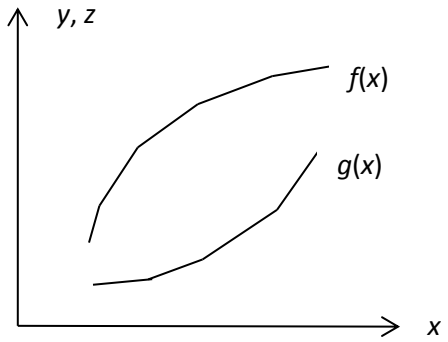


Figure 3. Best practice frontiers $f(x)$, $g(x)$ for desirable (y) and undesirable (z) outputs
Source: adaption from Sueyushi et al (2010)

found in all seven papers quoted above). Notice that the form of the functions in the simple case of Fig. 3 implies that the scale elasticities follow the development of scale elasticities assumed in Section 4 (se also Sueyushi and Goto 2013)[20]. The location of the two curves depends on the measurement units of y and z. It may well be the case that the $g(x)$ curve lays above the $f(x)$ curve, thus making the intersection of set T_1 and T_2 empty.

Murty et al. (2012) use a multi-equation model related to the model in Section 4 when analysing the trade-offs between the good and the bad, introducing explicitly generation of bads and estimating the model using DEA. Generation of bads are modelled following partly the factorially determined multioutput scheme in (6). However, abatement is introduced as a new output y^a appearing in the transformation relation for the intended output (y) and the bad (z):

$$\begin{aligned} &f(x_1, x_2, y, y^a) \leq 0 \\ &z \geq g(x_2, y^a) \end{aligned} \tag{36}$$

The inputs of type x_1 will not cause pollution, but the inputs of type x_2 will (cf. x_S and x_M in (6)). The overall technology set T is then specified as the intersection of the two sets based on the relations in (36). In the first relation we see that the abatement output compete with the output for resources, although there is no good reason for this at a micro level. In addition we do not see how the abatement output is actually produced as in (10). Another weakness in the model is that the signing of the partial derivatives in the first relation in (36) implies that there

[20] However, Sueyushi and Goto (2013) do not utilise right-hand and left-hand derivatives when illustrating scale elasticities in the DEA case of non-differentiable frontier points, see Førsund et al (2007).

is a trade-off between the bad and the good for a given amount of inputs (although it is called a correlation), but this violates the material balance equation.

9. Conclusions

When modelling the interactions between the production of desirable outputs and the natural environment a key foundation is the material balance telling us that mass in an economic activity cannot disappear but only takes on different forms. 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 sufficient realistic way. In the efficiency literature the most popular approach to empirical studies has been to assume a mathematical property of weak disposability of the production possibility set allowing for inefficient observations. This property blocks the degenerate case of using all resources on desirable outputs resulting in zero emission of residuals. However, a main result of the paper is that a trade-off between desirable and undesirable outputs, as implied by the weak disposability model, is not compatible with the material balance. An alternative model from 'classical' production theory that obeys the material balance is developed 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 type of model can easily be extended to cover abatement efforts of the end-of-pipe type. Abatement of residuals may also be added to the weak disposability model, but the increased complexity of the model, compared with the alternative model of the paper extended to cover inefficient operations, seems excessive.

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 through-put industries such as oil refineries, steel and iron, aluminium and other energy-intensive industries, as well as food-processing and cement, etc. 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) and 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.

Non-parametric DEA efficiency models are the only ones mentioned in the paper. However, parametric models may also be estimated (Färe et al 2013). As pointed out in Murty et al (2012) "…the extension to an econometric approach that models by-production is not difficult to foresee." Murty (2015) presents a set of comprehensive axioms for distance functions of emission-generating technology models, the fulfilling of which allows parametrically representation of more than one implicit production function. The paper by Kumbhakar and Tsionas (2015) and Malikov et al (2015) represents a start of an extension using parametric functions and Bayesian Markov Chain Monte Carlo methods to estimate inefficiencies in models of simultaneous generation of goods and bads. However, there may be a problem with reconciling stochastic frontiers with the material balance in general due to the latter relation being deterministic.

References

Aigner DJ, Lovell CAK and Schmidt P (1977) Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* 6(1), 21–37

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

Baumol WJ and Oates W (1975) *The theory of environmental policy*, Cambridge University Press, Cambridge (second edition 1988).

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

Charnes A, Cooper WW and Rhodes E (1978) Measuring the efficiency of decision making units. *European Journal of Operational Research* 2, 429–444

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

Coelli T, Lauwers L, Van Huylenbroeck, G (2007) Environmental efficiency measurement and the materials balance condition. *Journal of Productivity Analysis* 28, 3–12

Considine TJ and Larson DF (2006) The environment as a factor of production, *Journal of Environmental Economics and Management* 52, 645-662

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

Dakpo KH, Jeanneaux P and Latruffe L (2016) Modelling pollution-generating technologies in performance benchmarking: recent developments, limits and future prospects in the nonparametric framework. *European Journal of Operational Research* 250, 347-359

Färe R, Grosskopf S and Hernandez-Sancho F (2004) Environmental performance: an index number approach. *Resource and Energy Economics* 26, 343–352

Färe R, Grosskopf S and Margaritis D (2008) Efficiency and productivity: Malmquist and more. In: Fried HO, Lovell CAK, and Schmidt SS (eds) *The measurement of Productive Efficiency and Productivity Growth*. Chapter 5, 522–622, Oxford University Press, New York

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

Färe R, Grosskopf S and Pasurka CA (2001) Accounting for air pollution emissions in measures of state manufacturing productivity growth. *Journal of Regional Science*, 41(3), 381–409

Färe R, Grosskopf S and Pasurka P (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, 109–118. Amsterdam: Elsevier

Färe R, Grosskopf S, Noh D-W and Weber W (2005) Characteristics of a polluting technology: theory and practice. *Journal of Econometrics* 126, 469–492

Färe R, Grosskopf S and Pasurka CA (2014) Potential gains from trading bad outputs: the case of U.S. electric power plants. *Resource and Energy Economics* 36, 99-112

Färe R, Grosskopf S and Tyteca D (1996) An activity analysis model of the environmental performance of firms – application to fossil fuel-fired electric utilities. *Ecological Economics* 18, 161–175

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

Fisher AC and Peterson FM (1976) The environment in economics: a survey, *Journal of Economic Literature* 14(1), 1-33

Fried HO, Lovell CAK, and Schmidt SS (eds.) (2008) *The measurement of productive efficiency and productivity growth*. Oxford University Press, New York

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.* D. Reidel, Dordrecht

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 (1998) Pollution modelling and multiple-output production theory. *Discussion Paper* # D-37/1998, Department of Economics and Social sciences, Agricultural University of Norway

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)* XLVI, 1-34

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

Førsund FR (2010) Dynamic efficiency measurement. *Indian Economic Review* 45(2), 125-159

Førsund FR, Hjalmarsson L, Krivonozhko VE and Utkin OB (2007) Calculation of scale elasticities in DEA models: direct and indirect approaches. *Journal of Productivity Analysis* 28, 45–56

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, 457-473

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

Kneese AV, Ayres RU and d'Arge RC (1970) *Economics and the environment. a materials balance approach*, Johns Hopkins Press, Baltimore

Kumbhakar SC and Tsionas EG (2015) The good, the bad and the technology: endogeneity in environmental production models. *Journal of Econometrics*. In press http://dx.doi.org/ 10.1016/j.jeconom.2015.06.008

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

Malikov E, Kumbhakar SC and Tsionas EG (2015) Bayesian approach to disentangling technical and environmental productivity. *econometrics* 3, 443-465

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

Mishan EJ (1971) The postwar literature on externalities: an interpretative essay, *Journal of Economic Literature* IX(1), 1-28

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

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

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, 185-204

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, 211-223

Rødseth KL (2015) Axioms of a polluting technology: a material balance approach. *Environment and Resource Economics* DOI 10.1007/s10640-015-9974-1. Published online: 17 October 2015

Rødseth KL (2016) Environmental efficiency measurement and the material balance condition reconsidered. *European Journal of Operational Research* 250, 342-346

Shephard RW (1953). *Cost and Production Functions.* Princeton University Press, Princeton

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

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

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, 5902-5911

Sueyoshi T and Goto M (2011a) Methodological comparison between two unified (operational and environmental) efficiency measurements for environmental assessment. *European Journal of Operational Research* 210, 684-693

Sueyoshi T and Goto M (2011b) DEA approach for unified efficiency measurement: assessment of Japanese fossil fuel power generation. *Energy Economics* 33, 292-303

Sueyoshi T and Goto M (2012a) Returns to scale and damages to scale under natural and managerial disposability: Strategy, efficiency and competitiveness of petroleum firms. *Energy Economics* 34, 645-662

Sueyoshi T and Goto M (2012b) Environmental assessment by DEA radial measurement: U.S. coal-fired power plants in ISO (Independent System Operator) and RTO (Regional Transmission Organization). *Energy Economics* 34, 663-676

Sueyoshi T and Goto M (2012c) Data envelopment analysis for environmental assessment: Comparison between public and private ownership in petroleum industry. *European Journal of Operational Research* 216, 668-678

Sueyoshi T, Goto M and Ueno T (2010) Performance analysis of US coal-fired power plants by measuring three DEA efficiencies. *Energy Policy* 38, 1675-1688

Sueyoshi T and Goto M (2013) Returns to scale vs. damages to scale in data envelopment analysis: an impact of U.S. clean air act on coal-fired power plants. *Omega* 41, 164-175