Materials, Capital, Direct/Indirect Substitution and Mass Balance Production Functions

Jeroen C.J.M. van den Bergh[*]

Department of Spatial Economics
Vrije Universiteit
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
e-mail: jbergh@econ.vu.nl

[*] Fellow of the Tinbergen Institute Amsterdam/Rotterdam. This paper was partly written during a visit to the Department of Economics, Rensselaer Polytechnic Institute, Troy, USA, in October and November 1997. I am grateful to Ada Ferrer Carbonell, John Gowdy, Susan Mesner, Sigrid Stagl, Cees Withagen and an anonymous referee for helpful comments.

Abstract

The aim of this article is to provide a foundation for a correct and accurate analysis of the relationship between monetary values and physical dimensions in economic production processes that transform materials into products. This is relevant for a number of research topics, such as changes in materials use in production, the recycling of materials and products, “dematerialization” on a macro scale, and nonrenewable resource limits to growth. It is argued that the notions of “substitution” and “capital” are often cryptically used. In order to better understand and predict the changes that can occur in production, a distinction is proposed between direct and indirect substitution. Linked to this, a classification is offered of various types of substitutability and complementarity relationships between production factors. It is argued that the neoclassical production function may be consistent with mass balance, but is unsuitable to offer a detailed and accurate understanding of changes in production that influence materials use. For this purpose, a set of general formulations of production functions satisfying mass balance is presented, drawing upon the proposed classifications of substitution and production factors.

1. Introduction

Since a long time economists have been involved in the debate on whether future economic growth will be hampered by finite supplies of natural resources. Disagreement seems to persist, mainly because the relationship between value and physical dimensions is vague or remains unspecified, notably in models of economic production. This article proposes new concepts and formal descriptions of production that allow to provide for more detail and accuracy in examining material limits to economic processes. This is relevant for various other research topics as well. Recently, the issue of "dematerialization" is receiving a great deal of attention, which clearly requires analyses on the basis of models that are consistent with both economics and physics (Ayres and Ayres, 1996; de Bruyn and Opschoor, 1997; von Weizsäcker *et al.*, 1997). In addition, studies of material-product chains of activities, focusing on the recycling of materials and waste management, may generate reliable and satisfactory results when consistency with physics is strived for (Kandelaars and van den Bergh, 1996; Kandelaars, 1998).[1]

The discussion here pays attention to a number of fundamental aspects of such research. It is argued that "substitution" is subject to much confusion. Alternative interpretations of it are examined. Drawing upon the insights of Georgescu-Roegen (1971a), a classification of production factors is proposed that is linked to particular interpretations of substitution and complementarity. Based on these choices, the specification of general, possibly nonlinear, production functions satisfying mass balance is considered. This involves revisiting "neoclassical production functions" that are characterized by continuity and a high degree of substitution of inputs. It will be argued here that an analysis of substitution of materials, and limits to it, on a highly aggregate level, is unable to take into account the specific character of stocks, funds and flows, and the different types of substitution mechanisms and complementarity relationships that exist among these. As a result, such an aggregate analysis may give rise to incorrect or non-interpretable results.

In studies of materials flows and economic processes that are consistent with physics, mass balance provides a basic guiding principle. It denotes that materials can not be created or destroyed, and that material inputs in processes end up in either accumulated stocks or material output flows, the latter including both useful and waste output flows.[2] The discussion later on will

[1] "Value" can be based on revealed or stated preferences, and may include use and nonuse values. Alternative definitions of "value" do not affect the general findings here. The discussion is thus largely independent of assumptions such as "behaviour of individuals is motivated by utility maximization".

[2] The First Law of Thermodynamics or the principle of conservation of matter-energy, states that energy can be neither created nor destroyed. Mass balance is only an approximate and derived statement,

address two specific questions relating to materials use: whether mass balance is implicit or explicit in economic models of production, or not satisfied at all; and whether there is a limit to the amount of value that can be generated on the basis of a finite amount of material inputs in economic production.

Implications of the entropy law are not considered here. They require a separate treatment (e.g., Ayres, 1978; Berry *et al.*, 1978; Daly and Umaña, 1981; Georgescu-Roegen, 1971a; Peet, 1992; Ruth, 1993). One may argue that the openness of the biosphere in terms of energy justifies an exclusive focus on mass balance as a guiding principle for economic modelling of materials flows and processes (Young, 1991). What will be discussed here is the relationship between thermodynamics and complete and perpetual recycling, which has given rise to much debate. Recycling is implicit in aggregate production functions, such as are used in growth analysis. Here attention will be devoted to the issues of relevant time scale and waste mining in that context.

Although the use of materials balance models was propagated nearly three decades ago (see Ayres and Kneese, 1969; Kneese *et al.*, 1970), its operationalization has been mainly restricted to linear production models of the input-output type. The combination of non-linear models and mass balance conditions is rare in both theory and application. Few explicit models of the latter type have been formulated (see Georgescu-Roegen, 1971b; Gross and Veendorp, 1990; Ruth, 1993; and van den Bergh and Nijkamp, 1994). This article aims to contribute to an underpinning and further development of these more general models of production. Production function specifications will be offered that are general in the sense that they do not use particular mathematical assumptions, can represent non-linear processes, allow for substitution, explicitly address the transformation between physical and other units (functional, service, value), and are consistent with mass balance.

The organization of the article is as follows. Section 2 proposes to use a distinction between certain categories of production factors, and related to this, between direct and indirect substitution in production. Section 3 considers the notions of substitution and complementarity in relation to alternative views on capital and production factors. Section 4 examines whether often employed neoclassical production functions are inconsistent with mass balance, and links this to the wider debate on resource limits to growth. In particular, it is examined whether definite limits exist with regard to recycling opportunities and with regard to the value that can be derived from material inputs. Section 5 presents production function specifications that are explicitly consistent

---

relevant under “earthly conditions”, where the transformation between materials and energy is negligible. In other words, by approximation, on earth energy and materials are separately conserved.

with mass balance and the earlier classifications of production factors and substitution mechanisms. A final section concludes.

2. Direct versus indirect substitution and materials use in production

After presenting a general typology of production inputs, a distinction between direct and indirect substitution will be proposed. This may lead to a more subtle approach to correctly analyze and interpret changes that occur in production processes, either on a micro or macro scale, particularly changes that influence the use of materials in production. It will be argued that the present distinction between substitution processes and technological change is too crude to deal with specific relationships between the various categories of production inputs, including materials. Part of a more subtle approach is the distinction proposed by Georgescu-Roegen (1971a), between stocks, flows, funds and services. A fund differs from a stock in several respects: it generates non-durable, non-physical services; the process of generation of services does not "empty" but merely degrade the fund (tired worker, wear and tear of machines); and the allocation over time of services is restricted by the speed at which they can be generated. As opposed to this, the exploitation of a stock gives rise to a flow which has exactly the same characteristics (or quality) as the stock itself; and a stock can be discharged at any speed. Use of the stock generates an outflow that empties it. Moreover, the uses of a fund may be regarded as non-rival over time - provided its quality is maintained - while the use of a stock at any time conflicts with potential uses in the future. Finally, it should be noted that funds, flows and stocks are no absolute categories, but only defined so in a particular context.

A specific category of funds is "economic agents", which includes such aggregate sub-categories as capital and labor, and may be defined as operating to transform material and energy flows for economic purposes (production, consumption).[3] The flows are thus transformed into different qualities by the agents, thus generating economic value. The implication of mass balance is that material input flows end up as integrated or component parts of the product, or as waste output.

Adherence to these categories of inputs allows for a more precise treatment and interpretation of substitution between production factors. I would like to introduce the distinction between "direct substitution" and "indirect substitution". The first type refers to changes within a category of production factors that fulfill the same, or a similar, function in the production process. It may

[3] Note that natural resources or ecosystems that render services may be considered a subclass of funds. However, this is not of immediate interest to the present discussion.

be regarded as "replacement" of one type of production factor by another one with the same function, such as machines by labor[4], or one material by another. The second type, "indirect substitution", refers to a process involving multiple categories of production factors, which fulfill different - and often complementary - functions in the production process. In relation to materials this can be regarded as "saving" on the use of materials. This may result from working more accurately and less wastefully, via the avoidance of waste materials or the reduction of materials entering the valuable output. This may be related to longer working hours, more capital or labor input, re-organization of production, or use of more efficient techniques of production. Therefore, it is immediately related to an increase in the efficiency and productivity of the production process concerned. It should be noted that the addition of technical change does not alter this disaggregate factors/substitution framework, but just widens the choice spectrum of direct and indirect substitution, based on process or product innovation.

Absolute minimum limits to materials locked-up in products are hard to determine, as they are related to the exact service the product is to render, which in turn depends on the expected intensity of use, the quality of the product, and the time it is expected to remain in use. Applying mass balance to the case of indirect substitution implies a limited scope of change, i.e. for a given product. Irrespective of how much labor or capital is added to the production process, or how many new efficient production techniques come available, for a given product or output no more can be saved on materials use than the difference between materials entering the production process and materials locked up in the valuable output, i.e. material waste.

3. Substitutability, complementarity and definitions of capital

The purpose of this section is to argue that an aggregate analysis of substitution of materials, and limits to it, based on a distinction between natural and economic capital only, is unable to take into account the separate character of production inputs like stocks, funds and flows, as well as the various types of substitutability and complementarity relationships among these. It is quite common in environmental and ecological economics to use the opposition between substitutability and complementarity to distinguish between natural capital and economic (or produced or human-made) capital (e.g., Pearce and Turner, 1990; Costanza and Daly, 1992; Jansson *et al.*, 1994; Turner *et al.*, 1997). Sometimes other categories are added, some of which may be considered as sub-categories of economic capital, such as cultural capital (Berkes and Folke, 1992), social

[4] When they are substitutes; of course they can also be complementary, like in the case of people operating and monitoring machines.

capital, manufactured physical capital, human capital (labor and disembodied knowledge) and institutional capital. Natural capital is usually seen as encompassing functions, goods and services provided by natural environmental systems.

The degree of substitution between economic and natural capital has turned out to offer a level of abstraction at which many find it attractive to discuss the different perspectives on sustainability and sustainable development. The issue of substitutability versus complementarity of economic and natural capital is especially important to the distinction between weak and strong sustainability (see Pearce and Turner, 1990; Pearce and Atkinson, 1993). Weak sustainability is usually regarded to allow for much substitution between natural and economic capital, while strong sustainability is based on a high degree of complementarity in the context of production, consumption and welfare (see Cabeza-Gutés, 1996). An informative and correct analysis of substitution of materials on an aggregate level requires a distinction between "direct substitution", "indirect substitution", and complementarity among inputs. Related to this, the distinction of inputs should go beyond natural and economic capital, and be based on the categories funds, stocks and flows.

Regarding the environment as capital in a strict sense entails a purely economic approach (see Victor, 1991), which has been most clearly adopted and elaborated in bioeconomic approaches to fisheries and forestry management (Clark, 1990; Wilen, 1985). The capital approach to environmental stocks and funds accounts for natural resources and ecosystems in terms of the economic value of goods and services rendered by them over time. This approach has been criticized for several reasons, related to problems of aggregating natural system components measured in different physical units, and inadequately distinguishing between flows, stocks and funds in natural systems. Victor (1991) discusses the aggregation problem against the background of the famous "capital controversy" between the Cambridges, i.e. between the US neoclassical school and the UK Post-Keynesian school. Aggregate manufactured capital is assessed in value terms which assumes prices to "weigh" its different components. On an aggregate level the usual distinction between quantities and prices in economic neoclassical models is then impossible, since the unit of measurement is dependent on prices and income distribution. Instead, these four items should really be determined simultaneously, which means stepping away from the neoclassical model in which marginal products (co)determine prices which in turn (co)determine income distribution (see Harcourt, 1972). These difficulties are magnified in the notion of "natural capital", as it is not even clear in what units this type of capital is measured, so that any weighting is arbitrary. The suggestion that economic, monetary

valuation is a solution to formulating an indicator for aggregate natural capital is not very satisfactory. One reason is that it cannot cover all the indirect values related to complex and incompletely known relationships among spatially dispersed biophysical systems and processes. Likewise, alternative aggregation procedures based on specific physical, biological or environmental dimensions have the drawback of being partial or depending on unrealistic or arbitrary assumptions. Examples of such procedures are the "ecological footprint indicator of unsustainability" that tries to reduce all environmental impacts of economic activity to hypothetical land use (Wackernagel and Rees, 1996), and the "material intensity per service" (MIPS, "ecological rucksack", or intensity of material turnover) that reduces environmental impacts of a product to kilograms of any material used, processed and moved over its life-cycle, i.e. "from cradle to grave" (von Weizsäcker *et al.*, 1997, Section 9.2).

Table 1 offers a more disaggregate view on capital, substitutability and complementarity. It is based on various distinctions: between direct and indirect substitution; between stocks, flows, funds and services; and between more concrete types of substitution and complementarity related to categories of production inputs like materials, energy, throughput (energy and materials), agents (economic funds) and capital (natural and economic funds). Now the main problem with production functions and factors formulated at a very general and aggregate level seems to be that they cannot explicitly address the fact that two general types of production factors may be both complements and substitutes (e.g., labor and machines). The table illustrates that an aggregate treatment of processes and changes in the environment-economy system may miss the diversity of relationships and possible changes underlying any aggregate outcome. This is not meant to imply that aggregate analysis is irrelevant always, notably since some degree of aggregation is inevitable, certainly in any analytical approach. The point is merely that one should be careful to distinguish between various mechanisms and variables on the basis of the problem concerned. Although no absolute criterion for such a decomposition is available, some rules seem to make sense. For instance, a production function should in any case explicitly describe factors that are both complementary and variable, since production is incompletely described otherwise. One should avoid to ignore the information that is critical in prediction, interpretation or explanation. For instance, using a disaggregate logical systems structure can often be more useful for either purpose than lumping all effects together in a single elasticity coefficient, or blindly extrapolating correlation between variables over time.[5]

[5] Rose *et al.* (1996) perform a structural decomposition analysis of changes in materials use, based on comparative static changes in input-output table parameters. This can be regarded in terms of deriving

An important conclusion is that casting environmental problems in terms of the substitution between natural and economic capital on the most aggregate level, seems to neglect the essential differences between these factors of production. Economic analysis of substitution of materials should make use of the various types of substitutability and complementarity relationships as indicated in Table 1, so that a complete decomposition and understanding of changes in material use is possible.

[INSERT TABLE 1]

4. Neoclassical production functions, limits to growth and recycling

An intriguing question is whether the standard neoclassical growth models with resources are consistent with mass balance. Few dynamic models describing economic growth and change on an aggregate level have incorporated materials accounting (the exceptions, as far as I know, are d'Arge and Kogiku, 1973; Mäler, 1974; Gross and Veendorp, 1990; and van den Bergh and Nijkamp, 1994). This leads to the question whether production function specifications in neoclassical economic models denies mass balance. Daly (1997a,b) and Cleveland and Ruth (1997) think this is indeed the case. They seem to assume that an economic production function, notably the often used Cobb-Douglas (CD) function, translates physical input into physical output and therefore is inconsistent with mass balance. But production functions often translate value units into value units. This holds especially true in macroeconomics (e.g., growth theory), where all variables are in aggregate terms. The reason is that the components of aggregate variables are not homogeneous in general, so that aggregation needs to proceed via price or value weights. Microeconomic applications of production functions may translate physical into functional units (number of goods) or into value units. In either of these cases it can not be demonstrated that mass balance is harmed by using CD functions. It should be noted that Gross and Veendorp (1990) have shown that staightforward incorporation of mass balance in a Solow type growth model with production based on a non-renewable resource leads to a definite limit to growth, once all variables are interpreted in purely physical terms. Although their result is very useful as a benchmark, such a "physical approach" offers as little insight about the limits to growth as does the standard neoclassical approach using growth models with resources. The

---

characteristics from an implicit (aggregate) production function, without making stringent assumptions about the shape of the production function. The authors are capable of separating between thirteen different sources of change, including level and mix of demand, technical change associated with particular production inputs, and various substitution effects (see also Rose and Casler, 1996).

reason is that the interaction between physical and value dimensions, which is at the heart of the matter, is not really touched upon in either approach.[6]

The thesis that the neoclassical production function as used in growth theory models with environment and resources is not inconsistent with mass balance has several dimensions, some of them linking to aspects of the everlasting "growth debate". Dasgupta and Heal (1979) have used the terms "necessary" and "essential" to refer to the importance of materials in production when their supply depends on the presence of a nonrenewable resource. An input is "necessary" means that output is strictly positive only when the respective input is strictly positive (input and output are both nonnegative). A resource is "essential" when feasible consumption must necessarily decline to zero in the long run given that production uses materials from a nonrenewable resource. Note that "essential" implies "necessary" (though not vice versa).

In CD production functions formulated as $s*A^a*R^b$ inputs are necessary (s is a scale parameter, A is agents (capital or labor), R is materials inputs, and a and b are factor elasticities of output). The CD function has unbounded average products, so that it is unclear if the material input is essential. Two cases are important: if $a>b$ then the resource is not essential, otherwise the resource input is essential (cf. Solow, 1974; see also Dasgupta and Heal, 1979). These conditions do not seem inconsistent with mass balance, although they are minimal conditions in the sense that infinitesimal quantities of physical input are sufficient to produce some significant amount of output. Important, however, to this qualification is that output is in value terms (cf. the previous discussion on aggregate variables), whereas inputs are either in value or material units. The relationship between values and physical units remains a bit vague, and one can wonder if in addition to a production function a sort of transformation function is needed. This will be discussed in the next section.

[6] The notion of an aggregate (macro) production function such as used in growth theory is problematic anyway. The reason is that production factors are not independent on a macro level (as they are on a micro level). This means that an increase in one factor requires more of the other indirectly (as well of itself). This holds independently of whether we are considering aggregate labour, capital, materials or energy inputs. None of these is really a primary input, as each requires all others and itself to be produced: electricity generation requires labour, capital, materials and energy; raw materials extraction requires labour, capital and energy; etc. On a micro level independency can usually be assumed because buying an additional unit of each production factor will not measurably influence the prices or supply of each of the other factors. Since the entire economy is included in the aggregate production function, changes in the use of one aggregate factor will affect the relative prices of all other factors. The aggregate production function should reflect this. The crucial implication in the present context seems to be backward bending isoquants on a macro level. These directly imply limits to growth, as it is impossible to substitute between factors without bound, i.e. use infinitesemal positive amounts of one production factor. This idea was suggested to me by David Stern and Robert Kaufman (see Stern, 1997). Surprisingly, it seems not to have been treated in standard macroeconomic theory.

Another widely used production function specification in economics, namely the more general Constant Elasticity of Substitution (CES) function (Arrow *et al.*, 1961), is not consistent with mass balance over the entire range of its parameter values. When the CES elasticity of substitution (e.o.s.) is larger than 1 its inputs are not "necessary" (and therefore not "essential"), in which case the production function is: (i) not consistent with mass balance, namely when the actual use of materials is zero (in this case the material input is measured in physical or value terms); or, (ii) not realistic, namely when the actual use of materials is positive but its value zero, so that materials are evidently not scarce then, which is not realistic on a global scale (in this case the material input is measured in value terms). When the e.o.s. equals 1 the CES function reduces to the CD function, discussed above. And when the e.o.s. is smaller than 1 the inputs are "essential", so that mass balance is not harmed as long as the output is in value terms, irrespective of whether the inputs are in value or physical-material terms. In the latter case (e.o.s.<1) the CES function has decreasing marginal, and (asymptotically) bounded and decreasing average, products to factors. In other words, substitution of materials is very limited in this case. Dasgupta and Heal (1979, p. 200) argue that for this case economic growth based on a nonrenewable resource will come to an end. Within the class of CES production functions the "border case" of the CD function seems to leave the widest room for interpretations of the role of materials in production. It is not evident which of these static functions is most realistic as a long-run production function. Moreover, transitions between the various cases outlined, due to (technological) change, are perhaps possible - excluding the unrealistic CES function case where resources are not necessary.

Approaching the production function and mass balance from a more operational perspective, optimists will argue that a tendency to more "clean" services in GDP will improve the ratio "value of output" to "physical amount of inputs", however measured. It is difficult to provide factual or counterfactual evidence for a continuation of such a tendency (or the opposite) in the distant future. In the first place, there does not seem to be any absolute limit on the amount of economic value that can be derived from a fixed amount of throughput (energy and materials) of an economy, independent of whether "value" refers to utility or is measured via empirical indicators, such as green or sustainable GDP or ISEW (Daly and Cobb, 1989). One may argue that the marginal utility of an extra service may be decreasing for a particular type of service, but this does not necessarily carry over to all services together, i.e. on an aggregate level. Whether this requires continuous product (service) innovation is unclear; but one can imagine there is a limit to innovation in this sense. Apart from values, Ayres (1997) argues that there is no definite

finite upper limit to the service output of a given amount of materials due to the possibility of dematerialization, re-use, renovation, recovery and recycling. In conclusion, both the service output of materials processing and the value of this service output do not seem to be bounded by an identifiable absolute limit.

Let us consider some attacks on these views. From a more pragmatic perspective it may be argued that a minimum set of non-service sectors is needed to support a certain amount of final services (in value terms), and that even production of services requires a minimal amount of material inputs directly and indirectly, like a minimum amount of space, housing, physical infrastructure and energy use. This explains why there may be some shift rather than merely a reduction in environmental problems due to a transition to a more service oriented economy. The latter issue is addressed by the "Environmental Kuznets Curve" literature and Industrial Metabolism and Ecology approaches (e.g., Ayres and Ayres, 1996). The latter argue that environmental problems are not so much solved but merely shifted. In essence this is based on insights of applied materials accounting which implies, for instance, that reduction of pollutive emissions to air will, in the absence of changes in the use of virgin materials and energy resources, lead to an increase of emissions to other environmental mediums. This point has also been noted in the literature on transferable externalities, which seems consistent with, though not explicit about, mass balance (Bird, 1987; Shogren and Crocker, 1991).

Some authors, notably Georgescu-Roegen, have tried to counter the optimistic view via another route, namely by referring to the impossibility of complete and perpetual recycling (the "fourth law" of thermodynamics). Of course, if the neoclassical production function is aggregate as it is in growth theory, then it should include also recycling. Many authors have discussed the implications of thermodynamics for recycling at a very abstract level (for an overview, see Section 2.3 in van den Bergh, 1996). This discussion is hard to judge for non-experts as it is quite technical, but it seems clear that 100% recycling is impossible, if not for fundamental-theoretical reasons (physics), then in any case for all sorts of practical reasons (technology, economics, policy, information). However, this is not the end of the matter. On a more concrete level, there are two important elements to be considered.

First, as stated by Ayres (1997), the thesis of Georgescu-Roegen that materials recycling always goes along with losses, should be complemented by the insight that given the large availability of potential energy a large part of materials waste can be used as a resource ("waste mining"). Even though not all of the material waste can be recycled or used at the same time, a positive and non-negligible part of it can be set to use, where it should be noted that this part is

continuously changing. Of course, in order to find the net value derived from a certain use of materials, the costs of waste mining and recycling have to be included, which may rise to significant heights. Second, even as recycling becomes more difficult, energy-intensive and costly, it is still not clear whether this creates any serious physical limitations to the world's economy over a relevant period. Even Daly has admitted that his steady-state economy concept is not possible on an infinite time scale. Similarly, Daly's "optimal scale" should be defined for a specific finite time horizon. So, in essence, the facts that our earth is an energetically open system and that we are not really concerned with an infinite time horizon - for indeed this in any case means the all overruling heat death of the universe - together imply that thermodynamic limits are not so serious and relevant as might be judged from a conceptual-theoretical perspective.[7]

Related to the time horizon issue is the discussion between Daly (1997a,b) and Solow (1997) and Stiglitz (1997). The difference between growth "optimists" and "pessimists" may be largely due to different time horizons adopted[8]. Solow (in his answer to Daly's second question) indeed admits that optimism can be due to the short time horizon and fact that until now we have not yet reached severe limits ("the fault of extrapolation"). Stiglitz is even more clear, explicitly acknowledging the extremely limited time horizon of the growth theory models (60 years in his interpretation). This is a surprising response in view of the great deal of attention given to mathematical details related to infinite time horizons in growth theory. This "inconsistency" becomes even more clear when it is realized that some of the growth optimism has been fueled by moving theoretical neoclassical constructs to the extreme limits of substitution. This is certainly not consistent with rather short time horizons of less than 50 or 100 years.

Concluding, both on an abstract and a concrete level of relations between economic production and physical flows it is difficult to provide unambiguous support for two important theses: that there is a limit to the amount of value that can be derived from a finite amount of physical resources; and that mass balance is not satisfied in economic models. The next section provides new models of production that allow for a sort of decomposition of the various mechanisms underlying changes in the "materials/values relationship".

[7] Georgescu-Roegen certainly realized this but somehow seemed not to want to give too much attention to this. His underscoring of the openness of the earth in terms of energy has often been noted by critics (e.g., Young, 1991).

[8] This was our conclusion in a Dutch language paper in 1995, translated and revised in English as van den Bergh and de Mooij (1997).

5. Production factors and substitution in mass balance production functions

Mass balance models are not new but have remained restricted mainly to linear production models. Here I will devote attention to the combination of mass balance and a more general class of non-linear production functions, that represent transformations between value and physical units. Only few models have been developed along such lines, the one by Georgescu-Roegen (1971b) being closest to the approach adopted hereafter. The latter will use the earlier made classifications of production factors and substitution.

With respect to formalizing the mass balance principle in economic-ecological models, the following remarks are in order: either all variables should be in mass units, or transformations must be modeled between (variables in) mass units and other units; and, mass balance conditions can be specified for economic variables in the economic system, for ecological/physical variables in the environmental system, or as a supplement to descriptions of economic-environmental interactions. Below, different ways are shown to formulate production functions that satisfy the mass balance principle.[9] [10]

The production process may be envisioned as a transformation by agents of material inputs into goods and waste outputs. In line with the earlier discussion in Sections 2 and 3, one may expect considerable potential for substitution between sub-categories of agents of transformation - labor and capital - since they serve a similar function in the production process; therefore, they are aggregated into the variable agents (A).[11] Direct substitution is relatively easy within the category of resource inputs to production (R), i.e. direct substitution. Indirect substitution between the categories of agents and resources is limited. An increase in the use of agents may reduce the amount of waste output, i.e. the case of indirect substitution. This is not assumed a priori; it follows from application of the mass balance condition to the production function.

Mass balance can be formalized in a static sense as: (i) the inequality $R>Q$, as a minimal consistency condition, where R and Q denote the levels of material input and useful output from production (and , respectively; or a more strict condition, such as $R>Q+x$, based on knowledge of technical or physical constraints, where x is a lower bound on waste residuals from production; or, (ii) the equality $R=Q+W$, where W is the actual level of waste residuals from production.

---

[9] The following is a generalized version of the approach presented in the introduction to van den Bergh and Nijkamp (1994). The latter includes a discussion of inequality constraints and refers to optimization problems, and an application of mass balance production functions in a multisectoral growth model.

[10] All variables and functions presented take non-negative values only.

[11] In a dynamic analysis the distinction between the two types of actors is relevant to address processes of capital accumulation and labour/population dynamics, as well as accounting for locked-up resource material in capital (Faber *et al.*, 1987; van den Bergh and Nijkamp, 1994).

The most general way of formulating relationships between inputs and outputs of production is via an implicit function. An example is $G(Q,W,A,R,N,t)=0$, with Q goods output, W waste output, A agents input, R resource or material input, N environmental conditions (soil quality, climate, etc.) and t a time index, to reflect changes in the relationship over time. A less general but interesting formulation separates between inputs and outputs, for instance: $G(Q,W)=H(A,R,t)$. This can be regarded as an expression for a multi-output system, in this case with valuable or desired and waste output. Later on we will see more general expressions.

A mass balance condition may be added to the foregoing formulations of production relationships: $R=T(Q,A,t) + W$. The transformation function $T(\cdot)$ is added to take care of translation between physical and value or different value units. This specification is general in the sense that if Q is in value terms, than T translates this into material units. If Q is in material terms (in a micro context), then T is simply an identity function. Notice that when the effect of A on T is absent, then it is implicitly assumed that the relationship between materials and value is fixed, though not necessarily constant. If A and t have non-zero effects on the values T takes, then it is possible to allow for variable relationships. For instance, one can argue that an increase in the input "agents" gives rise to savings of materials (indirect substitution in terms of Section 2), i.e. an increase in A implies a decrease in R, for given Q. In the following specifications we will write T only as function of Q, to keep formulations simple, although clearly more general representations are possible.

The parameter t in the previous and following production and transformation functions denotes the change over time resulting from mainly technical progress, which has a different interpretation when output is in value terms than when it is in physical terms. The time parameter in the transformation function T has a similar meaning as in the production function F, i.e. it refers to more efficient (materials-savings) techniques over time. So it should be consistent with thermodynamics. However, the fact that there is no absolute limit to the amount of value and services derived from a given amount of materials, as discussed in the previous section, implies that T is never zero as long as Q is positive.[12]

The equations in (1) show a general explicit relationship between the output of goods Q, and all aggregate factors involved in its production, i.e. A, R and W. The mass balance principle is

[12] A complete dynamic representation of the production process given the present specification allows for a vintage approach that recalls the amount of materials that have entered the useful output produced at each point in time. Such a vintage approach is required when considering materials accounting over time in the presence of long delays between production and ultimate waste generation after consumption

stated as the equality condition (1b). In the formal representation *F(ɣ)* the conceptual difference between *A*, *R* and *W* is not made explicit, since all variables just enter the function as general arguments. Their difference becomes clear only after the mass balance constraint is added. All partial derivatives of *F(ɣ)* are positive. More agents (capital and labor) and more resources contribute positively to output (in value or physical terms). More waste is also assumed to have a positive direct effect on output, which may be interpreted as less concern for reducing waste making production easier or cheaper. "Easier" may imply more useful output in material terms, and "cheaper" the same in value terms.[13]

$$Q = F(A,R,W,t), \qquad \textbf{1a}$$

$$R = T(Q,A,t) + W. \qquad \textbf{1b}$$

A "recursive waste production function" can be derived from the relationships in (1), namely as $W = G(A,R,Q,t) = R - T(Q,A,t) = R - T(F(A,R,W,t),A,t)$. This can be considered as an "ex post" relationship between the production functions for useful output and waste, i.e. after the application of the mass balance condition. This relates to the discussion on indirect substitution in Section 2. The conditions in (1) may be substituted to obtain (2).

$$Q = F(A,R,R - T(Q,A,t),t) = F(A,T(Q,A,t) + W,W,t). \qquad \textbf{2}$$

This shows that the total description of the production system, including material flows, leads to an implicit relationship. This reflects the multi-output character of production, as well as the two dimensions, i.e. products (or values) and materials.

The second type of formulation of production subject to mass balance shown in (3) starts with two separate production functions for goods and for waste that are related via their dependence on *A* and *R*.

---

of the good (see Kandelaars and van den Bergh, 1997). This is especially relevant for durable goods like buildings, furniture, cars, refrigerators, etc. (see Noorman and Uiterkamp, 1998).

[13] Notice that the negative effect of waste or pollution on production is not dealt with here. That the partial derivative of *F* with respect to *W* is positive must be interpreted as that for a given combination of actors and resources, and in the absence of mass balance conditions, production can increase by working quicker and leaving more waste per unit of actor. The negative feedback of waste and accumulated waste (stock of pollution) is outside the scope of the production function discussion here, and would require a (dynamic) systems model of the economy, the environment and their mutual influences.

$$\begin{aligned} Q &= F_1(A,R,t), \\ W &= F_2(A,R,t), \\ R &= T(Q,A,t) + W. \end{aligned} \qquad \textbf{3}$$

This can be interpreted as a sequence of processes: an agent level *A* separately determines the levels of useful and waste outputs *Q* and *W*, and the sum of these in mass terms gives the resource requirement *R*. Waste as a function of *Q* and time, say *W=B(Q,t)*, is a special case of (3), namely for $F_2=B(F_1)$. This can be interpreted as choosing a level for *A*, which gives a value of *Q*, which in turn implies a value of *W*.

The set of equations in (4) shows disaggregation into separate agents A_1 and A_2 allocated to production and to an activity $F_3(\chi)$ that reduces the amount of waste resulting from the (main) production process that is now represented by the multi-output system { $F_1(\chi)$, $F_2(\chi)$ }.

$$\begin{aligned} Q &= F_1(A_1,R), \\ W &= F_2(A_1,R) - F_3(A_2), \\ A_1 + A_2 &= A \\ T(Q,A,t) + W &= R, \end{aligned} \qquad \textbf{4}$$

As waste is always non-negative, the condition $F_3 \pounds F_2$. holds for all combinations of A_1 and A_2.

A third type of formulation of mass balance production functions is by way of a function *F* that automatically satisfies the consistency restriction of mass balance, i.e.: $F(A,R) \pounds T^{-1}(R)$ for all values of $R>0$. This "integrated" the production, transformation and mass balance relationships. It can be accomplished in two ways. The first is shown in (5), and is characterized by taking the minimum of any general production function and some share $a(t)$ (with $0<a(\chi)\pounds 1$ and $a'>0$) of the resource input that is based on optimum technical efficiency of materials use (note that this is process-specific, first-law efficiency, given the state of technology at time *t*). This mass balance production function incorporates the complementarity between materials and other production factors as discussed in Table 1, and can be regarded as a generalized Leontief production function. The transformation function appears twice in (5) to allow for a comparison of the various units.

$$F(A,R,t) = T^{-1}(\mathrm{MIN}\{T(G(A,R,t)), a(t) * R\}, t). \qquad \textbf{5}$$

Technological change may affect either the output in functional or value units (via *t* in *G* or in T^{-1}) or in material terms (via *t* in *a( )*). It should be noted that *t* appears at three places in (5). The first one denotes changes in the potential for indirect substitution (saving of materials) due to technological change, learning, reorganization, etc. The second time index refers to changes in purely technological parameters (e.g. heat of processes) that determine absolute limits at each point in time (first-law efficiency); such changes are limited by absolute thermodynamic limits (i.e. process-independent, second-law efficiency). The third time index refers to broader changes in the relationship between functional or physical output and the value of output. The precise interpretation is also dependent on the scope of the production function. For instance, if it is aggregate like in growth models, it may include changes in the structure of the economy, new activities, recycling, etc.

The second "integrated" mass balance production function is shown in (6). It is based on a resource efficiency coefficient *r(ϰ)* which can be regarded as a variable coefficient that relates useful output to material input. This coefficient is increasing in *A* and *t*, and decreasing in *R*, and it takes values in between zero and one. An example is $r = aA/(A+bR)$, with *a* and *b* positive, and *a* smaller than one. The parameter *a* can be interpreted as the optimal efficiency of resource use in production, attainable only via extremely large input of actors. Resource efficiency in production is thus assumed to be improved either by increasing the intensity of the production activity factors relative to the resource input - which may be indicated by a higher activity-resource ratio *A/R* - or by technological progress (via *t* at multiple places, similar to 5).

$$Q = T^{-1}(r(A,R,t) * R, t). \qquad \mathbf{6}$$

Notice that in (6) R appears twice, which allows for increasing returns to material inputs, even if *r(ϰ)* is a concave function. This is consistent with the point made in Section 4, i.e. that there is no definite upper limit to the (value of a) service output or of a given amount of materials.[14]

Since the mass balance condition implies a linear (in)equality, it suits linear types of models well, such as fixed proportions (input-output or Leontief) and linear production functions. The distinction between the three general types of representations of production with mass balance conditions is useful in providing different insights about the specification and the relationship between production factors in the production process. It seems that the idea of the production

[14] This point was noted by a referee.

process as a system of parallel and sequential activities can provide a good starting point for further elaboration of these “mass balance production functions”. The specification in (4), for instance, has both parallel and sequential (serial) elements. A multi-output production system generating useful and waste output satisfies the parallel character, while “ex post” calculation of materials inputs or resource requirements as well as the reduction of waste satisfy the sequential character. This type of system may also be used to frame optimization decisions, for instance, related to minimizing or maximizing inputs or outputs of any type.

An interesting application of production specifications combined with mass balance is to analyze the implications of complementarity and direct and indirect substitution for long term development, with capital (agent) accumulation, technical progress, production with renewable and non-renewable resources, emission of waste residuals, and recycling. These steps can be made by adopting particular models as outlined above. Evidently, each of these models has specific advantages and disadvantages. The main purpose of this section was to show that a more explicit treatment of mass balance and different types of substitution is possible while not moving too far away from formulations and assumptions economists are familiar with.

6. Conclusion

The relation between values, substitution, production factors and thermodynamics in production functions is a somewhat neglected topic in environmental and ecological economics. The literature is mostly confined to abstract discussions hardly offering operational approaches. Moreover, there is generally a confusion in the use of the terms “substitution” and “capital”, which are often undefined, subject to cryptic use, and too aggregated to arrive at a precise and informative model of physical interactions between the environment and the economy. Here a distinction was proposed between direct substitution or “replacement”, and indirect substitution or “saving”. Linked to this distinction, a classification of substitutability and complementarity between flows, agents and capital was presented. The opposition of natural and economic capital seems too aggregate and abstract for an adequate analysis of substitution processes. It does not offer sufficient detail about stocks, flows, funds and services. The neoclassical production function was argued not to be necessarily inconsistent with mass balance. However, it provides little information on what types of mechanisms “de-link” value of output from materials inputs. On a more intuitive level, the growth debate was discussed against this background. In particular, the thesis of Georgescu-Roegen that complete and perpetual recycling is impossible (the “fourth law”) was examined. A safe conclusion seems to be that recycling may relax resource and

thermodynamic limits over a relevant time horizon, which may be extremely long but not infinite. To overcome the mentioned lack of realistic detail of neoclassical production functions, a number of general production functions were formulated that start from distinctions of production factors and substitution mechanisms, and are consistent with mass balance. These approaches should not be followed for all problems and questions. Their main advantage is where issues are at stake that relate to physical limits to economic processes and to policies regulating materials flows. A subsequent step is therefore to employ the proposed production formulations in analytical and empirical models and studies of economic growth, material-product chains, recycling, waste management, dematerialization and environmental policy.

References

Arrow, K.J., H.B. Chenery, B.S. Minhas and R.M. Solow, 1961. Capital-labor substitution and economic efficiency. *Review of Economics and Statistics*, vol. 43: 225-250.

Ayres, R.U., 1978. *Resources, Environment and Economics: Applications of the Materials/ Energy Balance Principle*. Wiley-Interscience, New York.

Ayres, R.U.., 1997. Comments on Georgescu-Roegen. *Ecological Economics*, vol. 22: 285-287.

Ayres, R.U. and A.V. Kneese, 1969. Production, consumption and externalities. *American Economic Review*, Vol. 59: 282-97.

Ayres, R.U., and L.W. Ayres (1996). *Industrial Ecology: Closing the Materials Cycle*. Edward Elgar, Cheltenham, UK, and Brookfield, USA.

Berkes, F., and C. Folke, 1992. A systems perspective on the interrelations between natural, human-made and cultural capital. *Ecological Economics*, vol. 5: 1-8.

Berry, R.S., P. Salamon and G. Heal, 1978. On a relation between economic and thermodynamic optima. *Resources and Energy*, Vol. 1: 125-137.

Bird, P.J.W.N., 1987. The transferability and depletability of externalities. *Journal of Environmental Economics and Management*, vol. 14: 54-57.

Cabeza-Gutés, M., 1996. The concept of weak sustainability. *Ecological Economics*, vol. 17: 147-156.

Clark, C.W., 1990. *Mathematical Bioeconomics: The Optimal Management of Renewable Resources.* 2nd ed. Wiley, New York.

Cleveland, C.J., and M. Ruth, 1997. When, where and by how much do biophysical limits constrain the economic process? A survey of Nicholas Georgescu-Roegen's contribution to Ecological Economics. *Ecological Economics*, vol. 22: 203-223.

Costanza, R. and H.E. Daly, 1992. Natural capital and sustainable development. *Conservation Biology*, vol. 6: 37-46.

Daly, H.E., 1997a. Georgescu-Roegen versus Solow/Stiglitz. *Ecological Economics*, vol. 22: 261-266.

Daly, H.E., 1997b. Reply to Solow/Stiglitz. *Ecological Economics*, vol. 22: 271-273.

Daly, H.E. and W. Cobb, 1989. *For the Common Good: Redirecting the Economy Toward Community, the Environment and a Sustainable Future.* Beacon Press, Boston.

Daly, H.E. and A.F. Umaña (eds.), 1981. *Energy, Economics and the Environment.* AAAS Selected Symposia Series, Westview Press, Boulder, Colorado.

d'Arge, R.C. and K.C. Kogiku, 1973, Economic growth and the environment. *Review of Economic Studies*, vol. 40: 61-77.

de Bruyn, S.M., en J.B. Opschoor (1997). Developments in the throughput-income relationship: theoretical and empirical observations. *Ecological Economics* 20: 255-268.

Faber, M., H. Niemes and G. Stephan, 1987. *Entropy, Environment and Resources: An Essay in Physico-Economics*. Springer-Verlag, Heidelberg.

Georgescu-Roegen, N., 1971a. *The Entropy Law and the Economic Process*. Harvard University Press, Cambridge, MA.

Georgescu-Roegen, N., 1971b. Process analysis and the neoclassical theory of production. In: N. Georgescu-Roegen, *Energy and Economic Myths*. Pergamon, New York, 37-52.

Gross, L.S. and E.C.H. Veendorp, 1990. Growth with exhaustible resources and a materials-balance production function. *Natural Resource Modeling*, Vol. 4: 77-94.

Harcourt, G.C., 1972. *Some Cambridge Controversies in the Theory of Capital*. Cambridge University Press, Cambridge, UK.

Hartwick, J.M., 1977. Intergenerational equity and the investing of rents from exhaustible resources. *American Economic Review*, Vol. 67: 972-974.

Jansson, AM., M. Hammer, C. Folke and R. Costanza (eds.), 1994. *Investing in Natural Capital: The Ecological Economics Approach Sustainability.* Island Press, Washington D.C.

Kandelaars, P.P.A.A.H. (1998). *Economic Analysis of Material-Product Chains: Models and Applications.* Ph.D. Thesis. Tinbergen Institute and Thesis Publishers, Amsterdam.

Kandelaars, P.P.A.A.H., and J.C.J.M. van den Bergh, 1996. Analysis of materials-product chains: Theory and application. *Environmental and Resource Economics*, vol. 8: 97-118.

Kandelaars, P.P.A.A.H., and J.C.J.M. van den Bergh, 1997. Dynamic analysis of materials-product chains: An application to window frames. *Ecological Economics*, vol. 22: 41-61.

Kneese, A.V., R.U. Ayres and R.C. D'Arge, 1970. *Economics and the Environment: A Materials Balance approach*, Johns Hopkins Press, Baltimore.
Mäler, K.-G., 1974. *Environmental Economics: A Theoretical Inquiry*. Johns Hopkins University Press, Baltimore.
Noorman, K.J., and T.S. Uiterkamp (eds.), 1998. *Green Households? Domestic Consumers, Environment and Sustainability*. Earthscan, London.
Pearce, D.W., and G.D. Atkinson, 1993. Capital theory and the measurement of sustainable development: an indicator of 'weak' sustainability. *Ecological Economics*, vol. 8: 103-108.
Pearce, D.W. and R.K. Turner, 1990. *Economics of Natural Resources and the Environment*. Harvester Wheatsheaf, New York.
Peet, J., 1992. *Energy and the Ecological Economics of Sustainability.* Island Press, Washington D.C.
Rose, A., and S. Casler, 1996. Input-output structural decomposition analysis: A critical appraisal. *Economic Systems Research* 8: 33-62.
Rose, A., C.-Y. Chen and G. Adams, 1996. Structural decomposition analysis of changes in material demand. Unpublished draft.
Ruth, M., 1993. *Integrating Economics, Ecology and Thermodynamics*. Kluwer Academic Publishers, Dordrecht.
Shogren, J.F., and T.D. Crocker, 1991. Cooperative and noncooperative protection against transferable and filterable externalities. *Environmental and Resource Economics*, vol. 1: 195-214.
Solow, R.M., 1974. Intergenerational equity and exhaustible resources. *Review of Economic Studies*, vol. 41: 29-45.
Solow, R.M., 1997. Reply. *Ecological Economics*, vol. 22: 267-268.
Stern, D.I., 1997. Limits to substitution and irreversibility in production and consumption: a neoclassical interpretation of ecological economics. *Ecological Economics*, vol. 21:197-216.
Stiglitz, J.E., 1997. Reply. *Ecological Economics*, vol. 22: 269-270.
Turner, R.K., C. Perrings and C. Folke, 1997. Ecological economics: paradigm or perspective. In: J.C.J.M. van den Bergh and J. van der Straaten (eds.), *Economy and Ecosystems in Change: Analytical and Historical Approaches.* Edward Elgar, Cheltenham, UK, and Brookfield, USA.
van den Bergh, J.C.J.M., 1996. *Ecological Economics and Sustainable Development: Theory, Methods and Applications.* Edward Elgar, Cheltenham, UK, and Brookfield, USA.
van den Bergh, J.C.J.M., and R. de Mooij, 1997. An assessment of the growth debate. In: J.C.J.M. van den Bergh (ed.), *Handbook of Environmental and Resource Economics*, Edward Elgar, Cheltenham, forthcoming.
van den Bergh, J.C.J.M., and P. Nijkamp, 1994. Dynamic macro modelling and materials balance. *Economic Modelling*, Vol. 11: 283-307.
Victor, P.A., 1991. Indicators of sustainable development: Some lessons from capital theory. *Ecological Economics*, Vol. 4, 191-213.
Wackernagel, M. and W. Rees, 1996. *Our Ecological Footprint: Reducing Human Impact on the Earth.* Illustrated by Phil Testemale. The new catalyst bioregional series vol. 9. Gabriola Island, BC and Philadelphia, PA: New Society Publishers.
Weizsäcker, E. von, A.B. Lovins en L.H. Lovins (1997). *Factor Four: Doubling Wealth - Halving Resource Use, A Report to the Club of Rome.* Earthscan, London.
Wilen, J.E., 1985. Bioeconomics of renewable resource use. In: A.V. Kneese and J.L. Sweeney (eds.), *Handbook of Natural Resource and Energy Economics*, Vol. 1. North-Holland, Amsterdam.
Young, J.T., 1991. Is the entropy law relevant to the economics of natural resource scarcity? *Journal of Environmental Economics and Management*, Vol. 21: 169-179.

Table 1. Substitutability, complementarity and production factors

Mechanism type	Input categories involved	Underlying production factors	Direct empirical testing	Examples
Substitutability				
1. Direct materials substitution	flows	multiple materials	possible, easy	substitution of one type of material for another, with the same or a similar function
2. Indirect materials substitution	funds and flows	agents and materials	possible, difficult	more efficient use of materials in production (saving) via more capital or labor use, or via new process technology
3. Direct energy substitution	flows/fluxes	multiple energy sources or fluxes	possible, easy to difficult	photo-voltaic cells instead of fossil fuel energy
4. Indirect energy substitution	funds and flows/fluxes	agents and energy flux	possible, difficult	new production technique that uses less energy per unit of output
5. Direct capital substitution	funds, flows, services	multiple agents	possible, easy	more machines, less labor
6. (In)direct capital substitution	funds	manufactured capital and natural capital	extremely difficult	Hartwick (1977) approach: investing revenues of resource extraction in economic capital;
Complementarity				
1. Agent complementarity	services or funds	multiple agents	possible, easy	machines operated by humans; land worked by tractors
2. Throughput complementarity	flows/fluxes	multiple materials or energy	possible, easy to difficult	composite materials; a fixed relation between energy and materials use, like in chemical processing
3. Factor complementarity	funds, flows and services	agents and materials or energy	possible, difficult	energy needed to operate machines; fixed machine-hours and material content of products
4. Capital complementarity	funds	manufactured capital and natural capital	very difficult to impossible	tractors, arable land and groundwater in agriculture; boats and nature in recreation