Aggregate Production in Macroeconomic Theory

Or: The Curious Case of the Missing Equation – A Macro Mystery

Andreas Pollak[*]

Department of Economics, University of Saskatchewan

December 21, 2025

Abstract

This paper argues that, for a firm sector to have the expected property of aggregate scale invariance while being consistent with our experience regarding economic growth, it requires a micro structure that endogenously determines the number of production units in a scale-invariant way. A simple and standard free-entry, competitive-market model meets this requirement, while offering several additional benefits compared to the firm sector models most commonly used in the macroeconomic literature. To demonstrate the advantages of the proposed approach, I present two example applications.

The first one develops a production sector that exhibits stable productivity growth resulting from firms' endogenous R&D investment. While only containing standard building blocks, the model structure is entirely new to the growth literature and offers a number of desirable features. It is the first R&D-based growth model inherently free of any unwanted scale effects or knife-edge conditions, while being far less complicated than most popular frameworks of endogenous growth. It does not rely on a linearity of production in an accumulable factor and retains the empirically well-supported medium-term capital dynamics associated with the neoclassical growth model.

As the second example, the proposed firm-sector structure is integrated into an otherwise standard two-sided labour market search framework. The resulting model is again simpler than the comparable Mortensen-Pissarides model, replaces the Nash-bargaining approach to wage determination not commonly used in other macro settings with standard marginal-product wages, and is easily calibrated to match empirical volatility patterns of labour market variables.

More generally, the paper shows that the notion that certain features of economies, such as R&D-driven growth or search unemployment, are inherently incompatible with a competitive setting, is not correct. A free-entry, competitive market setting can accommodate such model characteristics without the need for mark-ups, rents, market power or domain-specific modelling strategies such as scale effects in aggregate production or linearity in an accumulable factor.

Keywords: aggregate production, endogenous growth, search and matching

JEL Classification Numbers: E1, E2, J64

[*]Address: 9 Campus Dr., Saskatoon, SK, S7N 0P9, Canada; telephone: +1 (306) 966-5221; e-mail: *a.pollak@usask.ca*.

1 Introduction

The aggregate production function describes how much output national economies or similar units of aggregation generate in different circumstances. It does so by mapping a set of inputs used in production as well as possibly some productivity parameters to an aggregate measure of output. As such, it is primarily a formalization of the empirical patterns found in national income accounts and similar aggregate datasets. It captures the surprisingly simple and stable regularities observed in the average outcomes generated by a complex and ever-changing production sector, where a myriad individual producers engage in a variety of activities and interact with each other, the household sector, the government, and the international sector through a complex network of market and non-market relationships.

Yet, the use of the aggregate production function in macroeconomic theory goes far beyond this interpretation as a high-level description. Aggregate production outcomes are frequently identified with the production possibilities of a representative firm, or characteristics of production sector are taken as a blueprint for modelling individual firms, then often inferred to operate under a constant-returns-to-scale (CRTS) Cobb-Douglas technology. In other instances, ad-hoc assumptions are made either directly about the properties of the aggregate production function or the characteristics of the firm sector with immediate implications for aggregate production patterns. This is usually done to implement or support model mechanisms that rely on certain features of the production technology or particular firm behaviour.[1] This paper argues that these approaches can lead to inconsistencies with aggregate empirical patterns and significantly limit the ability to uncover mechanisms tied to the internal structure of the firm sector.

In what follows, we will be asking how patterns in aggregate outcomes combined with the hypothesis that at least part of observed productivity improvements result from deliberate investment in technology constrain the structure of the production sector. We will see that this approach points towards a set of specific firm sector characteristics that are

[1] Examples include monopolistic competition among a given mass of firms or the production sectors typically used in models of endogenous growth.

consistent with a simple model of perfect competition under free entry and exit. To demonstrate the usefulness of modelling production this way, we apply the idea to two areas of macroeconomics, economic growth and labour market search.

This paper is organized as follows. Section 2 discusses the characteristics we expect aggregate production to have; it then shows how these desirable characteristics constrain the structure of the firm sector under some simple assumptions and proposes a particular model that meets the requirements.

Section 3 applies the concept to the area of endogenous growth. We develop the core of a model where productivity growth is endogenously determined as the result of firms' optimal investment in technology improvements. Despite not using any unfamiliar or non-standard components, the model structure is entirely new to the literature and has many desirable features. The resulting model is, at the aggregate level, consistent with the properties of a standard neoclassical production sector, while delivering ongoing productivity growth driven by decisions at the firm level. There are no unrealistic scale effects, no knife-edge conditions. The model bridges the gap between the model worlds of exogenous and endogenous growth theory.

We then take a closer look at the Mortensen-Pissarides model of search in the labour market in section 4, identifying the linear firm-level production technology as the main factor requiring the unusual structure of the framework compared to other macroeconomic models. Modifying the firm sector to be consistent with the requirements derived in section 2, while leaving other model components such as the aggregate matching function and the job-separation process unchanged, not only considerably simplifies the model, but also makes it more regular and general. In addition to that, we show that the resulting search model can easily be calibrated to reproduce cyclical patterns of unemployment and vacancies.

Finally, section 5 briefly summarizes and discusses our results.

2 Aggregate Production

Our objective is to put constraints on the properties of a production sector that are suitable for the contexts in which macroeconomic models are commonly used. We start by establishing some criteria, which we group into two definitions. The first definition is entirely model agnostic and compiles a set of attributes we ought to require of an aggregate production function. The second definition focuses on a very specific model capability that has been difficult to reconcile with our expectations regarding aggregate properties.

With regards to the criteria relating to the aggregate production function, the first requirement is that production characteristics are independent of scale. It should be possible to reason about a firm sector irregardless of how large the economy under investigation is, or whether a single national economy, a part thereof or an aggregate of multiple counties is being studied.

Another source of constraints on aggregate behaviour are the stylized facts of economic growth. Kaldor's list of empirical regularities[2] includes a steady growth of labour productivity, a stable interest rate, a stable capital-output ratio, and steady factor shares. To put this differently, the economy exhibits a stable rate of labour productivity growth while the capital stock grows at the same rate as aggregate output. The notion that capital remains proportional to output as the economy grows is normally referred to as balanced growth. Among the many regularities that can be observed among growing economies, our focus will be on the parallel trends of output and capital given stable labour productivity growth, as these aspects of the data link available inputs to output at the aggregate level without imposing any restrictions on variables that may not be robust to lower-level modelling choices, such as the mechanism determining factor incomes.

Definition 1 (Scale-Invariant Production Function) *An aggregate production function that is strictly increasing in both capital and a non-capital input will be said to be scale invariant if it meets the following three criteria:*

[2] Kaldor (1961)

(C1) At any point in time, the aggregate production function exhibits static constant returns to scale in all factors that are rival in production.

(C2) It is consistent with balanced growth, in the sense that any capital-output ratio that is possible in one period can be maintained indefinitely while keeping all non-capital inputs stable.

(C3) While productivity may vary with time or depending on input ratios, any such changes must be independent of the scale of production.

Definition 1 requires three dimensions of scale independence.[3] First, everything else equal, the size or change of size of the economy being described measured in terms of labour input never matters for productivity. This means that the same, rich enough aggregate production function can be used in principle to describe a single national economy, a part of it or a group of countries, and irrespective of whether the size of the units under investigation changes. Second, to the extent that labour productivity changes over time, it must be possible that both variables measured in units of goods, production and capital, scale proportionally. Here, capital should be considered a broad measure encompassing the value of all the wealth used in production. Finally, any productivity variation resulting from changes in the production function must also be invariant to whether one large homogeneous economy is under consideration or a structurally identical subset of it.

While criteria (C1) to (C3) as axiomatic requirements are not testable per se, their relevance and suitability for the most common use cases for aggregate production functions is an interesting issue, which is addressed in appendix B.

Broadly speaking, definition 1 requires aggregate production to be described by a CRTS production function, optionally with labour-augmenting technological progress that is independent of the scale or growth of labour input.[4] Any plausible specification would likely look like a variation of the neoclassical production function with exogenous technological

[3] A more formal version of this definition, followed by a characterization of aggregate production functions that meet its requirements, is presented in appendix A.

[4] See appendix A.

progress. There is no scope for any features supporting endogenous growth at the level of the aggregate production function. This, however, is not a problem, as we will argue below that mechanisms leading to productivity enhancements will be located at a level of abstraction below aggregate production, *within* the firm sector.

Definition 2 (Growth-Compatible Production Sector) *A model of the production sector will be said to be growth-compatible if its aggregate characteristics are consistent with Definition 1, and:*

(C4) The model structure is compatible with R&D-driven growth at a stable rate. Any given improvement in technology requires expenditure in the private sector, the amount of which is independent of the scale of production.

While definition 1 is an entirely structure-independent description of empirical features of a production sector, definition 2 adds a requirement pertaining to actual model mechanics. Note that the definition only requires *compatibility* with R&D-driven, balanced growth; there is no need for a model economy to actually exhibit productivity changes. Section 4 will show an example of a setting unrelated to economic growth that benefits from modelling the production sector in line with definition 2.

Definition 2 is focused on growth-related characteristics of a production sector. The reason for this is not that every macroeconomic model needs to be a growth model. The rationale is rather that growth patterns are among the most robust aggregate observations we have, while at the same time being hard to incorporate into standard production frameworks, making them particularly attractive as a criterion for selecting models. If a firm sector model is *structurally incompatible* with important empirical patterns, it is unlikely to be a good representation of its real-world counterpart and may be suspect of relying on mechanisms that differ from the ones governing the true data-generating process. This may affect outcomes and predictions even if the model is applied to questions entirely unrelated to growth. In a Popperian sense, everything else equal, an available model consistent with a wider set of observations is preferable, even if the criteria for model selection are not directly related to the research question at hand.

The remainder of this section discusses what constraints definition 2 imposes on the *internal* structure of the production sector, in particular with regards to how the size and number of individual production units[5] must scale over time and with the size of the economy.

2.1 Production Microstructure

Consider a firm sector composed of a number (or mass) μ of identical firms. The firms' production function in period t is $f_t(k, l; \kappa)$. There are two types of production factors. The scalar k is the value of the capital used in production, including both physical capital and intangible assets, and l is a vector of non-capital inputs[6]. We will interpret and refer to l as labour. The parameter κ, which stands for the aggregate capital-labour ratio, is included to allow for a capital externality.[7] The production function can change between periods.

Aggregate production in the economy is then given by $Y = F_t(K, L) = \mu f_t(k, l; \kappa)$ for aggregate factors $K = \mu k$ and $L = \mu l$. We will sometimes normalize factors by a measure of labour input $n = |l|$ and $N = |L| = \mu n$, $(\kappa, \ell) = \frac{1}{n}(k, l) = \frac{1}{N}(K, L)$.[8]

[5] See appendix C for comments on the relevant unit.

[6] This could include different qualities of labour as well as other inputs.

[7] This can be the direct positive productivity effect of aggregate capital on individual firms assumed in Romer's influential AK model (Romer (1986)), but it can also account for a range of other mechanisms. See appendix D for a brief discussion.

[8] Apart from the obvious limitation of perfect symmetry in firm sector, this specification is quite general. It allows for any positive number of different non-capital inputs. Production functions can change over time in any arbitrary way. Since we will only be interested in a single history of production functions at a time, specifically histories that form a balanced growth path, any time-varying parameters affecting production should be thought of as directly incorporated into period production functions. The only exception to this are parameters that are direct functions of aggregate factors, which are considered variables rather than time-specific constants for the purpose of specifying the aggregate production function. As we will not allow scale effects, only normalized versions of such factors, i.e input ratios, will actually be relevant. The one we specifically allow for, the capital-output ratio κ, is a sufficient statistic of any such normalized factors, as all other factor ratios remain constant along a balance growth path.

It would certainly be possible to generalize this model further, for example by allowing for stable distributions of firm characteristics. For our purpose of *motivating* the use of a particular firm sector framework, such extensions do not seem relevant.

2.1.1 Static CRTS – Criterion (C1)

Let $\mu_t(N\kappa, N\ell)$ be the number of firms as a function of aggregate factor endowment in a period. Using N as our scaling variable while holding (κ, ℓ) constant, we impose linear homogeneity on the production sector by setting the elasticity of output with respect to N equal to unity, $Y'(N)\frac{N}{Y} = 1$, where $Y(N) = \mu_t(N\kappa, N\ell) f_t(\frac{N}{\mu_t(N\kappa,N\ell)}\kappa, \frac{N}{\mu_t(N\kappa,N\ell)}\ell; \kappa)$. This condition simplifies to

$$(\sigma_t - 1)(\nu_t - 1) = 0, \tag{1}$$

where σ_t and ν_t are the local scale elasticities of f_t and μ_t, respectively.

Static aggregate constant returns to scale exist if, for any factor input combination, at least one of these elasticities is one. Disregarding odd scenarios where the economy-wide scaling mechanism switches based on factor inputs, this means that either individual production units operate a CRTS technology, or the number of firms expands proportionally with the size of the economy so that the scale and output of production units is independent of the size of the economy, or both.

2.1.2 Balanced Growth– Criterion (C2)

Criterion (C2) defines a balanced growth path as a sequence of aggregate factor input and output combinations such that the capital-output ratio as well as the input ratios of all non-capital factors remain constant. What does this imply for the sequence of production functions f_t of an individual firm? To facilitate intertemporal comparisons of these production functions, we start by locally approximating each of them by an isoelastic function at the actual factor input combination and capital externality (k_t, l_t, κ_t) in this period for the growth path under investigation.

$$f_t(k, n\ell; \kappa) \approx \tilde{f}_{\ell,t}(k, n; \kappa) = b_t k^{\alpha_t} n^{\beta_t} \kappa^{\gamma_t}$$

Here, α_t, β_t and γ_t are the local elasticities of output with respect to capital, the remaining factors l_t with constant input ratios, and the capital externality κ. b_t is a productivity constant ensuring equality between the true production function and its approximation at the relevant factor input combination, $f_t(k_t, n_t\ell; \kappa_t) = \tilde{f}_{\ell,t}(k_t, n_t; \kappa_t)$.

Aggregation yields

$$Y_t = \mu_t y_t = \mu_t \tilde{f}_{\ell,t}(\frac{K_t}{\mu_t}, \frac{L_t}{\mu_t}; \frac{K_t}{N_t}) = b_t \mu_t^{1-(\alpha_t+\beta_t)} K_t^{\alpha_t+\gamma_t} N_t^{\beta_t-\gamma_t}. \tag{2}$$

This already makes it clear we need $\alpha_t + \gamma_t > 0$ and $\beta_t - \gamma_t > 0$ to meet the requirement in definition 1 that output increases in both capital and a non-capital factor.

Using the characteristic of a stable capital-output ratio on a balanced growth path, $K_t = hY_t$ for a constant $h > 0$, as well as the definition of aggregate labour productivity $P_t = \frac{Y_t}{N_t}$, equation (2) can be written as

$$P_t^{1-(\alpha_t+\gamma_t)} = b_t h^{\alpha_t+\gamma_t} \mu_t^{1-(\alpha_t+\beta_t)} N_t^{\alpha_t+\beta_t-1}.$$

Log-differentiating this equation with respect to time[9] and using the notation $\hat{x}_t$ for the growth rate of a variable x_t, we arrive at

$$(1 - (\alpha_t + \gamma_t))\hat{P}_t = (1 - (\alpha_t + \beta_t))(\hat{\mu}_t - \hat{N}_t) + \hat{a}_t, \tag{3}$$

where $\hat{a}_t = \hat{b}_t + \dot{\alpha}_t \frac{K_t}{\mu_t} + \dot{\beta}_t \frac{N_t}{\mu_t} + \dot{\gamma}_t \frac{K_t}{N_t}$ is rate of TFP growth at the firm level, i.e. the growth rate of output for an unchanged factor input combination and externality (k_t, n_t, κ_t).

Equation (3) describes the relationship between empirical labour productivity growth $\hat{P}_t$, firm-level TFP growth $\hat{a}_t$, labour force growth $\hat{N}_t$, and the change in the number of active firms $\hat{\mu}_t$ that must always hold on a balanced growth path. As discussed in more detail in section 3, it is general enough to nest all the major growth frameworks as special cases for

[9] Here, we require time to be continuous and f_t, production factors, and μ_t to be differentiable with respect to time. While these implicit assumptions simplify the derivation of our result, they are obviously non-essential to the main point.

carefully chosen parameter combinations.

2.2 Proposed Base Framework

How can we satisfy the requirement of scale-invariance of aggregate production while allowing for a production sector that is not structurally inconsistent with investment-driven growth, as per definition 2?

Suppose the number of production units were not allowed to vary, $\hat{\mu} = 0$. Then, equation (1) implies that firm-level output must be linear homogeneous in factor inputs, $\sigma = \alpha + \beta = 1$, at any point in time. With this, equation (3) simplifies to $(1 - (\alpha_t + \gamma_t))\hat{P}_t = \hat{a}_t$, showing that any productivity growth is the result of firm-level TFP improvements.[10]

However, the notion of CRTS production with TFP being the only source of growth is hard to square with definition 2, which calls for a model structure that allows for growth resulting from private-sector expenditure. Firms operating under CRTS that pay factors their marginal product have no capacity to spend on productivity improvements. Moreover, the nature of the costs of a given productivity improvement as independent of production according to (C4) makes R&D expenditure inconsistent with scale invariance for CRTS producers. These and similar problems are the reason why the setting of CRTS production with TFP improvements is firmly associated with an exogenous growth trend, while models of endogenous growth have explored alternative production frameworks.

If we want a production sector model *does* satisfy definition 2, we almost certainly need to allow the number of firms to scale flexibly. If we choose a model where statically, μ scales proportionally with N, we satisfy equation (1) without needing any further restrictions regarding the firm-level production function. According to equation (3), a differential growth of firm numbers and overall scale, $\hat{\mu} - \hat{N}$, may then contribute to overall productivity

[10] It might appear that the special case $\alpha + \gamma = 1$ can remove the link between $\hat{P}$ and $\hat{a}$ by reducing equation (3) to $0 = \hat{a}$. This case, howerver, which corresponds to the Romer (1986) AK model, is incompatible with the requirement in definition 1 that aggregate output increase in both capital *and* another factor. $\alpha + \gamma = 1$ combined with static CRTS $\alpha + \beta = 1$ implies that the elasticity of output with respect to labour $\beta - \gamma$ is zero, see equation (2).

changes.[11]

There is a familiar and extremely simple model that meets the requirement of $\mu \propto N$: The standard model of competitive markets with free entry.

Individual firms have access to the same production technology that is characterized by a well-defined efficient scale: For any relevant combination of factor prices, there is a positive amount of output for which average costs are minimal. In equilibrium, all firms produce at this efficient scale and at zero economic profits.[12] Free entry and exit of firms ensures that any supply of factors will be absorbed, or equivalently, any overall demand will be met.

The economy as a whole always operates under static constant returns to scale – doubling the available amount of factors will result in twice the number of firms in equilibrium. The scale of individual firms is always well defined and independent of the size of the economy. It may, however, respond to changes in aggregate factor input ratios or, equivalently, factor prices. How individual firms respond to such changes will ultimately determine the properties of the aggregate production function.

The efficient scale of firms may also change as the economy gets more productive over time. Individual production units may get larger, for example to take better advantage of the benefits of the division of labour or in response to increasing R&D costs associated with the development of new products or production processes, or they may get smaller, for example if new technologies allow for a more local supply goods and services. Either way, entry and exit ensure that there is always the right number of firms to meet current demand. This feature of the free-entry model helps eliminate constraints on various production sector parameters and resulting knife-edge conditions, as suggested by equation (3) above and discussed in more detail in the next section.

Even though the technology used by production units is not CRTS, in equilibrium firms *locally* operate at a scale elasticity of unity. As firms produce at minimum average costs, i.e

[11] Note that equation (1) only requires μ to *statically* scale one-for-one with N; comparing different points in time may involve different mappings μ_t from scale to the number of firms.

[12] This is true for as long as the scale of each firm is much smaller than total output; for the purpose of modelling aggregate production, this seems like a reasonable assumption to make in general.

the average costs curve is flat and average costs equal marginal costs, average costs do not change for small deviations from the optimal scale. Equivalently, for the firm to have zero profits, the value of its factor payments must equal the value of its output. Since factors are paid their marginal product, we have $y = \sum_{i=1}^{n} x_i \frac{\partial y}{\partial x_i}$ for output y and production factors $x_1, ..., x_n$, which is true if and only if the local scale elasticity of y is one.[13]

This way of modelling aggregate production aligns well with experience. In most industries, there appears to be an efficient scale of production. This scale differs widely by industry and often changes over time, but production is clearly organized into units of comparable characteristics. Larger economies have more of these production units.

The suggestion to replace an ad-hoc CRTS aggregate production function in a general equilibrium model with a free-entry model that immediately aggregates into an equivalent CRTS aggregate production function may appear trivial and pointless. In the remainder of the paper, I will argue that following this approach does have value in certain situations, as it can lead to more consistent and robust models, make it easier to design problem-specific versions of a production sector, and make relevant mechanisms more obvious by specifying the production problem at the appropriate granularity and directly exposing the relevant economic unit in the firm sector of a macroeconomic model.

2.3 Discussion

This subsection discusses how and when it makes a difference to explicitly integrate the suggested micro structure into the production sector of a general equilibrium model.

2.3.1 Model Structure and the Missing Equation Problem

One big structural difference between our suggested approach and the standard assumption of an opaque CRTS production sector is that the free-entry model generally provides us with one additional meaningful condition that allows us to pin down one extra variable. In our setting, if firms make n (marginal) decisions, we get a system of $n + 1$ equations describ-

[13] This is a local application of Euler's homogeneous function theorem.

ing firm behaviour, including the free-entry or zero-profit condition. In a CRTS model, the same $n+1$ equations can be derived, but Euler's homogeneous function theorem implies that only n of them are independent. I will refer to this property of CRTS production functions as *the missing equation problem.*

The extra equation in the free-entry setting fundamentally determines the scale of a production unit or, equivalently, the number or mass of active units. While this might not always be a variable one would be particularly interested in, there are situations when it is useful.

First, there are numerous dimensions of firm behaviour that are inherently tied to the scale of the operation, many of which are relevant to macroeconomic outcomes. These are related to decisions where either the cost or the benefit is tied to the firm scale, market share, number of customers or similar. Examples include marketing or advertising expenditure, R&D efforts, the quality of internal training programs offered, the geographic reach of recruitment activities, local constraints regarding the availability of human capital or other resources, the ability to differentiate products from competitors, access to credit, lobbying activities, and regulatory compliance. As these aspects of firm activity are difficult to incorporate in models with a standard CRTS production sector, alternative approaches such as monopolistic competition have to be employed. Depending on the particular setup, these models may be inconsistent with the static and dynamic empirical regularities (C1) to (C3) discussed above.

Second, some variables of interest are closely linked to firm scale and can be calculated from it given some other available data, even though we may not think of them in this way. This includes excess capacity, the underutilization of capital, vacancies, and labour hoarding. Section 4 provides an example.

Finally, being able to differentiate between intensive and extensive margins and keeping track of entry and exit in the firm sector can be useful in the presence of frictions. With labour-market frictions, for example, a rise in firm exit might result in an uptick of unemployment. With capital-related frictions such as putty-clay investment or adjustment costs, changes in the optimal firm scale might affect asset values.

2.3.2 Modelling Firm Behaviour

For most applications, it should be much more straightforward to integrate model features at the level of the production unit than when working directly with a CRTS production function at a higher level of aggregation.

When defining an appropriate cost function for an individual firm, the only formal requirement is that average costs have a global minimum for a positive output level for any relevant price vector. Model extensions applied to a CRTS production function, on the other hand, would typically have to amount to a scale-invariant modification in order not to break the linear homogeneity of output. Ideally, the modeller would have to predict how an extension would manifest itself at the aggregate level if aggregation were done from first principles. This is likely to be difficult in general.

Starting at the level of production units with well-defined efficient scales provides more flexibility. It is possible, for example, to model aspects of the economic environment that are often inherently scale dependent, such as licensing fees, regulatory costs, or union coverage. The approach is also more robust. It is problematic to impose a fixed costs on firms operating under CRTS, but is entirely straightforward in the free-entry setting.

In practice, macroeconomic models that require more complex firm behaviour often rely on specifications based on an exogenously determined number of firms that do not necessarily produce under CRTS.[14] While equally convenient as our proposed model, this approach risks being inconsistent with important characteristics of real-world economies.

2.3.3 Understanding Productivity Shocks

Understanding the sources of productivity disturbances requires a disaggregated model of the firm sector, in which entry, exit and firm sizes have an important role to play.

Any aggregate production function consistent with definition 1 inherently obfuscates

[14] This allows for heterogeneity in the firm sector while keeping aggregation simple.

mechanisms leading to changes in productivities by aggregating them into a single productivity variable.[15] The causes of productivity variation are not explainable or understandable at this level of aggregation. Different types of disturbances within the firm sector will all manifest themselves as indistinguishable fluctuations in aggregate productivity.

Suppose, for example, the economy is subject to a change that does not immediately affect the factor endowment, but does have an impact on firms. Production units will reoptimize, and if they end up producing at a different scale, there will be entry or exit. After all adjustments are made, the economy will continue producing with the same factor inputs, but output may be different. We will assess that total factor productivity (TFP) has changed. Ultimately, any shock to the firm sector that does not involve inputs to the aggregate production function will be observed as a TFP shock at the aggregate level, if at all. A more disaggregated model of the firm sector, like the one proposed here, makes it possible to study a more differentiated range of shocks.[16]

2.3.4 Additional Complexity

Replacing a CRTS aggregate production function with a model of individual firms operating under free entry seems to add complexity. This does not, however, have to be an issue, as the following example shows.

Suppose each production unit requires a capital stock of exactly k, the value of the production facilities specifically designed to produce at minimal cost given current technological knowledge and expected factor prices over the lifetime of the capital goods. We know that

[15] See appendix 1.

[16] TFP shocks are, of course, the most popular way of modelling the source of cyclical fluctuations. If this is done in a frictionless setting such a real business cycle models based on Kydland and Prescott's seminal work (Kydland and Prescott (1982) and Prescott (1986)), where the role of the production sector is merely to accommodate the plans of the household given aggregate productivity, using this level of abstraction is perfectly sound. Once, however, there are frictions within the firm sector so that shocks can lead to observable changes in the economy beyond productivity, a more disaggregated approach may be required. This is because TFP may not be a sufficient statistic for the macroeconomic effects of a shock anymore, and because the aggregate production function is a black box with regards to the more differentiated responses of individual firms. The most obvious example is that of labour market frictions, where the unemployment response to a shock would likely depend mostly on the induced firm-level employment adjustment, entry and exit, not on its ultimate effect on productivity.

aggregate production is well-described by the CRTS production function $Y = F(K, N)$. If we define the firm-level production function as $y = f(k, n) := kF(1, \frac{n}{k})$, the overall output will aggregate to F under symmetry:

As the number of firms is given by $\mu = \frac{K}{k}$ and factor market clearing implies $\frac{n}{k} = \frac{N}{K}$, we have $Y = \mu y = \frac{K}{k} kF(1, \frac{N}{K}) = F(N, K)$. The factor prices obtained directly from the firm's profit function for the optimal choice of the labour input n and the zero-profit condition are consistent with those derived directly from the aggregate production function F based on marginal productivities.[17]

Integrating additional features into such a firm-level model is straightforward. They could, for example, affect the production function or enter the profit function as additional costs. In contrast to dealing with the aggregate production function or a representative firm directly, there is no need for these features to be specified in a scale-independent way.

For a range of applications, macroeconomists do model the production sector as composed of individual firms.[18] Often in this case, firms can earn economic profits or rents, and firm values can change in response to updates in rate-of-return expectations resulting from idiosyncratic shocks or varying aggregate conditions. This can complicate the firm-sector model substantially, because in general, firm behaviour can be forward-looking, as current decisions may affect future profits.[19] Replacing such a production sector model with a free-entry competitive market can greatly simplify the firms' optimization problem. As there are no profits in any state and firm values are always zero, the firms' profit maximization problem becomes entirely static. The labour market search model presented in section 4 is a good example of this. There, firms' decisions are so simple that it is possible to derive

[17] There is no firm-level marginal condition for capital, as the required capital stock is fixed. The return on capital is pinned down by the zero profit condition.

[18] This often involves settings where market power, product differentiation, or firm-level shocks or frictions are relevant and make a deeper and more complex model of firm behaviour desirable. Examples include New Keynesian monetary theory (see e.g. Woodford (2003)), international trade (e.g. Krugman (1979)), real aspects of business cycles (e.g. Chevalier and Scharfstein (1996)), aggregate demand (e.g. Blanchard and Kiyotaki (1987)) and financial frictions (e.g. Midrigan and Xu (2014)). Job search and endogenous growth will be discussed in more detail below.

[19] As long as the environment remains stable and non-stochastic, as in a steady-state or on a balanced growth path, firm-sector models can still remain tractable. Once aggregate shocks or adjustment processes are considered, however, solving them can become numerically very challenging.

a full analytical solution for the production sector dynamics that is valid universally. This contrasts with the otherwise structurally similar Mortensen-Pissarides model, for which, to the best of my knowledge, a general analytical solution is only available for the steady state.

3 Application: Endogenous Growth

In this section, we develop a model of endogenous growth that satisfies definition 1. This means it delivers stable, endogenously determined balanced growth without exhibiting static scale effects. Moreover, the specification is not subject to the knife-edge conditions that characterize popular models of endogenous growth. Specifically, we will not assume that aggregate output is linear in an accumulable factor.

I will start by explaining how various popular growth frameworks satisfy the balanced-growth condition (3) in different ways by choosing particular factor elasticities for the aggregate production function. I will discuss how our growth model can endogenize R&D-driven productivity growth without being subject to the same constraints on production function parameters thanks to the inherent properties of the free-entry, competitive market model we employ.

We will then develop the actual model and derive its aggregate static and growth characteristics. Finally, model features will be compared between our model and the major growth frameworks.

3.1 Growth Frameworks, Categorized

The relationships between the popular growth frameworks favoured in the literature can be understood by mapping them to equation (3), which makes clear their unique structural choices and particular growth characteristics. In their canonical forms, all of them model the production sector either directly through an isoelastic aggregate production function or a CRTS representative firm. This means that entry and exit have no role to play, and thus $\hat{\mu} = 0$.

The neoclassical growth model[20] imposes CRTS on the firm sector, typically setting $\alpha+\beta = 1$, $\gamma = 0$. With this, equation (3) simplifies to $\hat{P} = \frac{1}{\beta}\hat{a}$. This means that productivity growth is *always* ultimately determined by firm-level TFP growth $\hat{a}$, which is taken to evolve exogenously. As is well understood, the diminishing marginal product of capital implied by CRTS makes it impossible for the economy to grow indefinitely by means of investing in an accumulable factor. The economy will always reach a steady state.[21]

Raising the scale elasticity above unity while still maintaining diminishing returns to the accumulable factor, $\alpha + \beta > 1$, $\alpha < 1$, $\gamma = 0$ does not have a big impact on how labour productivity growth is determined if we maintain $\hat{\mu} = 0$. Equation (3) becomes $\hat{P} = \frac{1}{1-\alpha}\left((\alpha + \beta - 1)\hat{N} + \hat{a}\right)$. Labour supply growth does matter now, and it has a positive impact on productivity, as one would expect when economies of scale exist. Broadly speaking, the positive growth effects of both the TFP improvement $\hat{a}$ and size of the economy in terms of population increase with the scale elasticity $\alpha + \beta$. The resulting economy still has a time-variant steady state and transition dynamics just like the neoclassical growth model; the difference is that this steady state now scales with both N and a rather than a only.[22]

This range of parameters is associated with semi-endogenous growth models.[23] Even for $\hat{a} = 0$, long-term growth can be explained as a result of scale effects for positive population growth. One focus of this literature has been to rationalize sustained productivity growth in an environment where investment in new technologies do affect productivity ($\alpha+\beta > 1$), but the cost of implementing a given productivity improvement rises with the level of existing technological knowledge, i.e. there are diminishing returns to technology ($\alpha < 1$). Much of the motivation for this scenario comes from empirical findings on R&D costs.[24]

[20] This category encompasses any growth models based on a standard, CTRS firm sector, importantly the Solow-Swan (Solow (1956), Solow (1957), and Swan (1956)) and the Ramsey-Cass-Coopmans (Ramsey (1928), Cass (1965), and Koopmans (1963)) models.

[21] Our definition of balanced growth (C2) is not designed to allow for individual non-capital inputs to grow at different rates. Letting some rival inputs grow at exogenous rates different from both capital and labour will generally introduce a scale dependence of productivity growth. See for example Nordhaus (1992) for a discussion of the growth drag resulting from the reliance on finite resources.

[22] Structurally, this case is very similar to that of CRTS with a fixed factor, except that the sign of the contribution of labour growth is different.

[23] See for example Jones (1995a), Segerstrom (1998), and Jones (2005).

[24] Kortum (1993) discusses seemingly contradicting evidence on patent creation and R&D expenditure

Increasing α toward 1 results in infinities for non-zero $\hat{a}$ or $\hat{N}$. Yet, the special case of linearity of output in the accumulable factor is the realm of the the macroeconomic frameworks commonly referred to as endogenous growth models.

In the absence of entry or exit ($\hat{\mu} = 0$), (3) implies that productivity growth is generally linear in both $\hat{a}$ and $\hat{N}$, which is undesirable when attempting to explain growth endogenously, $\hat{a} = 0$. Endogenous growth models operate by removing productivity growth from this constraint, setting $\alpha + \gamma = 1$. This makes it possible in principle to determine $\hat{P}$ independently of labour-related variables. Doing so, however, requires the inclusion of an appropriate mechanism for this purpose in the model, which functions, in a sense, independently or on top of the core production sector, the operation of which is already described by what remains of equation (3). Different endogenous growth models follow different approaches, but all of the more commonly used ones are characterized by significant additional complexity attributable to the need to determine productivity growth outside of the core production sector.[25]

The AK model due to Romer (1986) is structurally very simple. Starting with a standard CRTS neoclassical model ($\alpha + \beta = 1$), long-term endogenous growth is achieved thanks to a delicately calibrated externality of the capital stock on productivity, $\gamma = 1 - \alpha$. What makes this model very tractable is that investment in capital is determined by firms operating under CRTS individually, so that aggregate productivity growth is really a by-product of the mechanisms familiar from the neoclassical growth model.

The two most popular growth frameworks, the "varieties" model of Romer (1990) and the models of creative destruction due to Grossman and Helpman (1991) and Aghion and Howitt (1992), do not rely on an externality and employ a much more sophisticated struc-

data. Segerstrom (1998) and Kortum (1997) develop theoretical models to explain these observations and conclude that it must be getting harder over time to innovate. Klette and Kortum (2004) model the relationship between firms' R&D expenditure, firm size and growth.

[25] If linearity of output in an accumulable factor were a true feature of our growth experience, one would expect it to emerge endogenously in growth models, based on a robust mechanism that operates under a range of specific model assumptions and for a variety of functional forms. In practice, however, endogenous growth models operate by assuming *exact* proportionality in a few important aspects of economic activity, with any slight deviations from these assumptions eliminating the possibility that the model delivers stable, balanced growth.

ture. Production of the final good is linear in a productivity parameter. Productivity can be increased at a cost, and the incentive for doing so comes from a mechanism that rewards whoever is responsible for a TFP improvement. Details differ between models, but most commonly the reward consists in a monopoly rent tied to ownership of the intellectual property that enables a productivity improvement. In order to have balanced growth, it is necessary that aggregate R&D costs are proportional to output for a given growth rate and that the costs and benefits associated with developing productivity improvements are such that these happen at a steady rate. The literature presents various ways of ensuring that theses requirements are met, but this generally requires imposing a very specific micro structure on the R&D sector, which makes such models both complex and dependent on numerous micro-level assumptions that macroeconomic models are otherwise often able to abstract from.

In addition to the knife-edge assumption of linearity of output in capital, early endogenous growth models were criticized for their tendency to have scale effects in population and their incompatibility with population growth. Under these parameters of $\alpha + \gamma = 1$, equation (3) requires that effective labour $a^{\frac{1}{\beta-\gamma}} N$ be constant for $\hat{\mu} = 0$.[26] This issue has been addressed by Young (1998), who set $\hat{\mu} = \hat{N}$[27], thus ensuring that the constraint (3) is always met. Others have used similar approaches.[28]

Growth models with capital coefficients beyond unity have not seen serious consideration. They would tend to suffer from explosive growth and have other problematic properties.

With the whole range of plausible production function parameters $1 - \beta \leq \alpha \leq 1$ already occupied by the growth literature, how can we build a model with new characteristics? The

[26] One response to the scale effects inherent in many frameworks of endogenous and semi-endogenous growth has been to reinterpret the labour force, the stock of technological knowledge and the growth rate as worldwide variables, or to argue that growth should be studied at a global scale (see, for example, Jones (2002), section V). In contrast, our approach in this paper is to *require* scale invariance and then investigate if and how we can still model endogenous growth. Appendix B offers some brief comments on the historical variability of the scale of the part of the wold contributing to the development of technological improvements.

[27] Note that while it is necessary to do this in order eliminate $\hat{N}$ from equation (3) to avoid scale issues in endogenous growth settings, criterion (C1) can be satisfied by ensuring that *statically*, μ_t scales with N_t, a much weaker condition that still allows these two variables to grow at different rates over time.

[28] Dinopoulos and Thompson (1998), Howitt (2000), and Peretto (1998)

answer lies in the endogenous change in the number of firms μ under competition with free entry.

We will assume that investment-driven growth opportunities exist in the long run but that at any given time horizon, the ability of firms to increase their productivity is limited, as accelerating the development of new production techniques is costly. This will mean that at any point in time, there is a positive but finite optimal R&D investment per firm and a finite optimal scale. Free entry guarantees, under these conditions, static constant returns to scale, while at the same time, it automatically accommodates any changes in optimal firm size and available aggregate labour supply as the economy grows.

3.2 Endogenous Growth without Scale Effects or Knife-edge Conditions

In following the recommendation regarding the free-entry structure of production sector, I aim to develop as simple an endogenous growth model as possible. Some additional notation will, however, result from the desire to present the relevant ideas and concepts in a way that is consistent with the existing literature on endogenous growth.[29]

Let the production function of an individual firm or production unit for a given level of technological knowledge be

$$\bar{f}(i,n) = \bar{a}(i)k(i)^{\bar{\alpha}}n^{\beta}. \tag{4}$$

Here, i is the stock of ideas available to the firm when devising its production process, and $\bar{a}(i) = \phi_1 i^{\chi_1}$ and $k(i) = \phi_2 i^{\chi_2}$ are, respectively, the TFP associated with the best production process the firm can use given its knowledge as well as the corresponding value of the capital required to produce. The assumption that all parameters $\phi_1, \phi_2, \chi_1, \chi_2$ are positive will ensure that productivity and capital inputs both increase with the stock of knowledge i.

Statically, the productivity of the firm is limited by the amount of technological knowledge

[29] The objective of this chapter is to present a *simple* example of a growth model that is built on the ideas presented in section 2. Many aspects of growth, R&D, knowledge spillovers and complex firm dynamics discussed in the literature are intentionally abstracted from.

i it has access to during the current period. This knowledge includes publicly available stock of ideas $\bar{i}$, which we assume to be the highest level of i used in production and thus revealed by any firm during the previous period.[30] In addition to that, a firm can add to this stock of ideas by devoting a share $\tau \in [0, 1]$ of the duration of the current period to R&D efforts rather than production. We will assume that

$$i = i(\tau) = (1 + \iota\tau^{\psi})\bar{i} \tag{5}$$

for constants $\iota > 0$ and $\psi \in (0, 1)$. The parameter range for ψ implies that small improvements can be made cheaply, but there are diminishing returns to R&D: increasing the rate at which technological advancements are made is disproportionally costly.

We will rewrite the production function 4 entirely in terms of capital and labour to make it more consistent with the notation used outside of the realm of endogenous growth models. We use the definitions of $k(i)$ and $\bar{a}(i)$ to write the firm's production function in the simple form

$$f(k, n) = ak^{\alpha}n^{\beta}, \tag{6}$$

for constants $a = \phi_1 \phi_2^{-\frac{\chi_1}{\chi_2}}$ and $\alpha = \bar{\alpha} + \frac{\chi_1}{\chi_2}$. Defining $\underline{k} = k(\bar{i})$ as the capital stock associated with the currently available public stock of ideas and $\hat{\underline{k}} = \frac{k}{\underline{k}} - 1$ as the rate of improvement over that level, we can use equation (5) and the definition of $k(i)$ to express τ as

$$\begin{aligned} \tau = \tau(\hat{\underline{k}}) &= \left(\frac{1}{\iota}\left(\frac{i}{\bar{\bar{i}}} - 1\right)\right)^{\frac{1}{\psi}} = \left(\frac{1}{\iota}\left(\left(\frac{k}{\underline{k}}\right)^{\chi_2} - 1\right)\right)^{\frac{1}{\psi}} \\ &= \left(\frac{1}{\iota}\left(\left(\hat{\underline{k}} + 1\right)^{\chi_2} - 1\right)\right)^{\frac{1}{\psi}} \approx \left(\frac{\chi_2}{\iota}\hat{\underline{k}}\right)^{\frac{1}{\psi}}, \end{aligned} \tag{7}$$

[30] This assumption identifies the period length with the time for which new ideas are exclusively available to the organization that developed it. It serves as a convenient way of simplifying the firm's problem and is a trick that is occasionally used in the growth literature. How long such a period would be is not clear. The quality-ladder literature tends to attribute significant importance to legal protection of intellectual property and might take the duration of patent protection as a guide, resulting in a timeframe of many years to decades. At the other extreme, one might consider it as the time it takes competitors to reverse-engineer a product and develop a similar design, which may be months.

thus rewriting it in terms of capital as well. Note that $\tau(\hat{\underline{k}})$ is strictly convex and unbounded above with $\tau(0) = \tau'(0) = 0$.

With this, we can represent a firm's output in the current period as

$$y = \underline{f}(k, n) = (1 - \tau(\hat{\underline{k}}))f(k, n) = (1 - \tau(\hat{\underline{k}}))ak^{\alpha}n^{\beta} \tag{8}$$

with $\hat{\underline{k}}$ a function of k.[31] Firms choose how much labour n to hire and what level of capital k to employ, which implies the share τ of resources that needs to be devoted to R&D in order to raise k from $\underline{k}$ to the required level.[32] In addition to that, profits must be zero in any equilibrium with free entry and exit.

Solving the model is surprisingly simple. The firm's profit in the current period is $\pi = y - (r + \delta)k - wn$ for factor prices r and w and a depreciation rate δ. The first-order conditions characterizing the optimal levels of factor inputs

$$\beta \frac{(1 - \tau)f(k, n)}{n} - w = 0$$
$$\alpha \frac{(1 - \tau)f(k, n)}{k} - \tau'(\hat{\underline{k}}) \frac{1}{\underline{k}} f(k, n) - (r + \delta) = 0$$

can be used to eliminate factor prices from the free-entry condition

$$(1 - \tau)f(k, n) - (r + \delta)k - wn = 0,$$

which, after some straightforward simplifications, yields[33]

$$\frac{(1 + \hat{\underline{k}})\tau'(\hat{\underline{k}})}{1 - \tau(\hat{\underline{k}})} = \alpha + \beta - 1. \tag{9}$$

[31] We are using an underscore to mark functions and variables tied to the current production technique.

[32] The capital stock serves as an indicator of the level of technology. Its role in the production function should be thought of encompassing the effects of physical capital, intellectual property, and the general level of productivity. This is why α is greater than $\bar{\alpha}$, which excludes TFP changes tied to the level of knowledge I.

[33] for $y = (1 - \tau)f(n, k) > 0$

Equation (9), which determines $\hat{\underline{k}}$ as a function of model parameters, is interesting. It is easy to verify that for positive output, i.e. $\tau < 1$, there is a unique growth rate of capital $\hat{\underline{k}}$ associated with any scale elasticity $\alpha + \beta \geq 1$, which is zero in the CRTS case $\alpha + \beta = 1$ and strictly increases in $\alpha + \beta$.[34]

As all firms have access to the same production technology, there is perfect symmetry in the firm sector, and the capital stock associated with the highest level of technology previously employed by *any* firm, $\underline{k}$ is identical to the capital stock previously employed by *every* active firm. We can thus identify the rate of capital growth relative to the broadly available technology, $\hat{\underline{k}}$, with the growth rate of firm-level capital $\hat{k}$.

In the case of CRTS, $\alpha + \beta = 1$, there is no benefit to raising the firm's capital stock, and therefore no resources are expended on changing the production technique. This is the scenario of the neoclassical growth model, where there is no scope for endogenous growth.

If there is a potential to increase productivity by investing in R&D, $\alpha + \beta > 1$, firms will devote resources to improve their production process, which has the effect of raising the capital stock, $\hat{k} > 0$.[35] For $\hat{k} \ll 1$, equation (9) can be approximated as $\tau'(\hat{k}) \approx \alpha + \beta - 1$, making it clear that firms are equating the marginal cost $\tau'(\hat{k})y$ of their R&D efforts to their marginal benefits, the improvement of profits $(\alpha + \beta - 1)y$.

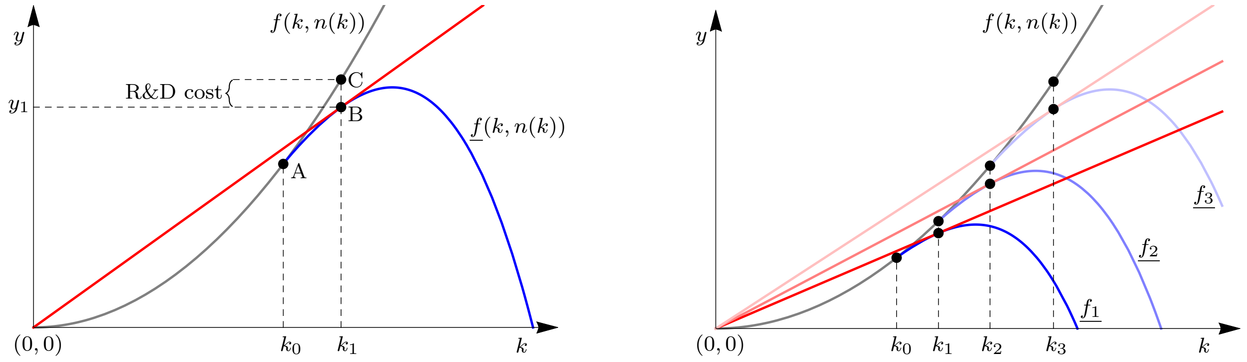
Figure 1 illustrates how the free-entry condition, which requires firms to produce at minimum average costs, leads to sustained growth given the ability of firm to achieve limited productivity improvements each period.

Having established that the capital stock at the firm level grows at a constant rate $\hat{k}$, all that remains to do is to find aggregate output. For aggregate factor endowments K and N, the number of firms in the economy is $\mu = \frac{K}{k}$ and the labour input of individual firms is

[34] Let $T(x) = \frac{(1+x)\tau'(x)}{1-\tau(x)}$ and $\bar{x} = \tau^{-1}(1)$. Given the definition of the function τ, we clearly have $T(0) = 0$. Moreover, it is straightforward to verify that T' is strictly positive on the open interval $(0, \bar{x})$ and that $\lim_{x \nearrow \bar{x}} T(x) = \infty$. The strict monotonicity combined with $T([0, \bar{x})) = \mathbb{R}^+$ implies the claim.

[35] By choosing to substitute k for the more abstract concept of a stock of ideas i in the production function, we made firm-level capital our productivity index, which will be convenient for aggregation later, as capital is one of the factors observed at the aggregate level. Alternatively, we could have written our model in terms of a stock of ideas i, total factor productivity b or any other index.

Figure 1: R&D Expenditure, Capital Intensity and Growth



(a) Capital adjustment and productivity improvement in a single period

(b) Growth over three periods

Notes: To facilitate a two-dimensional representation, labour $n = n(k)$ is adjusted in accordance with balanced-growth requirements. The grey line $f(k, n(k))$ shows the long-term production function, which does not include the costs of changing the production technique, the blue line corresponds to within-period production possibilities accounting for the option of changing the production technique to a more capital-intensive one through R&D. The slope of the red ray through the origin measures average productivity.

Panel (a) shows capital adjustments in a single period. Firms can use the production technique with capital k_0 at point A for free without performing R&D. Spending resources on R&D, capital intensities can be increased at an output cost corresponding to the difference between the grey and blue lines. Competitive firms choose the most cost-effective production technique available within this period, point B at $k = k_1$, at which average productivity is maximized. In the following period, the resulting production technique will be public knowledge and usable without additional R&D efforts, so firms will be able to produce at point C. The vertical distance between points B and C measures R&D expenditure within the period.

Panel (b) shows three consecutive periods of growth. The increasing slope of the red line shows the improvements in average productivity and, given zero profit, the rise in average factor costs and thus the ability to pay higher wages over time.

$n = \frac{N}{\mu} = \frac{Nk}{K}$. Aggregate output can then be calculated as

$$\begin{aligned} Y =& \mu y = \mu(1-\tau) a k^{\alpha} n^{\beta} = \frac{K}{k}(1-\tau) a k^{\alpha} \left(\frac{Nk}{K}\right)^{\beta} \\ =& (1-\tau) a k^{\alpha+\beta-1} K^{1-\beta} N^{\beta}. \end{aligned}$$

Finally, adding time indices to all variables and defining TFP in period t as $A_t = (1 - \tau(\hat{k})) a \left(k_0 (1+\hat{k})^t\right)^{\alpha+\beta-1}$ for a firm-level capital stock of k_0 in period $t = 0$, we can write output in period t the familiar form

$$Y_t = A_t K_t^{1-\beta} N_t^{\beta}, \tag{10}$$

where TFP grows at the constant rate $\hat{A}_t = (1+\hat{k})^{\alpha+\beta-1}$. Economy-wide R&D expenditure is $\frac{\tau}{1-\tau} Y$, and the rate of firm entry or exit is $\hat{\mu} = \hat{K} - \hat{k}$.

3.3 Comparison

We have developed a model that allows firms to invest in improved productivity and fully endogenizes the growth rate, without requiring technology to be linear in capital. Our model looks, at the aggregate level, exactly like the neoclassical growth model with exogenous technological progress, and as such allows for the medium-term adjustment dynamics familiar from the Solow-Swan and Ramsey-Cass-Coopmans models, while fully accounting for ongoing R&D investment. All this is at odds with the widely held belief that balanced, endogenously determined growth at a stable rate that is independent of labour supply growth is only possible if output is linear in accumulable asset.[36]

How is it possible for our model to exhibit long-term growth without being linear in capital? The answer is that, as is well understood, even the neoclassical model already allows for balanced growth as long as labour productivity improves. In the Solow model, this productivity improvement is assumed to happen automatically. Semi-endogenous growth models leverage the same mechanics by assuming aggregate scale effects, thus tying improvements in labour productivity to population growth. Typical endogenous growth models abandon this mechanism in favour of a structure where output is always automatically proportional to the capital stock. They then impose additional model structure on top of the core production sector model that regulates capital accumulation, which directly drives growth. Our model, in contrast, directly endogenizes the TFP or labour productivity parameter already found in the neoclassical growth model.[37] Doing so requires specifying production possibilities at the firm rather than aggregate level, as we have seen before.[38] This is in line with the message

[36] Interestingly, many of the main mechanics underpinning our model have previously been employed in the endogenous growth literature. The AK model is built on the idea that growth-relevant investment decisions are made at the firm level and affect the wider production sector; in our setting, the fact that technology eventually becomes public information allows for free entry and exit. The idea that entry is possible has been employed by Young (1998) and others to counteract scale effects.

[37] Alternatively, one could put the difference between the two approaches like this: Is economic growth fundamentally driven by the accumulation of capital, or is the ability to accumulate more capital an automatic by-product of productivity improvements that are achieved?

[38] Two fundamental questions when designing a model of R&D-driven growth are (1), what is the source of the returns on R&D expenditure? and (2), what mechanism regulates these expenditures and returns to achieve stable growth? Our model solves both of these issues by linking R&D activity to a well-defined and endogenously determined firm scale.

of section 2, which emphasized that *mechanisms* explaining productivity will not be found at the level of the aggregate economy, but at the lower level of individual production units; only their effects manifest themselves in the form of a single aggregate variable, TFP.

The fact that it has been so difficult to develop simple endogenous growth models without knife-edge conditions that are consistent with the static and dynamic regularities discussed above is likely the consequence of the missing-equation problem. When designing an endogenous growth model, one would want to start by building on an existing, successful foundation such as the Ramsey model. As we have seen, this is bound to fail due to the assumption of a CRTS representative firm. Ideally, firms would choose how much to optimally spend on production factors and productivity improvements, the benefits of which would depend on their scale of operation, which would in turn be pinned down by a zero-profit condition. Under CRTS, however, factor payments for static production *always* account for the firm's entire value-added revenue, leaving no room for any productivity-enhancing investments. Zero profits are implied, so that the free-entry condition does not determine the firm's size.[39, 40]

Table 1 compares our growth model to established frameworks. Where relevant, the "endogenous growth" column refers to the characteristics of the popular quality ladder and varieties frameworks rather than alternative specifications.

4 Application: Two-sided Search

This section applies our proposed modelling strategy to the realm of labour market search. For some time, the Mortensen-Pissarides (MP) model[41] was the standard framework for integrating search frictions into macroeconomic models. It has, however, seen challenges

[39] Semi-endogenous growth models do relax the assumption of CRTS, but do not impose a zero-profit condition as relevant mechanisms typically operate at the level of an aggregate production function.

[40] Endogenizing firm sizes in our model opens up the opportunity of a more natural integration of macroeconomic production with IO models, which has been an important focus of the recent literature on Schumpeterian growth (Acemoglu, Akcigit, et al. (2018), Akcigit and Kerr (2018), Klette and Kortum (2004), Lentz and Mortensen (2008), and Acemoglu and Cao (2015)). Our approach can add new aspects to this literature, in particular what happens to growth in an industry once individual firms reach a size that leads to an oligopolistic market structure.

[41] Mortensen and Pissarides (1994), Pissarides (2000), chapter 1.

Table 1: Growth Frameworks – Comparison

model feature	model			
	neoclassical growth	semi-endogenous	endogenous growth	*this
endogenous growth	no	yes	yes	yes
growth rate determined by	exogenous	employment growth	technology and market structure	technology
capital coefficient	$0 < \alpha < 1$	$0 < \alpha < 1$	$\alpha = 1$	$\alpha > 0$
knife-edge conditions	(scale elasticity)	none	capital coefficient	none
employment growth	irrelevant	determines productivity growth	problematic without extension	irrelevant
static CRTS	yes	no	no	yes
balanced growth	yes	yes	yes	yes
adjustment dynamics	possible	possible	none	possible
flexible/rising relative R&D costs	n/a	key feature	not possible	possible
market structure requirements	none	none	monopoly pricing on innovations	free entry and exit
consistent with competitive markets	yes	unclear	no	yes

in the past 15 years, starting with Shimer's influential paper[42]. Since then, there have been efforts to address the perceived shortcomings of the MP model, combined with a search for alternatives. Models of directed search have gained popularity, and at the current time, there does not seem to be a widely accepted standard for modelling labour market frictions.[43]

Compared to many alternatives, the MP model has desirable features. The use of an aggregate matching function as a black-box stand-in for much more complex micro mechanisms in individual labour markets offers a level of abstraction that often seems appropriate for general equilibrium models. In contrast, popular versions of directed-search models require specifying micro-level structure that one may want to abstract from and that could call into question the generality of the approach.

The recent challenges to the MP model originally stem from the difficulty to replicate empirical characteristics of labour market dynamics. There are, however, additional model properties that may merit further attention. Among those are the relative difficulty of integrating the MP model with standard production sector models due to its unusual atomistic structure as well as complexity of using it in non-stationary settings, which requires solving for the expected lifetime values of both employers and workers because of the forward-looking

[42] Shimer (2005)

[43] Rogerson, Shimer, and Wright (2005) is an older but comprehensive survey of the search literature.

nature of the assumed wage bargaining process.

In the following subsection, we will take a closer look at the structure of the MP model. I will argue that virtually all of the identified shortcomings are the result of the wage-bargaining structure and the mechanisms directly tied to it.

I will then develop a simple alternative matching-function-based labour market model. Calibrated versions of this model will be used to demonstrate its ability to match empirical patterns in stochastic settings.

Finally, there will be a comparative overview of our model and the MP model.

4.1 Deconstructing the Mortensen-Pissarides Model

I will be discussing the MP model in its canonical form, as presented in Pissarides (2000)[44]. While numerous extensions exist, including many advanced by Pissarides in his textbook, the basic version is the one that has seen the widest adoption in the literature, both for direct policy analysis and as a starting point for developing custom versions of the model.[45]

The core mechanic of the model operates as follows. Flows into employment are determined based on the current number of vacancies and unemployed workers by an aggregate matching function. Unemployment is history dependent and updated based on new job matches and exogenous job destruction. Vacancies are created by forward-looking firms under free entry such that the expected ex-ante value of a vacancy is zero. This ex-ante value is the combination of a cost of filling the vacancy, which depends on how long the position is expected to remain unfilled, and the share of an expected rent that can be earned for the duration of the resulting job. This rent is the difference between the productivity of the job and the worker's outside option, and is typically assumed to be split between the worker and the firm by Nash-bargaining.

To understand the construction of the model better, consider how a version *without* any

[44] Ibid., chapter 1.

[45] The following exposition does assume some familiarity with the framework.

labour market frictions and wage bargaining would determine supply and demand in the labour market. Any number of firms would be willing to hire workers at a wage w equal to their productivity p, so labour demand would be fully elastic. Workers would accept any job that pays at least their outside option b. Labour supply would be inelastic at the number of workers $\bar{n}$ for $w > b$. The frictionless labour market equilibrium thus exists at employment $n = \bar{n}$ and $w = p$ as expected, as long as $b < p$, which is always assumed.

If we now do allow for search frictions, there will be unemployment $u > 0$ so that $n+u = \bar{n}$. We need an extra equation to pin down the additional endogenous variable u. A matching function mapping unemployment to a resulting job-creation rate will generally suffice for this purpose. In a one-sided search model, a simple hazard rate for unemployment is enough to endogenize u. In a two-sided search setting, however, we introduce vacancies v along with the unemployment rate as a factor affecting the matching rate. We thus need one more equation to solve our model. This is where the MP model brings in the Nash-bargaining wage equation.

Even though it may not be obvious given the unusual structure of the MP model involving atomistic firms, the apparent need for the wage equation stems from a case of the missing equation problem, very similar to the endogenous growth scenario discussed in the previous section. To see why this is the case, think about how we would go about constructing a model of two-sided search where labour market characteristics result from firms choosing how many vacancies to open under free entry, when there is a cost to hiring. For a fixed productivity p of every worker, which is a CRTS assumption, a firm paying an employee their marginal product always makes zero profits from production. There is no scope for paying any additional hiring costs, and the free entry condition is redundant.[46] The MP solution to this problem is to replace marginal product wages with a lower wage rate that is determined differently. This has the two effects of (1) creating the capacity of firms to pay hiring costs out of rents and (2) effectively adds an equation by changing the redundant

[46] This is exactly the same problem as when integrating R&D investment into a model where firms operate at CRTS, as discussed in the previous section.

marginal product wage equation into an alternative specification that is independent of the zero-profit condition.

The generic wage equation imposed in the MP model brings with it an additional parameter, the relative bargaining power of the parties negotiating the wage. With this added degree of freedom, it is possible to set up the MP model to match any desired steady-state levels of vacancies and unemployment. Any model dynamics, however, will now also have to be in line with the constraints imposed by the wage equation. While there are still enough moving parts in the model generate a wide range of dynamics for certain subsets of endogenous variables, the difficulty of making the model match the data is part of what has been causing frustration with the framework, and indicates that the choice of wage equation, while plausible and convincingly argued, may not align well with real-world mechanisms.

Shimer (2005) pointed out that then-typical calibrations of the MP model resulted in unrealistically low volatilities of unemployment and vacancies when productivity shocks were used as a source of labour market fluctuations. Various solutions to this problem have been proposed.[47] I will focus on the contribution by Hagedorn and Manovskii (2008), as it is a different calibration strategy for the standard MP model that does not require new model extensions or additional mechanisms. Ljungqvist and Sargent (2017) and Ljungqvist and Sargent (2021) explain that alternative solutions to the Shimer challenge ultimately rely on the same mechanics used by Hagedorn and Manovskii (2008).

Putting it simply, the problem with traditional MP calibrations was that moderate productivity shocks would only lead to very small changes in firms' expected rents, thus only supporting minor changes in vacancy rates. The solution proposed by Hagedorn and Manovskii (2008) operates through choosing a calibration that amplifies the volatility of the firms' rent share substantially. In the Nash bargaining process, households are assumed to have an outside option that is almost as high as average productivity (in excess of 95%), but almost no bargaining power (about 5%). This means that the firm practically captures the part of

[47] A popular approach is to allow for wage stickiness, see Hall (2005). Ljungqvist and Sargent (2017) and Ljungqvist and Sargent (2021) survey approaches and papers that all exploit similar model mechanics to amplify the response of labour market variables to shocks.

output that varies with productivity and not much else, so that the resulting highly leveraged rents are extremely volatile and result in sufficient entry fluctuation to create realistic labour market fluctuations.

4.2 A Simple Alternative Model of Two-sided Search

This section develops an alternative firm-sector block for a two-sided search model. It endogenizes the wage rate and vacancies for given labour market flows and household labour supply, and can serve as a replacement for the free-entry/vacancy-cost/wage-equation block of the MP model, while leaving unemployment transitions, the matching function and labour supply unchanged.

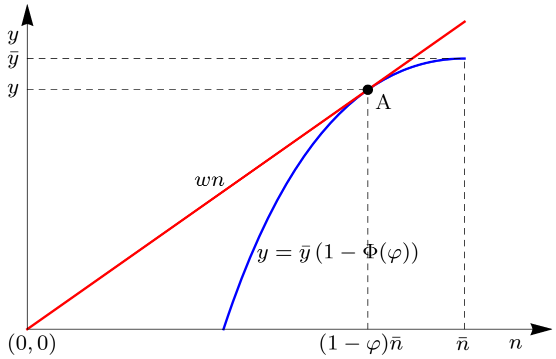
There are certainly many ways of incorporating vacancies into a firm-level model. Ultimately, most approaches likely boil down to specifying the costs of vacancies and possibly hiring relative to the benefits. In what follows, I choose a model setup that is both very simple and relatively general and generic.

Firms produce output using only labour. A fully staffed firm can produce an output of $\bar{y} = a\bar{n} > 0$ for the fixed scale of maximum employment $\bar{n}$. Due to labour turnover, a share $\varphi \in [0, 1]$ of firms' positions is vacant. Vacancies reduce firms' productivity, so that actual output is given by $y = \bar{y}\,(1 - \min\{\Phi(\varphi), 1\})$. We will assume that the cost or output reduction Φ associated with vacancies is continuously differentiable, convex, and that $\Phi(0) = \Phi'(0) = 0$. A small share of vacancies is associated with negligible output effects, but costs grow superlinearly with the extent of understaffing. $\Phi(1) > 1$ guarantees that the firm requires a certain positive amount of employment to produce anything at all, i.e. there is maximum possible vacancy rate $\bar{\varphi} \in (0, 1)$, beyond which firms cannot produce.

Let the productivity effect of vacancies be of the specific functional form $\Phi(\varphi) = -\left(\ln(1-\varphi) + \varphi\right) a^{-\omega} d$ for parameters $d > 0$ and $\omega \in \mathbb{R}$.[48] With this, the firm's profit

[48] This specification meets all requirements regarding the shape of Φ for *any* combination of parameters a, d and ω and results in a particularly simple closed-form solution of the model.

Figure 2: The Optimal Vacancy Rate



Notes: A fully staffed firm produces output $\bar{y}$ with labour $\bar{n}$. The blue line shows output as a function of employment within the production unit. Under free entry, firms produce at maximum average productivity at point A for a vacancy rate of φ. At A, the marginal and average products of labour are equal and account for all factor costs, which are represented by the red ray.

for $\varphi \geq \bar{\varphi}$ and a wage rate w is given by

$$\pi = \big(a + a^{1-\omega} d \left(\ln(1-\varphi) + \varphi\right) - w(1-\varphi)\big)\,\bar{n}. \tag{11}$$

The optimal plan specifies a vacancy rate φ for which the firm is willing to pay the prevailing wage rate w. The corresponding first-order condition combined the free-entry condition $\pi = 0$ determines the two endogenous variables[49]

$$\varphi = 1 - \exp(-\frac{a^{\omega}}{d}), \tag{12}$$

$$w = a^{1-\omega} d \frac{\varphi}{1-\varphi}. \tag{13}$$

Figure 2 illustrates equilibrium firm behaviour.

Labour market flows can be determined exactly as in the MP model. Separations between workers and firms happen at an exogenous rate λ, and new matches are given by a CRTS aggregate matching function M. The scale-independence of M makes it possible to specify the model in terms of rates rather than levels, specifically the unemployment rate u and the aggregate vacancy rate $v = \frac{\varphi+\lambda}{1-\varphi}(1-u)$. We will assume that the matching function is of the

[49] Notice that for small $\frac{a^{\omega}}{d}$ and thus vacancy rates, the following solution can be approximated as $\varphi = 1 - \exp(-\frac{a^{\omega}}{d}) \approx \frac{a^{\omega}}{d}$.

Cobb-Douglas form $M = M(v,u) = mv^{\eta}u^{1-\eta}$ and abstract from changes in the labour force, which would otherwise enter into the transition function for u. Substituting the expression for the aggregate vacancy rate, the matching rate can be written as $M = m\left(\frac{\varphi+\lambda}{1-\varphi}\right)^{\eta}(1-u)^{\eta}u^{1-\eta}$.

In a steady state, the flow into unemployment must equal the number of matches, $\lambda(1-u_{ss}) = M = m\left(\frac{\varphi+\lambda}{1-\varphi}\right)^{\eta}(1-u_{ss})^{\eta}u_{ss}^{1-\eta}$. The steady-state unemployment rate is thus $u_{ss} = \frac{\Psi}{1+\Psi}$ with $\Psi = \left(\frac{\lambda}{m}\left(\frac{1-\varphi}{\varphi+\lambda}\right)^{\eta}\right)^{\frac{1}{1-\eta}}$. It is easy to show that this steady-state is stable.

4.3 Model Dynamics and Calibration

We will now consider unemployment and vacancy dynamics. In contrast to the basic MP model, our model is, in principle, capable of endogenizing separations in a rather natural way. If there is market exit, production units shutting down could release some or all of their workers into unemployment rather than reassigning them to other projects. If conditions change in such a way that the optimal firm-level vacancy rate rises above the current one, workers could be released as well. While these options are interesting, we will not consider them here to maintain comparability with the MP model, and assume that labour is reallocated between production units.[50]

The two model parameters a and d directly affect firms' vacancy rate; shocks changing the values of these coefficients thus have implications for labour market outcomes. Innovations to a are TFP shocks that directly affect firms' labour productivity even in the absence of a behavioural response. For such disturbances, the elasticity ω determines the extent to which the vacancy rate responds to the productivity shock. Changes to d, on the other hand, only affect desired staffing without any direct effect on base productivity. Both types of shocks, however, can have significant *indirect* effects on average labour productivity. Again, in remaining consistent with common practice, I focus on TFP shocks.

[50] Allowing for this type of separation gives the model features similar to Pissarides's search model with endogenous job destruction (Pissarides (2000), chapter 2), where shocks can trigger an instantaneous release of a mass of labour into unemployment. This might be interesting for studying business cycle dynamics. Firstly, in the presence of ongoing separation, only large shocks would trigger mass separation, introducing a nonlinearity. Second, the mechanism only ever increases unemployment, leading to asymmetric cyclical effects.

Table 2: Calibration

scenario	parameter							
	m	λ	$\bar{a}$	d	ω	η	ρ	σ_a
Shimer	$\frac{4}{52}$	$\frac{0.2}{52}$	1	20	13.7	0.537	$\frac{0.775}{52}$	0.0068
Hagedorn et al.					14.7	0.549	$\frac{1.62}{52}$	0.0065

I calibrate the model to match the higher moments of vacancies v, unemployment u, market tightness $\theta = \frac{v}{u}$ and labour productivity $p = \frac{y}{1-\varphi}$ emphasized by Shimer (2005) and also targeted by Hagedorn and Manovskii (2008). TFP a varies such that we have $a_t = \bar{a}\exp(e_{a,t})$ for a constant $\bar{a} > 0$ and the AR(1) shock $e_{a,t+1} = (1-\rho)e_{a,t} + \varepsilon_{a,t}$, where normally distributed innovations $\varepsilon_{a,t}$ arrive at an average rate of 8 per year,

$$\varepsilon_{a,t} = \begin{cases} x \text{ with } x \sim \mathcal{N}(0, \sigma_a^2) & \text{with probability } 8\Delta t \\ 0 & \text{with probability } 1 - 8\Delta t \end{cases}$$

for a period length Δt measured in years.[51]

Some parameters are inconsequential for the purpose of matching higher moments and are fixed at plausible levels[52], but four of the parameters are used to target data moments.[53] I set the variance of the TFP innovation σ_a and the persistence parameter ρ to target the volatility and autocorrelation of productivity p. Then, the vacancy elasticity ω and the matching function parameter η are used to match the standard deviations of u and v to the empirical moments. I present two calibrations, matching both the empirical moments reported in Shimer (2005) and those targeted in the Hagedorn and Manovskii (2008) calibration. The model is simulated at weekly frequency, $\Delta t = \frac{1}{52}$. Table 2 summarizes the two sets of parameter values.

[51] The fixed stochastic arrival rate of shocks was chosen to make the characteristics and interpretation of the shock less dependent on the period length chosen for the simulation.

[52] The matching rate m of 4 p.a. results, for $v \approx u$, in an average unemployment duration of 3 months. The job destruction rate λ is chosen such that jobs last for five years on average. Productivity $\bar{a}$ is normalized to one so that $d = 20$ gives a vacancy rate of about 5%. The actual averages of u and v are $u \in (4.3\%, 4.2\%)$ and $v = 5.4\%$ in the Shimer and Hagedorn et al. calibrations.

[53] These are parameters for which there is no obvious single empirical value to set them to, and their interpretation is closely tied to the moments being matched.

Table 3: Data and Simulations

variable	Shimer		*this	Hagedorn et al.		*this
	data	model		data	model	
standard deviations						
u	0.190	0.009	0.190	0.125	0.145	0.125
v	0.202	0.027	0.202	0.139	0.169	0.139
θ	0.382	0.035	0.388	0.259	0.292	0.254
p	0.020	0.020	0.020	0.013	0.013	0.013
autocorrelations						
u	0.936	0.939	0.960	0.870	0.830	0.923
v	0.940	0.835	0.883	0.904	0.575	0.773
θ	0.941	0.878	0.937	0.896	0.751	0.877
p	0.878	0.878	0.878	0.765	0.765	0.765
correlations						
u, v	-0.894	-0.927	-0.879	-0.919	-0.724	-0.813
u, θ	-0.971	-0.958	-0.933	-0.977	-0.916	-0.932
v, θ	0.975	0.996	0.972	0.982	0.940	0.958
p, u	-0.408	-0.958	-0.883	-0.302	-0.892	-0.803
p, v	0.364	0.995	0.995	0.460	0.904	0.996
p, θ	0.369	0.999	0.955	0.393	0.967	0.943

Table 3 compares the results of the two calibration scenarios to the data and model output reported in Shimer (2005)[54] and Hagedorn and Manovskii (2008).[55]

The simulation results show than our model can match the empirical moments emphasized by Shimer (2005) and Hagedorn and Manovskii (2008) about as well as a carefully calibrated MP model. The point of this exercise is not to conduct a performance comparison;[56] it is rather to show how easy it is, using our proposed modelling approach, to construct an alternative two-sided matching model, which, despite it simplicity, is entirely viable in terms of ability to replicate empirical patterns that are considered important.

[54] I focus on the productivity shock scenario. The job-destruction shock simulations are not reproduced here.

[55] The model is simulated at weekly frequency, then statistics are aggregated to quarterly frequency by averaging. The reported moments are based on log deviations from an arithmetic mean. Each simulation run begins in the non-stochastic steady state and represents a total of 130,000 periods, i.e. 10,000 quarterly data points or 2,500 years.

[56] Even though Hagedorn and Manovskii (2008) are able to calibrate one more parameter than we do, for all practical purposes, their model has the same number of degrees of freedom as ours, because the discount rate, despite being an important factor in determining the stochastic properties of the MP model, cannot be chosen freely. Their calibration strategy is also different in that they target the elasticity of wages with respect to productivity, which then reduces the flexibility in matching labour market volatilities.

4.4 Comparison

Similar to the area of endogenous growth, labour market models with two-sided search have been a setting in which otherwise uncommon and relatively complicated modelling approaches have been used. The MP model in particular imposes wage-setting through a worker-firm negotiation process. The very specific assumptions regarding the mechanisms driving the vacancy-opening decisions in this model have been widely considered necessary for such a setting.

Our model demonstrates that, once we build a two-sided search model with an aggregate matching function on top of the framework advocated in this paper, the apparent necessity to resort to non-standard modelling techniques disappears. We specified a search model that has a similar ability to match relevant empirical patterns as the MP model while being simpler and having other desirable properties.

These include the ability to account for larger production units, to differentiate between average and marginal effects, to determine wages in a more standard way as a marginal product, and to greatly simplify the interaction between households and firms. Where for the MP model we need to calculate households' expected present value of utility from a job compared to an outside option, which may be exceedingly difficult in the presence of aggregate or even idiosyncratic fluctuations or more general utility functions, in our setting, hiring is a much more standard market interaction, where the behaviour of market participants and thus supply and demand can be determined independently of each other. Moreover, job creation behaviour is not inherently tied to a single mechanism. The MP model relates vacancy posting to the relationship between typically fixed vacancy costs and the ability to recoup these set-up investments through earning rents later, possibly over a long time. Our approach should provide more flexibility to include a variety of factors affecting firm behaviour.

Despite all this, our model remains very close to the MP approach for the determination of almost all relevant labour market variables. The history-dependent unemployment rate is driven by the same flows, which in turn are determined in exactly the same manner: exogenous job destruction and worker-firm matching according to an aggregate matching function. Vacancy rates are, in both cases, ultimately determined by free entry. The main difference is that we can rely on marginal-product wages thanks to nonlinearities in the firm-level technology that are not present in the MP model.[57]

As Ljungqvist and Sargent (2017) and Ljungqvist and Sargent (2021) explain, versions of the MP model that are in line with empirical volatility patterns all operate through leveraging a very narrow margin. In this context, two points are worth considering. First, we attribute *all* of the labour market fluctuations to firms' expectations regarding their ability to recover some moderate job setup costs by capturing a risky future rent amounting to just a few percent of their revenue over an expected job duration of years or decades. Ideally, one may want to be able to consider additional, robust mechanisms that complement this one. Second, calibrations or model versions that "work well" appear to inherently undermine the originally intended wage bargaining mechanics. Households are assumed to have an outside option that is almost as high as the productivity. They have minimal bargaining power, so they essentially capture this outside option as their wage. Quantitatively, Nash bargaining becomes irrelevant and the exogenously imposed outside option is almost indistinguishable from the market wage.[58]

In the light of these observations, the structure of our model can be interpreted as a generalization of the MP model as commonly calibrated today. It replaces the exogenous outside option as the primary determinant of the wage rate with the firm's marginal labour productivity, which is easier to justify and reason about.[59] Removing Nash bargaining sub-

[57] Hagedorn and Manovskii (2008) explicitly state that they consider the linearity and homogeneity of the MP model a challenge and that they interpret it as a linear approximation to a model with curvature. As we have seen, modelling this curvature explicitly removes many of the design constraints that make the MP model different and more complex compared to other popular frameworks used in macroeconomics.

[58] Hagedorn and Manovskii (2008) point out that their calibration is very close to what one would expect to find in a frictionless competitive environment (ibid., p. 1703). Under this interpretation, Nash bargaining only remains in the model as a technically necessary formality. The outcome is reminiscent of the Diamond (1971) paradox.

[59] For the purpose of calibration the outside option parameter, Hagedorn and Manovskii (2008) explicitly appeal to the idea that the MP model should be interpreted as a linear approximation to a richer, nonlinear model.

stantially simplifies the model without changing outcomes qualitatively, while transparently acknowledging the existence of a general market wage. The inherent nonlinearities in our model make it possible to include various mechanisms that affect vacancy rates in a flexible way, including the vacancy costs emphasized by the MP model.

5 Conclusions

This paper argues that the structure of the firm sector included in macroeconomic models can have important implications both for available model-design choices and aggregate characteristics of the model economy. At this time, a variety of different production sector frameworks are in widespread use, many of which have significant limitations.

The concept of the aggregate production function is somewhat vague and fuzzy. While the CRTS-representative firm setting is straightforward and logically consistent, it is a unique corner case that is fundamentally incompatible with endogenous growth and suffers from the missing-equation problem. Of course, the notion that under CRTS, firm size is irrelevant and profits are zero is widely appreciated; still, the examples presented in sections 3 and 4 do suggest that this model characteristic can significantly impact model design options even in situations when the connection is not obvious. Alternative specifications of the aggregate production function with increasing returns to scale appear to break the link to a representative-firm metaphor, making it more difficult to derive production choices. Models that employ such an aggregate production technology generally include additional structure determining the factors that enter into aggregate production.[60]

Even if the firm sector is modelled in a more disaggregated way, not all commonly made assumptions are in line with the empirical long-term behaviour of economies. In section 2, we argued that a mechanism to adjust the number of firms over time is likely a necessary feature for the production sector to be compatible with the notion of endogenous growth. For any setting involving production dynamics, it is then desirable to allow for an appropriate

[60] Examples of this related to economic growth were given in section 3.

form of firm entry and exit.

The aggregate production function is not merely an abstract concept, it should be considered an actual aggregate of an underlying firm sector. This means that, on the one hand, any aggregate production function needs to be consistent with a plausible microstructure of a firm sector. On the other hand, any suitable firm sector microstructure needs to aggregate into a production sector object that is consistent with the empirical regularities we observe at higher levels of aggregation. It might be desirable to view determining aggregate production as the end point of the process of specifying a production sector, not the starting point. Ad-hoc assumptions about the outcome of aggregation involve the risk of being incompatible with a plausible microstructure of the production sector.

This paper proposes a more axiomatic approach to production sector design. We put forward a set of simple and empirically well-supported requirements that a model should meet and discuss how they can be met in principle. The approach is constructive in that it includes a specific recommendation for a firm sector model that is simple, well-understood and part of our standard microeconomic toolbox. Using this approach allows us to derive additional model features such as CRTS endogenously rather than assuming them. Creating "deeper" models like this alleviates limitations to our ability to uncover economic structure. Directly assuming a model characteristic like CRTS not only makes it impossible to find the mechanism that leads to this outcome, it also deprives us of the opportunity to understand any phenomena tied to that hidden mechanism. The recommendation to use more feature-rich production sector models can be seen as an application of the Lucas critique[61] to a different domain: If we want to ensure that our model predictions remain plausible as we take them beyond immediate scenario they were calibrated to, we need to be able to rely on a credible, deep microstructure to deliver the correct outcomes.[62]

The example applications of sections 3 and 4 show that following our proposed modelling approach has advantages beyond ensuring consistency with aggregate phenomena. Maybe

[61] Lucas (1976)

[62] Better matching data moments by adding more complexity to an inherently inappropriate model is like adding epicycles; it will work well, but little can be learned from it about the true nature of the world.

surprisingly, the resulting models were simpler than existing alternatives.[63] Part of the reason is that the firm sector is guaranteed to aggregate into a regular CRTS complex. Another factor is that it is often easier to express relevant model features at the firm level – in the form of a cost function or production technology – rather than integrating them into an aggregate production function directly or building additional model structure on top or around an aggregate production function. Finally, the resulting domain-specific models are likely to be more consistent and compatible with each other. Combining the models presented in the previous two sections into an endogenous growth model with unemployment should be entirely straightforward, much easier than integrating an MP model into one of the popular endogenous growth frameworks. I anticipate that following this approach will enable us to find simpler and more intuitive alternatives to existing macro frameworks, and there will likely be fewer areas deemed to require unusual and more complex modelling approaches.

As the applications show, our approach challenges the belief that rents or mark-ups are necessary to deliver basic economic outcomes such as job creation or R&D investment. In our models, all firm behaviour is regulated in a standard way through competitive markets. In the literature, deviations from the benchmark setting of perfect competition are often tied to situations where additional costs cannot be absorbed by firms due to an assumption of linearities in production.

Building a firm sector model around a production unit with a well-defined optimal scale enables us to look into relevant dimensions of firm behaviour that are otherwise often inaccessible. This includes differentiating between marginal and average effects, between intensive and extensive margins and between firm-level and aggregate adjustments, none of which are possible in pure CRTS settings due to the missing equation phenomenon.

Our preferred firm sector model of perfect competition and free entry is a simple special case, but it is an attractive benchmark setting and starting point. A number of relevant

[63] In both examples, our models were both more general and regular than typical existing frameworks. The ability to specify model features in a direct and unconstrained way enabled us to follow standard approaches do determine firm decisions and factor incomes while also freeing us from the need to make very specific structural assumptions.

extensions could be integrated into the model in a fairly straightforward way, although they would tend to complicate the model, possible considerably. This includes barriers to entry and exit, and related to that market power, rents, changes in firm values and (temporary) profits and losses. Even though our simple firm-sector model may not be the ideal framework for every setting, I would argue that our examples have shown that the way production is represented in macroeconomic model frameworks deserves a closer look. The models developed in the previous sections, simple as they may be, at the very least demonstrate the potential of our approach by already delivering model features that have previously widely been considered impossible.[64]

References

Acemoglu, Daron, Ufuk Akcigit, et al. (Nov. 2018). "Innovation, Reallocation, and Growth". In: *American Economic Review* 108.11, pp. 3450–91. DOI: `10.1257/aer.20130470`.

Acemoglu, Daron and Dan Cao (2015). "Innovation by entrants and incumbents". In: *Journal of Economic Theory* 157.C, pp. 255–294. DOI: `10.1016/j.jet.2015.01.001`.

Aghion, Philippe and Peter Howitt (1992). "A Model of Growth Through Creative Destruction". In: *Econometrica* 60.2, pp. 323–351. DOI: `10.2307/2951599`.

Ahmad, Mumtaz, John G. Fernald, and Hashmat Khan (Sept. 2019). *Returns to Scale in U.S. Production, Redux.* Carleton Economic Papers 19-07. Carleton University, Department of Economics.

Akcigit, Ufuk and William R. Kerr (2018). "Growth through Heterogeneous Innovations". In: *Journal of Political Economy* 126.4, pp. 1374–1443. DOI: `10.1086/697901`.

[64] In Pollak (2025), I show that given the same environment of a competitive, free-entry production sector advocated in this paper, and under a very general and plausible assumption regarding the completeness of investment markets, the rate of labour productivity growth on a balanced growth path is a function only of the capital share and the interest rate, with the predicted growth rate matching the empirical value of about 2 percent per annum. The relevant mechanism ensuring this result is another example of an interesting effect that is masked and thus undiscoverable under common assumptions regarding aggregate production due to the missing equation problem (in the case of CRTS production units) or the absence of a structurally meaningful zero-profit condition (in the case of non-competitive settings).

Basu, Susanto and John G. Fernald (1997). "Returns to Scale in U.S. Production: Estimates and Implications". In: *Journal of Political Economy* 105.2, pp. 249–283. DOI: 10.1086/262073.

Blanchard, Olivier Jean and Nobuhiro Kiyotaki (1987). "Monopolistic Competition and the Effects of Aggregate Demand". In: *The American Economic Review* 77.4, pp. 647–666.

Bodkin, Ronald G. and Lawrence R. Klein (1967). "Nonlinear Estimation of Aggregate Production Functions". In: *The Review of Economics and Statistics* 49.1, pp. 28–44.

Cass, David (1965). "Optimum Growth in an Aggregative Model of Capital Accumulation". In: *The Review of Economic Studies* 32.3, pp. 233–240. DOI: 10.2307/2295827.

Chevalier, Judith A. and David S. Scharfstein (1996). "Capital-Market Imperfections and Countercyclical Markups: Theory and Evidence". In: *The American Economic Review* 86.4, pp. 703–725.

Diamond, Peter A (1971). "A model of price adjustment". In: *Journal of Economic Theory* 3.2, pp. 156–168. DOI: https://doi.org/10.1016/0022-0531(71)90013-5.

Dinopoulos, Elias and Peter Thompson (1998). "Schumpeterian Growth without Scale Effects". In: *Journal of Economic Growth* 3.4, pp. 313–335. DOI: 10.2307/40215991.

Grossman, Gene M. and Elhanan Helpman (1991). "Quality Ladders in the Theory of Growth". In: *The Review of Economic Studies* 58.1, pp. 43–61. DOI: 10.2307/2298044.

Hagedorn, Marcus and Iourii Manovskii (Sept. 2008). "The Cyclical Behavior of Equilibrium Unemployment and Vacancies Revisited". In: *American Economic Review* 98.4, pp. 1692–1706. DOI: 10.1257/aer.98.4.1692.

Hall, Robert E. (Mar. 2005). "Employment Fluctuations with Equilibrium Wage Stickiness". In: *American Economic Review* 95.1, pp. 50–65. DOI: 10.1257/0002828053828482.

Howitt, Peter (Sept. 2000). "Endogenous Growth and Cross-Country Income Differences". In: *American Economic Review* 90.4, pp. 829–846. DOI: 10.1257/aer.90.4.829.

Jones, Charles I. (1995a). "R & D-Based Models of Economic Growth". In: *Journal of Political Economy* 103.4, pp. 759–784. DOI: 10.2307/2138581.

Jones, Charles I. (1995b). “Time Series Tests of Endogenous Growth Models”. In: *The Quarterly Journal of Economics* 110.2, pp. 495–525. DOI: 10.2307/2118448.

— (Mar. 2002). “Sources of U.S. Economic Growth in a World of Ideas”. In: *American Economic Review* 92.1, pp. 220–239. DOI: 10.1257/000282802760015685.

— (2005). “Chapter 16 - Growth and Ideas”. In: ed. by Philippe Aghion and Steven N. Durlauf. Vol. 1. Handbook of Economic Growth. Elsevier, pp. 1063–1111. DOI: 10.1016/S1574-0684(05)01016-6.

Kaldor, Nicholas (1961). “Capital Accumulation and Economic Growth”. In: *The Theory of Capital: Proceedings of a Conference held by the International Economic Association.* Ed. by D. C. Hague. London: Palgrave Macmillan UK, pp. 177–222. DOI: 10.1007/978-1-349-08452-4_10.

Klette, Tor Jakob and Samuel S. Kortum (Oct. 2004). “Innovating Firms and Aggregate Innovation”. In: *Journal of Political Economy* 112.5, pp. 986–1018. DOI: 10.1086/422563.

Koopmans, Tjalling C. (1963). *On the Concept of Optimal Economic Growth.* Cowles Foundation Discussion Papers 163. Cowles Foundation for Research in Economics, Yale University.

Kortum, Samuel S. (1993). “Equilibrium R&D and the Patent–R&D Ratio: U.S. Evidence”. In: *The American Economic Review* 83.2, pp. 450–457. DOI: 10.2307/2117707.

— (1997). “Research, Patenting, and Technological Change”. In: *Econometrica* 65.6, pp. 1389–1419. DOI: 10.2307/2171741.

Krugman, Paul R. (1979). “Increasing returns, monopolistic competition, and international trade”. In: *Journal of International Economics* 9.4, pp. 469–479. DOI: https://doi.org/10.1016/0022-1996(79)90017-5.

Kydland, Finn E. and Edward C. Prescott (1982). “Time to Build and Aggregate Fluctuations”. In: *Econometrica* 50.6, pp. 1345–1370. DOI: 10.2307/1913386.

Lentz, Rasmus and Dale T. Mortensen (2008). “An Empirical Model of Growth through Product Innovation”. In: *Econometrica* 76.6, pp. 1317–1373. DOI: 10.2307/40056508.

Ljungqvist, Lars and Thomas J. Sargent (Sept. 2017). “The Fundamental Surplus”. In: *American Economic Review* 107.9, pp. 2630–65. DOI: 10.1257/aer.20150233.

— (2021). “The fundamental surplus strikes again”. In: *Review of Economic Dynamics* 41. Special Issue in Memory of Alejandro Justiniano, pp. 38–51. DOI: https://doi.org/10.1016/j.red.2021.04.007.

Lucas, Robert E. (1976). “Econometric policy evaluation: A critique”. In: *Carnegie-Rochester Conference Series on Public Policy* 1, pp. 19–46. DOI: https://doi.org/10.1016/S0167-2231(76)80003-6.

Midrigan, Virgiliu and Daniel Yi Xu (Feb. 2014). “Finance and Misallocation: Evidence from Plant-Level Data”. In: *American Economic Review* 104.2, pp. 422–58. DOI: 10.1257/aer.104.2.422.

Mortensen, Dale T. and Christopher A. Pissarides (1994). “Job Creation and Job Destruction in the Theory of Unemployment”. In: *The Review of Economic Studies* 61.3, pp. 397–415.

Nordhaus, William D. (1992). “Lethal Model 2: The Limits to Growth Revisited”. In: *Brookings Papers on Economic Activity* 23.2, pp. 1–60. DOI: https://doi.org/10.2307/2534581.

Peretto, Pietro F. (1998). “Technological Change and Population Growth”. In: *Journal of Economic Growth* 3.4, pp. 283–311.

Pissarides, Christopher A. (2000). *Equilibrium Unemployment Theory.* Cambridge, MA: MIT Press.

Pollak, Andreas (2025). “Investment and Economic Growth”. manuscript.

Prescott, Edward C. (1986). “Theory ahead of business-cycle measurement”. In: *Carnegie-Rochester Conference Series on Public Policy* 25, pp. 11–44. DOI: 10.1016/0167-2231(86)90035-7.

Ramsey, F. P. (1928). “A Mathematical Theory of Saving”. In: *The Economic Journal* 38.152, pp. 543–559. DOI: 10.2307/2224098.

Rogerson, Richard, Robert Shimer, and Randall Wright (Dec. 2005). “Search-Theoretic Models of the Labor Market: A Survey”. In: *Journal of Economic Literature* 43.4, pp. 959–988. DOI: 10.1257/002205105775362014.

Romer, Paul M. (1986). “Increasing Returns and Long-Run Growth”. In: *Journal of Political Economy* 94.5, pp. 1002–1037. DOI: 10.2307/1833190.

— (1990). “Endogenous Technological Change”. In: *Journal of Political Economy* 98.5, S71–S102. DOI: 10.2307/2937632.

Segerstrom, Paul S. (1998). “Endogenous Growth without Scale Effects”. In: *The American Economic Review* 88.5, pp. 1290–1310. DOI: 10.2307/116872.

Shimer, Robert (Mar. 2005). “The Cyclical Behavior of Equilibrium Unemployment and Vacancies”. In: *American Economic Review* 95.1, pp. 25–49. DOI: 10.1257/0002828053828572.

Solow, Robert M. (1956). “A Contribution to the Theory of Economic Growth”. In: *The Quarterly Journal of Economics* 70.1, pp. 65–94. DOI: 10.2307/1884513.

— (1957). “Technical Change and the Aggregate Production Function”. In: *The Review of Economics and Statistics* 39.3, pp. 312–320. DOI: 10.2307/1926047.

Swan, T. W. (1956). “Economic Growth and Capital Accumulation”. In: *Economic Record* 32.2, pp. 334–361. DOI: 10.1111/j.1475-4932.1956.tb00434.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1475-4932.1956.tb00434.x.

Woodford, Michael (2003). *Interest And Prices: Foundations Of A Theory Of Monetary Policy.* Princeton, NJ: Princeton University Press.

Young, Alwyn (1998). “Growth without Scale Effects”. In: *Journal of Political Economy* 106.1, pp. 41–63. DOI: 10.1086/250002.

Appendix

A Scale-Invariant Production Functions

Let $K > 0$ be aggregate capital and L a tuple of aggregate inputs of other factors $|L| > 0$. K encompasses the total value of all assets used in production measured in units of output, including physical capital, intellectual property and anything else that contributes to the overall value of the firm sector. L will be interpreted as primarily comprising of labour, and could, for example be a combination of labour inputs by quality, skills, training, or similar, along with other non-capital factors; in the case where homogeneous labor is the only input besides capital, L would effectively be a scalar.

Define $N = |L|$ as a scalar measure of the quantity of labour input and, based on this, the capital-labour ratio $\kappa = \frac{K}{N}$ and the remaining factor input ratios, $\ell = \frac{L}{N}$.[65]

We will use an index t to refer to discrete time periods starting at t_0. The aggregate production function in period t is F_t, so that aggregate output is given by $Y_t = F_t(K_t, L_t, \upsilon_t)$, where υ_t is a parameter vector. The production function can differ between periods. Let Σ_t be the value of state object in period t. The transition function T updates the state and the production function, possibly depending on current inputs to aggregate production, $(\Sigma_{t+1}, F_{t+1}) = T(\Sigma_t, K_t, L_t, \upsilon_t)$. This specification is extremely general, and allows for almost any sequence of production functions.[66]

Definition 1A (Scale-Invariant Production Function) *A sequence of aggregate production functions F_t, $(t - t_0) \in \mathbb{N}_0$, with F_t strictly increasing in the inputs capital and labour for all t that is generated by a transition function T will be said to be scale invariant if it meets the following three criteria:*

(C1) For all t, $F_t(K, L, \upsilon)$ is linear homogeneous in the inputs (K, L), i.e. $\forall t, K, L, \upsilon : q >$

[65] If there is only one other production factor, $\dim L = 1$, then $\ell = (1)$.

[66] Notice that the initial state Σ could include a full description of the world economy, and that υ_t could be the vector of all stochastic micro-level shocks in period t, meaning there are no real limitations to the complexity of the patterns production function sequence inherent in this specification.

$0 \Rightarrow F_t(qK, qL, \upsilon) = qF_t(K, L, \upsilon)$.

(C2) It is consistent with balanced growth, in the sense that any capital-output ratio that is possible in one period can be maintained indefinitely while keeping all non-capital inputs and parameters stable, i.e. $\forall s, t, K, L, \upsilon_s, \upsilon_t : \exists q > 0 : qF_s(K, L, \upsilon_s) = F_t(qK, L, \upsilon_t)$.

(C3) Changes in the production function must be independent of the scale of production, i.e. for any initial state Σ_t and any sequence of production function arguments $(K_{t+s}, L_{t+s}, \upsilon_{t+s})_{s\in\mathbb{N}_0}$, the transition function T must produce the same sequence of production functions $(F_{t+s})_{s\in\mathbb{N}}$ as for the normalized parameter sequence $(\kappa_{t+s}, \ell_{t+s}, \upsilon_{t+s})_{s\in\mathbb{N}_0}$.

The following result summarizes how definition 1A constrains the shape and evolution of the aggregate production function.

Lemma 3 (Scale-Invariant Production Function Characteristics)

1. *Any sequence of aggregate production functions consistent with definition 1A can be written in the form $F_t(K, L, \upsilon) = F(\kappa, p_t(\kappa, \ell)\ell, \upsilon)N$, where F is linear homogenous in its first two arguments and p_t is a time-varying productivity parameter that depends on the factor input ratios (κ, ℓ) and is updated according to a transition rule $(\Sigma_{t+1}, p_{t+1}) = T_p(\Sigma_t, \kappa_t, \ell_t, \upsilon_t)$ for a state Σ_t.*

2. *If F_t is differentiable for all t and we are willing to make the following assumption, then p_t is a constant for each period t, i.e. $\forall \kappa, \ell : p_t(\kappa, \ell) = \bar{p}_t > 0$: Along any balanced growth path, i.e. a sequence of production levels achieved in different periods with the same non-capital inputs and production function parameters and capital input levels proportional to period output, the elasticites of substitution between capital and* all *non-capital inputs remain constant.*

Proof. Part 1 of the lemma is s straightforward rewrite of the production function process that specifically uses the CRTS requirement (C1) twice. Part 2 is readily shown by deriving the corresponding marginal products from the production function expression $F(K_t, p_t(\frac{K_t}{N_t}, \frac{L_t}{N_t})L_t, \upsilon)$ at $K_t = p_t(\kappa, \ell)K,\ L_t = \ell N_t$ for fixed κ and ℓ and noting that any non-zero derivatives of p_t violate the requirements of unchanging elasticities of substitution. ■

Part 1 of Lemma 3 states that, as long as an economy moves along a balanced growth path, even one that involves a changing scale of non-accumulable factor inputs, any variation in output per labour are the result of the same productivity changes across all non-accumulated production factors.

Under the additional assumption of part 2 of the lemma, definition 1A imposes a much stronger constraint on the changes of the production function, now requiring a single relative productivity change to apply to the non-accumulable inputs *uniformly across all input ratios.* The simple intuition behind this result is that any non-uniform productivity change would deform isoquants in such a way that elasticities of substitution would be affected. Therefore, the requirement of locally stable substitutability implies stability of the global shape of the CRTS production function over time.

In the particularly relevant case of capital and labour being the only production factors of interest, this means that all growth is driven by improvements in labour productivity, and would thus be classified as Harrod-neutral. The definition does not impose any restrictions on how this factor productivity should change over time, except that any variations must be independent of the scale of output.

B Relevance

I will comment on the relevance of the three criteria comprising definition 1 by asking three questions. First, do they make sense from a methodological or theoretical point of view? Second, given the common use cases for aggregate production functions, do they align with the relevant data? Third, are they in line with common modelling practice?

B.1 Methodology

Criterion (C1) formalizes the expectation that, for a geographically homogeneous economy, each subregion should be describable as a scaled-down version of the whole. Of course, this only makes sense in reality as long as the scale is large enough for the homogeneity to be maintained.[67] Combined with (C3), this ensures that the principle of scale invariance applies across time as well, i.e. if relative scales of subregions change either due to differential changes in factor endowments or changes in the definitions of the subregions themselves.

(C2) addresses another dimension of scale invariance, focusing on how the economy changes with productivity. From a purely theoretical perspective, compatibility with balanced growth is an appealing feature of a production sector, as it supports stable ratios between income, consumption and wealth, which would be expected to a certain extent in a setting with finitely lived households that wish to accumulate and decumulate lifecycle savings.

B.2 Evidence

Static CRTS in production are a widely appreciated and uncontroversial feature of the data. Estimates of scale elasticities for national economies typically find values close to unity.[68] Of course, any empirical test of scale independence will require specific assumptions to disentangle scale effects from time-varying TFP.[69]

Similarly, the notion that the capital stock expands proportionally with income is well supported empirically and has been noted since at least Kaldor (1961). It is directly implied by the combination of a stable interest rate with stable factor shares, which, despite some

[67] E.g. world, free-trade area, country, region, metropolitan area, but not city district or household.

[68] An example of an earlier study is Bodkin and Klein (1967), who estimate a scale elasticity of 1.2, and provide a brief overview of the challenges in estimating the production function at the time. More recently, Basu and Fernald (1997) finds constant or slightly decreasing returns to scale at the industry level in the US and discusses issues around aggregation. Ahmad, Fernald, and Khan (2019) update these estimates with very similar results.

[69] The empirical success of models that build on a CRTS production sector, including the Solow model and the real business cycle framework, speak to the fact that scale invariance is an important feature of the data we aim to explain.

short and medium term fluctuations in both variables, appears to be widely accepted as an important macroeconomic pattern.

The only somewhat unconventional aspect of definition 1 is the insistence that growth patterns as measured in labour productivity be independent of scale, in particular labour input and labour supply growth. One of the most stunning macroeconomic patterns is the stability of the rate of productivity growth in economies near the technological frontier. For the past 150 years at least – as long as reliable records on output go back – the growth rate has remained close to 2% per year.[70] This observation is particularly surprising considering how much the world has changed during this long time span. Demographics are very different now from what they were in the past, with population growth dropping dramatically during the 20th century from its previously high levels. The openness of economies to international trade and investment has moved from largely unrestricted in the 19th century to protectionist in the first half of the 20th century and back to rather globalized today, likely impacting the effective scale of the markets supporting technological progress. Similarly, the part of the world that participates in the free exchange of goods, assets and ideas is much larger today than in the decades following World War II. As a much simpler example, regions within a country usually grow at similar rates as the whole economy on average over long time spans, even if they experience different rates of population growth.

B.3 Practice

The most commonly used production sector specification, that of the neoclassical growth model, best exemplified by a CRTS Cobb-Douglas technology with exogenous TFP growth at a constant rate, is fully consistent with definition 1. In fact, appendix A argues that any aggregate production function that meets our criteria is a straightforward generalization of such a technology.

Moreover, CRTS are a widely used production function characteristic that is applied not

[70] This point has been made a in particular by proponents of semi-endogenous growth models, see for example Jones (1995b) and Jones (2002).

only at the aggregate level, but also to firm-level production.[71]

The most important exceptions from the use of CRTS aggregate production exist in the area of growth, where both frameworks of endogenous and semi-endogenous growth regularly exhibit scale effects in levels and growth rates, thus being inconsistent with (C1) and (C3). The reasons for this are explained in detail in 2.1.2 and 3.1. Some remedies for the scale effects arising in endogenous growth models have been available for some time[72], and are employed occasionally, but many of the unusual characteristics of these model frameworks have been accepted over time. Semi-endogenous growth models inherently link long-term growth to population growth in the long run, although they can also be modified to yield population-independent growth for a long time.[73]

All popular model frameworks focused on explaining or describing long-term growth are consistent with balanced growth, criterion (C2), as the existence of a balanced growth path is widely considered to be a requirement for a growth model to be empirically plausible.

C The Relevant Unit

When discussing the composition of the firm sector, the level of granularity will be the efficient (or otherwise practically relevant) unit *of production* for individual goods. This deviates somewhat from the majority of the literature, which most commonly refers to the building blocks of the production sector as firms. The difference is that while the firm is the level at which production *decisions* are made, actual production is divided into units that are designed for cost-effective manufacturing of the relevant output.

[71] The example discussed in detail in this paper is search models of the labour market, see section 4.

[72] Young (1998) and Dinopoulos and Thompson (1998)

[73] It has been argued that the ability of the world economy to grow in the long term may be inherently tied to the scale of the of the economy, in particular if the cost of further technological advance increases with the level. This seems to suggest that it is impossible to develop a useful model of endogenous growth at any scale below that of the world economy. This, however, need not be the case. It is common to implicitly take linkages of an economy with the rest of a large world as given. Just as it makes sense to model aggregate production in Luxembourg in the same way as aggregate production in the USA, even though it is clear that a country with a population of less than 700,000 could not achieve its level of productivity in isolation, it should be possible to model the growth of Luxembourg endogenously, knowing that the economy shares in the knowledge of a much larger world.

In order to explain aggregate production possibilities and the response of aggregate output to changes in available factors, these efficient production units seem to be the right starting point. In a sufficiently competitive environment, who owns or runs a production unit should not matter much. Even in the presence of market power, one would expect production to still be organized efficiently, although the overall scale of output may be affected by the decisions of firms.

Firms are often large and complicated constructs that may be horizontally or vertically integrated. Even if the size of the efficient production unit is fixed, a firm may effectively produce at constant returns to scale by operating multiple of these units. For example, a microprocessor manufacturer may be a large firm, because chip design and R&D are very expensive and because there are network effects in the use of the product. At the same time, the manufacturing of the actual physical good could take place in smaller units that are either operated by the same or a different firm. For the purpose of capturing the response of aggregate output to changes in factor availability, the value-added contributions of each part of the firm, in this example the single design and R&D unit as well as multiple manufacturing plants, should at least conceptually be modelled as independent, efficiently operated production units.

All of this may be a side issue, but it could be relevant for choosing how to model the production sector under some circumstances. In this paper, I will be using the terms production unit and firm interchangeably.

D Aggregate Capital in the Firm-level Production Function

Suppose a firm produces under a labour productivity p, which depends on the technological knowledge i available in the firm as well as the economy-wide average level of knowledge

$\bar{i}$,

$$p = p(i) = \phi_1 i^{\phi_2} \bar{i}^{\phi_3}. \tag{14}$$

It is possible to increase this knowledge at a marginal cost $c(i) = \phi_4 i^{\phi_5} \bar{i}^{\phi_6}$, which in turn may depend (positively or negatively) on the firm's own and everyone else's level of technological insight. This specification allows for the standing-on-shoulders and stepping-on-toes effects discussed in the endogenous growth literature[74], as well as the idea that effecting a given relative improvement in productivity gets harder and more costly with the technology level rises, a notion that has been emphasized in the literature on semi-endogenous growth.[75]

To recast knowledge i as a regular capital good that is measured units of investment expenditure, we will write it in terms of its replacement value.

$$k(i) := \int_0^i c(j)dj$$

Solving the integral yields $k(i) = \frac{\phi_4}{\phi_5+1} i^{\phi_5+1} \bar{i}^{\phi_6}$. Let $k = k(i)$ be the firm's knowledge capital. We can write i in terms of this capital and $\bar{i}$.

$$i = \left(\frac{\phi_5 + 1}{\phi_4} k \bar{i}^{-\phi_6} \right)^{\frac{1}{\phi_5+1}} \tag{15}$$

Under symmetry, the typical firm in the economy has a productivity $\bar{p} = p(\bar{i}) = \phi_1 \bar{i}^{\phi_2+\phi_3}$, which, on a balanced growth path, is proportional to its capital-labour ratio $\kappa = \frac{\bar{k}}{\bar{n}} = \frac{K}{N}$, $\bar{p} = q\kappa$. We thus have

$$\bar{i} = \left(\frac{q}{\phi_1} \kappa \right)^{\frac{1}{\phi_2+\phi_3}}. \tag{16}$$

Using equations (15) and (16) to eliminate i and $\bar{i}$ from the definition of labour productivity (14), we get $p = \psi_1 k^{\psi_2} \kappa^{\psi_3}$ for constants ψ_i, which aligns with our choice of a unit-level production function in section 2.1.

[74] The popular varieties and Schumpeterian models are frameworks where such effects prominently arise, see Romer (1990) and Aghion and Howitt (1992).

[75] See footnote 24 above.