Technical Appendix
for Policy Paper
Estimating economic benefits of the Single Market for European countries and regions

Giordano Mion*
Dominic Ponattu†

May 2019

*University of Sussex, UK; CEP, UK; CEPR, UK; CESifo, Germany; NIESR, UK.
†Bertelsmann Stiftung, Germany.

1 Model

Our model builds on Behrens, Mion, Murata, and Südekum (2014, 2017). There are $r = 1, 2, \ldots, K$ countries or regions. For simplicity, we henceforth always use the term regions. Region r has a population of L_r workers who are also consumers. Each individual supplies inelastically one unit of labor. Labor is the only factor of production, and it is supplied locally (i.e., there is no cross-regional commuting). However, we assume that individuals are mobile across countries and regions, i.e., they are free to pick their place of residence.

We first set up the model in Sections 1.1 and 1.2, and then analyze the firm-level outcomes and the 'short run' equilibrium — when people are immobile and do not relocate across regions — in Section 1.3. We spell out the details concerning labor mobility in Section 1.4.

1.1 Preferences and demands

There is a continuum of horizontally differentiated varieties of final consumption goods and services. Consumers have identical preferences that display 'love of variety' and give rise to demands with variable elasticity. One key property of our model is that the marginal utility at zero consumption is bounded. Hence, consumers will not demand varieties for which the price (including trade costs) is too high. Those varieties are not traded across regions/countries, as is often the case for numerous services. Our model thus naturally applies to the analysis of the aggregate economy in which goods and services co-exist.

Let $p_{sr}(i)$ and $q_{sr}(i)$ denote the price and the per capita consumption of variety i when it is produced in region s and consumed in region r. The utility maximization problem of a representative individual in region r is given by:

$$\max_{q_{sr}(j),\, j \in \Omega_{sr}} \quad U_r \equiv \sum_s \int_{\Omega_{sr}} \left[1 - \mathrm{e}^{-\alpha q_{sr}(j)}\right] \mathrm{d}j \qquad \text{s.t.} \quad \sum_s \int_{\Omega_{sr}} p_{sr}(j) q_{sr}(j) \mathrm{d}j = E_r, \quad (1)$$

where $\alpha > 0$ is a utility parameter, and where Ω_{sr} denotes the endogenously determined set of varieties produced in s and consumed in r. As shown in Appendix A.1, solving (1) yields the following demand functions:

$$q_{sr}(i) = \frac{E_r}{N_r^c \overline{p}_r} - \frac{1}{\alpha} \left\{ \ln \left[\frac{p_{sr}(i)}{N_r^c \overline{p}_r} \right] + h_r \right\}, \quad \forall i \in \Omega_{sr}, \quad (2)$$

where N_r^c is the mass of varieties consumed in region r, and

$$\overline{p}_r \equiv \frac{1}{N_r^c}\sum_s \int_{\Omega_{sr}} p_{sr}(j)\mathrm{d}j \quad \text{and} \quad h_r \equiv -\sum_s \int_{\Omega_{sr}} \ln\left[\frac{p_{sr}(j)}{N_r^c\overline{p}_r}\right]\frac{p_{sr}(j)}{N_r^c\overline{p}_r}\mathrm{d}j$$

denote the average price and the differential entropy of the price distribution, respectively.

As explained before, demand does not need to be positive if the price charged for the variety is too high. Formally, as can be seen from (2), the demand for the local variety i (resp., the distant variety j) is positive if and only if the price of variety i (resp., variety j) is lower than some *choke price* p_r^d: $q_{rr}(i) > 0$ if and only if $p_{rr}(i) < p_r^d$; and $q_{sr}(j) > 0$ if and only if $p_{sr}(j) < p_r^d$, where $p_r^d \equiv N_r^c\overline{p}_r \mathrm{e}^{\alpha E_r/(N_r^c\overline{p}_r)-h_r}$ depends on the price aggregates $\overline{p}_r$ and h_r.

Using the definition of the choke price allows us to express the demands for local and distant varieties concisely as follows:

$$q_{rr}(i) = \frac{1}{\alpha}\ln\left[\frac{p_r^d}{p_{rr}(i)}\right] \quad \text{and} \quad q_{sr}(j) = \frac{1}{\alpha}\ln\left[\frac{p_r^d}{p_{sr}(j)}\right]. \tag{3}$$

The price elasticity of the local variety i (resp., the distant variety j) is given by $1/[\alpha q_{rr}(i)]$ (resp., $1/[\alpha q_{sr}(j)]$). Thus, if individuals consume more of those varieties, which is for instance the case when their expenditure increases, they become less price sensitive (see, e.g., Simonovska, 2015). Hence, the model allows us to take into account the fact that richer consumers are less price sensitive than poorer consumers.

Last, since $\mathrm{e}^{-\alpha q_{sr}(j)} = p_{sr}(j)/p_r^d$, the indirect utility in region r is given by

$$U_r = N_r^c - \sum_s \int_{\Omega_{sr}} \frac{p_{sr}(j)}{p_r^d}\mathrm{d}j = N_r^c\left(1 - \frac{\overline{p}_r}{p_r^d}\right). \tag{4}$$

Expression (4) will prove useful to compute the equilibrium utility in the subsequent analysis and to assess the consequences of changing trade costs.

1.2 Technology and market structure

The production side of the model features heterogeneous firms as in Melitz (2003) and Melitz and Ottaviano (2008). Prior to production, firms decide in which region they enter and they engage in research and development. The labor market in each region is perfectly competitive, so that all firms take the wage rate as given. Entry in region r requires a fixed amount F_r of labor paid at the market wage w_r. Each firm i that enters in region r discovers its marginal labor requirement $m_r(i) \geq 0$ only after making this irreversible entry decision. We assume that $m_r(i)$ is drawn from a known, continuously differentiable distribution G_r.[1] In what follows we assume, for simplicity, that firms' productivity draws $1/m$ follow a Pareto distribution

$$G_r(m) = \left(\frac{m}{m_r^{\max}}\right)^k,$$

with region-specific upper bounds, $m_r^{\max} > 0$, and a common shape parameter, $k \geq 1$. The Pareto distribution has been extensively used in the previous literature on heterogeneous firms (e.g., Bernard et al., 2007b; Helpman et al., 2008; Melitz and Ottaviano, 2008). It provides a good approximation of the distribution of firm sizes.

Shipments from region r to region s are subject to trade costs $\tau_{rs} > 1$ for all r and s, which firms incur in terms of labor. Put differently, the firm has to hire $\tau_{rs} - 1$ additional workers in order to ship the good from region r to region s. Those additional costs include, e.g., transportation costs of the good per se, but also different indirect trade costs including, for example, non-tariff barriers (NTB).

Since entry costs are sunk, firms will survive (i.e., operate) provided they can charge prices $p_{sr}(i)$ above marginal costs $\tau_{rs} m_r(i) w_r$ in at least one region. This usually includes the region the firm is located in, but the model allows for situations where a firm can survive only because of its distant demand and does not sell anything to its local market. While this situation seems not very relevant in an international context, it clearly is at smaller geographic scales such as the interregional context that we will focus on in what follows. The surviving firms produce in the region where they enter. We assume that firms do not relocate, i.e., once location choices have been made there is no relocation. Adding relocation makes the model complicated since it requires to deal with the spatial sorting of firms along productivity (see, e.g., Gaubert, 2015, for a model dealing with that question).

In line with empirical evidence, we assume that product markets are segmented,

[1] Differences in the sunk entry costs F_r and the productivity distributions G_r across regions/countries thus reflect production amenities such as startup costs, technology, and local knowledge that are only partly transferable across space, as well as differences in the institutional environments in which firms operate. Firms take those differences into account when making their entry decisions. Note that differences in start-up costs and institutions across countries are large (see, e.g., the World Bank's "Doing business" report; World Bank, 2016). Our model allows us to recover an implicit measure of these technological and institutional differences across regions/countries in equilibrium.

i.e., resale or third-party arbitrage is sufficiently costly, so that firms are free to price discriminate between regions. This is even a feature of a priori integrated economic areas such as the EU. There is substantial evidence that international product markets are segmented so that firms can do 'pricing-to-market' (see, e.g., Haskel and Wolf, 2001; Simonovska, 2015). While regions are a priori more integrated than countries, deviations from the law-of-one-price also apply there, as, e.g., seen from interregional border effects (Wolf, 2000).

The operating profit of a firm i located in region r is then as follows:

$$\pi_r(i) = \sum_s \pi_{rs}(i) = \sum_s L_s q_{rs}(i) \left[p_{rs}(i) - \tau_{rs} m_r(i) w_r \right], \tag{5}$$

where $\pi_{rs}(i)$ is the operating profit in market s, and $q_{rs}(i)$ is given by (3). Each surviving firm maximizes (5) with respect to its prices $p_{rs}(i)$ separately. Because there is a continuum of firms, no individual firm has any impact on p_r^d, so that the first-order conditions for (operating) profit maximization are given by:

$$\ln \left[\frac{p_s^d}{p_{rs}(i)} \right] = \frac{p_{rs}(i) - \tau_{rs} m_r(i) w_r}{p_{rs}(i)}, \quad \forall i \in \Omega_{rs}. \tag{6}$$

A price distribution satisfying (6) almost everywhere is called a *price equilibrium*. Equations (3) and (6) imply that $q_{rs}(i) = (1/\alpha)[1 - \tau_{rs} m_r(i) w_r / p_{rs}(i)]$. Thus, the minimum output that a firm in r may sell in market s is given by $q_{rs}(i) = 0$ at $p_{rs}(i) = \tau_{rs} m_r(i) w_r$. This, by (6), implies that $p_{rs}(i) = p_s^d$. Hence, a firm located in r with draw $m_{rs}^x \equiv p_s^d / (\tau_{rs} w_r)$ is just indifferent between selling and not selling to s, whereas all firms in r with draws below m_{rs}^x are productive enough to sell to s. In what follows, we refer to $m_{ss}^x \equiv m_s^d$ as the *local cutoff* in region s, whereas m_{rs}^x with $r \neq s$ is the *'export' cutoff* from region r to region s. Export and local cutoffs are linked by the following relationship:

$$m_{rs}^x = \frac{\tau_{ss}}{\tau_{rs}} \frac{w_s}{w_r} m_s^d. \tag{7}$$

Expression (7) reveals how trade costs and wage differences affect firms' abilities to break into different markets. In particular, when wages are the same in the two regions ($w_r = w_s$) and trade is costless ($\tau_{ss} = 1$), all export cutoffs must fall short of the local cutoffs since $\tau_{rs} > 1$. Breaking into market s is then always harder for firms in $r \neq s$ than its local competitors in s, which is the standard case considered in the literature (e.g., Melitz, 2003; Melitz and Ottaviano, 2008).[2]

Given the cutoffs (7) and the mass of entrants N_r^E, only $N_r^p = N_r^E\, G_r\left(\max_s \{m_{rs}^x\}\right)$ firms survive, namely those which are productive enough to sell at least in one market (which, as mentioned before, need not be their local market). The mass of varieties *consumed* in region r is given by

$$N_r^c = \sum_s N_s^E G_s(m_{sr}^x), \tag{8}$$

which is the mass of all firms that are productive enough to sell to market r. Utility changes in region r will be intimately linked to changes in N_r^c because consumers value variety in consumption.

1.3 Equilibrium with immobile labor

We now solve for the general equilibrium of our multi-regional trade model with heterogeneous firms. To do so, we first need to derive the firm-level outcomes in terms of prices, quantities, and profits. We relegate that part of the analysis and the corresponding technical details and expressions to Appendix A.2. We next need to consider three sets of equilibrium conditions. First, for each region, zero expected profit holds. Using equation (5), the zero expected profit condition (henceforth, ZEP) is given by

$$\sum_s L_s \int_0^{m_{rs}^x} \left[p_{rs}(m) - \tau_{rs} m w_r\right] q_{rs}(m) \mathrm{d}G_r(m) = F_r w_r. \tag{9}$$

Second, since there is no interregional commuting, local labor markets clear in each region. The labor market clearing condition (henceforth, LMC) requires that

$$N_r^E \left[\sum_s L_s \tau_{rs} \int_0^{m_{rs}^x} m q_{rs}(m) \mathrm{d}G_r(m) + F_r\right] = L_r. \tag{10}$$

Condition (10) states that the labor hired by firms to produce for both the local and the different distant markets, including the labor used to overcome trade costs and the labor hired to pay for the sunk entry costs (irrespective of whether the firm survives subsequently or not), sums to the regional labor endowment. The

[2] However, in the presence of wage differences and intra-regional trade costs τ_{rr}, the local cutoff need not be larger than the export cutoff in equilibrium. The usual ranking $m_s^d > m_{rs}^x$ prevails only when $\tau_{ss} w_s < \tau_{rs} w_r$.

latter will be endogenously determined by the location decisions of interregionally mobile individuals.

Last, trade must balance for each region, which is equivalent to saying that each consumer's budget constraint is satisfied with equality in each region. The trade balance condition (henceforth, TBC) for region r requires that the total value of exports equals the total value of imports, and it is given by

$$N_r^E \sum_{s \neq r} L_s \int_0^{m_{rs}^x} p_{rs}(m) q_{rs}(m) \mathrm{d}G_r(m) = L_r \sum_{s \neq r} N_s^E \int_0^{m_{sr}^x} p_{sr}(m) q_{sr}(m) \mathrm{d}G_s(m). \quad (11)$$

The $3 \times K$ general equilibrium conditions (9)–(11) depend on $3 \times K$ unknowns: the wages w_r, the masses of entrants N_r^E, and the local cutoffs m_r^d. Once the local cutoffs and the wages have been determined, the export cutoffs m_{rs}^x can be computed by using (7). Before proceeding, we simplify the general equilibrium conditions by using the Pareto parametrization and the results from Appendices A.2 and A.3. Using those results, the ZEP, LMC and TBC conditions can be rewritten as follows:

$$\mu_r^{\max} = \sum_s L_s \tau_{rs} \left(\frac{\tau_{ss}}{\tau_{rs}} \frac{w_s}{w_r} m_s^d \right)^{k+1}, \quad (12)$$

$$N_r^E \left[\frac{\kappa_1}{\alpha \left(m_r^{\max}\right)^k} \sum_s L_s \tau_{rs} \left(\frac{\tau_{ss}}{\tau_{rs}} \frac{w_s}{w_r} m_s^d \right)^{k+1} + F_r \right] = L_r, \quad (13)$$

$$\frac{N_r^E w_r}{\left(m_r^{\max}\right)^k} \sum_{s \neq r} L_s \tau_{rs} \left(\frac{\tau_{ss}}{\tau_{rs}} \frac{w_s}{w_r} m_s^d \right)^{k+1} = L_r \sum_{s \neq r} \tau_{sr} \frac{N_s^E w_s}{\left(m_s^{\max}\right)^k} \left(\frac{\tau_{rr}}{\tau_{sr}} \frac{w_r}{w_s} m_r^d \right)^{k+1}, \quad (14)$$

where $\mu_r^{\max} \equiv \left[\alpha F_r \left(m_r^{\max}\right)^k\right] / \kappa_2$ is a bundle of parameters that captures 'technological possibilities'. Note that $\mu_r^{\max}$ is region-specific and depends on both the sunk entry costs F_r and the upper bounds of the underlying productivity distribution $m_r^{\max}$. Thus, this bundle of parameters captures the local production amenities that are not transferable across space. It also subsumes aspects of the institutional environment of the region/country.

Combining (12) and (13), we obtain

$$N_r^E = \frac{\kappa_2}{\kappa_1 + \kappa_2} \frac{L_r}{F_r}, \quad (15)$$

which implies that more firms choose to enter in larger markets and in markets with lower entry requirements. Adding the term in r that is missing on both sides of (14), and using (12) and (15), we then obtain the following relationship:

$$\frac{1}{(m_r^d)^{k+1}} = \sum_s L_s \tau_{rr} \left(\frac{\tau_{rr}}{\tau_{sr}} \frac{w_r}{w_s} \right)^k \frac{1}{\mu_s^{\max}}. \quad (16)$$

The $2 \times K$ conditions (12) and (16), obtained by substituting out the equilibrium masses of entrants N_r^E, summarize how wages and cutoffs are related in general equilibrium, given the regional population sizes, technological possibilities, and trade costs.

Using these expressions, we can furthermore show that — in equilibrium — the mass of varieties consumed in region r is inversely proportional to the domestic cutoff, while the (expenditure share) weighted average of markups that consumers face is proportional to the local cutoff (see Appendix A.4 for the derivations):

$$N_r^c = \frac{1}{\kappa_1 + \kappa_2} \frac{\alpha}{\tau_{rr} m_r^d}, \quad (17)$$

$$\overline{\Lambda}_r^c \equiv \frac{\sum_s N_s^E \int_0^{m_{sr}^x} \frac{p_{sr}(m) q_{sr}(m)}{E_r} \Lambda_{sr}(m) \mathrm{d}G_s(m)}{\sum_s N_s^E G_s(m_{sr}^x)} = \frac{\kappa_3 \tau_{rr} m_r^d}{\alpha}, \quad (18)$$

where κ_1, κ_2, and κ_3 are positive constants that depend only on the common shape parameter k of the Pareto distribution.[3]

Under the Pareto parametrization, average productivity in region r is simply proportional to the inverse of the local cutoff:[4]

$$A_r = \frac{1}{G_r(m_j^d)} \int_0^{m_r^d} \frac{1}{m} \mathrm{d}G_r(m) = \frac{k_r}{k_r - 1} \frac{1}{m_r^d}. \quad (19)$$

Finally, the indirect utility in region r can be expressed as

$$U_r = \left[\frac{1}{(\kappa_1 + \kappa_2)(k+1)} - 1 \right] \frac{\alpha}{\tau_{rr} m_r^d} = \left[\frac{1}{(\kappa_1 + \kappa_2)(k+1)} - 1 \right] \frac{\kappa_3}{\overline{\Lambda}_r^c}, \quad (20)$$

[3] It can be seen from (17) and (18) that there are pro-competitive effects in our model, since $\overline{\Lambda}_r^c = [\kappa_3/(\kappa_1 + \kappa_2)](1/N_r^c)$ decreases with the mass of competing firms in region r.

[4] Alternatively, we can use the average (variable) labor productivity

$$\widetilde{A}_j = \left[\int_0^{m_j^d} q_j(m) \mathrm{d}G_j(m) \right] \cdot \left[\int_0^{m_j^d} m q_j(m) \mathrm{d}G_j(m) \right]^{-1} = \left(\frac{k_j + 1}{k_j} \right)^2 \frac{1}{m_j^d},$$

which generates quantitatively the same percentage productivity changes as A_j.

which implies that tougher selection (lower m_r^d) and fiercer competition (lower $\overline{\Lambda}_r^c$) both translate into higher utility in region r.[5]

1.4 Labor mobility and spatial equilibrium

Until now, we have taken the regional population sizes L_r as given. We now endogenize them by allowing individuals to move across regions to exploit differences in real incomes. To this end, we introduce taste heterogeneity in residential locations into our model. This is done for two reasons. First, individuals in reality choose their location not only based on wages, prices, and consumption diversity that result from market interactions, but also based on non-market features such as amenities (e.g., climate or landscape) and local social networks. The relatively low interregional mobility in Europe suggests that regional attachment is an important feature of individual location choices, and regional amenities and social networks certainly play a key role there (e.g., Faini et al., 1997; Faini, 1999). Second, individuals do not necessarily react in the same way to regional gaps in wages and cost-of-living. Such taste heterogeneity offsets the extreme — and counterfactual — outcome that often arises in typical agglomeration models with mobile individuals, namely that *all* mobile economic activity concentrates in a single region (Tabuchi and Thisse, 2002; Murata, 2003).

We assume that the location choice of an individual ℓ is based on a linear random utility $V_r^\ell = U_r + A_r + \xi_r^\ell$, where U_r is given by (20) and A_r subsumes region-specific amenities that are equally valued by all individuals. We usually do not observe A_r (or observe it only very imperfectly). The random variable ξ_r^ℓ captures idiosyncratic taste differences in residential location, subsuming many unobserved features such a social networks, amenities, and family ties. Following McFadden (1974), we assume that the ξ_r^ℓ are i.i.d. across individuals and regions according to a double exponential distribution with zero mean and variance equal to $\pi^2\beta^2/6$, where β is a positive constant. Since β has a positive relationship with variance, the larger the value of β,

[5] Alternatively, we have $U_r = [1/(k+1) - (\kappa_1 + \kappa_2)]N_r^c$, i.e., the indirect utility is proportional to the mass of varieties consumed. The utility gains come from imported varieties (Broda and Weinstein, 2006), as the mass of domestic varieties $N_r^E G_r(m_r^d)$ decreases when trade integration reduces the cutoff m_r^d. This finding is in line with those by Feenstra and Weinstein (2017), who show that new import varieties have contributed to US welfare gains even when taking into account the displaced domestic varieties.

the more heterogeneous are the consumers' attachments to each region. This makes, everything else equal, the population less sensitive to differences in regional utility differences that stem from differences in prices and wages.

Given the population distribution, an individual's probability of choosing region r can then be expressed as a logit form:

$$\mathbb{P}_r = \Pr\left(V_r^\ell > \max_{s \neq r} V_s^\ell\right) = \frac{\exp((U_r + A_r)/\beta)}{\sum_{s=1}^{K} \exp((U_s + A_s)/\beta)}. \tag{21}$$

For the distribution of population across regions to be non-degenerated, we assume that $\beta > 0$ in the subsequent analysis.[6] A *spatial equilibrium* is defined as a distribution of population across regions such that

$$\mathbb{P}_r = \frac{L_r}{\sum_{s=1}^{K} L_s}, \quad \forall r. \tag{22}$$

In words, a spatial equilibrium is a fixed point where the choice probability of each region is equal to that region's share of the economy's total population. This is a direct consequence of the law of large numbers. In theory, there can be multiple regional population distributions satisfying (22). However, this is not an issue given the aim of our paper. Indeed, in Section 2.3, when we fit our model to data, we plug the observed regional population shares into the right-hand side of (22) and uniquely back out $(U_r + A_r)/\beta$ such that this population distribution is a spatial equilibrium.

2 Quantification

To take our model to the data, we first derive a system of gravity equations and restate the general equilibrium conditions of the model. The gravity equation is required to estimate the trade frictions for goods and service trade from the data, whereas the general equilibrium conditions are required to take the model structurally to the data and to simulate the counterfactual impacts of the changes in trade barriers.

[6] If $\beta \to 0$, which corresponds to the case without taste heterogeneity, people choose their location based only on $U_r + A_r$, i.e., they choose the region with the highest $U_r + A_r$ with probability one. By contrast, if $\beta \to \infty$, individuals choose regions with equal probability $1/K$. In that case, regional tastes are extremely heterogeneous, so that $U_r + A_r$ does not affect location decisions at all.

2.1 Gravity equation system

We now derive a system of gravity equations that will be useful for taking the model to the data. The value of exports from region r to region s is given by

$$X_{rs} = N_r^E L_s \int_0^{m_{rs}^x} p_{rs}(m) q_{rs}(m) \mathrm{d}G_r(m).$$

Using (7), (A-3), (15), and the Pareto distribution for $G_r(m)$, we obtain the following gravity equation:

$$X_{rs} = L_r L_s \tau_{rs}^{-k} \tau_{ss}^{k+1} (w_s/w_r)^{k+1} w_r \left(m_s^d\right)^{k+1} (\mu_r^{\max})^{-1}. \tag{23}$$

As can be seen from (23), the value of shipments depend on bilateral trade costs τ_{rs}, internal trade costs in the destination region τ_{ss}, origin and destination regional wages w_r and w_s, the destination region cutoff m_s^d, and the origin region's technological possibilities $\mu_r^{\max}$. It is also increasing with the destination region's number of consumers, L_s, and the origin region's labor supply. A higher relative wage w_s/w_r raises the value of exports as firms in r face relatively lower production costs, whereas a higher absolute wage w_r raises the value of exports by increasing export prices p_{rs}. Furthermore, a larger m_s^d raises the value of exports since firms located in the destination are on average less productive. Last, a lower $\mu_r^{\max}$ implies that firms in region r have higher expected productivity, which raises the value of their exports. From the ZCP and the ZEP conditions, we further obtain the following general equilibrium conditions:

$$\mu_r^{\max} = \sum_s L_s \tau_{rs} \left(\frac{\tau_{ss}}{\tau_{rs}} \frac{w_s}{w_r} m_s^d\right)^{k+1}, \tag{24}$$

$$\frac{1}{\left(m_r^d\right)^{k+1}} = \sum_s L_s \tau_{rr} \left(\frac{\tau_{rr}}{\tau_{sr}} \frac{w_r}{w_s}\right)^k \frac{1}{\mu_s^{\max}}. \tag{25}$$

The $2 \times K$ general equilibrium conditions (24) and (25) summarize the interactions between the endogenous variables, namely the K wages and the K cutoffs. These conditions are reminiscent of those in Anderson and van Wincoop (2003), who argue that general equilibrium interdependencies need to be taken into account when conducting a counterfactual analysis based on the gravity equation.

Interestingly, the gravity equation system (23)–(25) is, indeed, akin to that in Anderson and van Wincoop (2003). To see this, let $Y_r = w_r L_r$ be the labor income of region r. Define the total labor income in all regions as $Y_w = \sum_r w_r L_r$ and the labor income share of region r as $\sigma_r = Y_r / Y_w$. Also define the multilateral resistance terms as follows:

$$\Phi_r^{-k} = \sigma_r^{k+1} \mu_r^{\max} L_r^{-k-1} \tag{26}$$

$$\Psi_s^{-k} = \sigma_s^{-k} \tau_{ss}^{-k-1} (m_s^d)^{-k-1} L_s^k. \tag{27}$$

Then, our gravity equation system (23)–(25) can be rewritten as follows:

$$X_{rs} = \frac{Y_r Y_s}{Y_W} \left(\frac{\tau_{rs}}{\Phi_r \Psi_s} \right)^{-k} \tag{28}$$

$$\Phi_r^{-k} = \sum_v \sigma_v \left(\frac{\tau_{rv}}{\Psi_v} \right)^{-k} \tag{29}$$

$$\Psi_s^{-k} = \sum_v \sigma_v \left(\frac{\tau_{vs}}{\Phi_v} \right)^{-k}, \tag{30}$$

which is the same as the gravity equation system (9)–(11) in Anderson and van Wincoop (2003), except that their exponent capturing the elasticity of substitution is replaced by the shape parameter k of the Pareto distributions. Assuming that $\tau_{rs} = \tau_{sr}$, i.e., trade costs are symmetric as in Anderson and van Wincoop (2003), we know that (29) and (30) yield a solution $\Phi_r = \Psi_r$ that solves the equations

$$\Phi_r^{-k} = \sum_v \sigma_v \tau_{rv}^{-k} \Phi_v^k. \tag{31}$$

We will use this property in the subsequent analysis as it greatly simplifies our quantification procedure.

2.2 Data

In order to make our model operational we need data on trade costs as well as on GDP and population. In order to recover trade costs we build on a gravity approach consistent with (28) and use data on trade in goods (services) coming from the COMTRADE (ITS) database provided by the United Nations (Eurostat) for the period 2010-2016. We also consider the usual set of gravity equation covariates provided by the Centre d'Etude Prospectives et d'Informations Internationales (CEPII): distance (d_{rs}), an ex-colony dummy ($Colony_{rs}$), a common language dummy ($Lang_{rs}$), a common border dummy ($Border_{rs}$) as well as a dummy indicating whether countries/regions r and s belong to the European Economic Area or not (EEA_{rs}).

In practical terms, we obtain the trade costs equivalent of the EEA and the SM from the estimation of a trade gravity equation from which we recover the parameter corresponding to the dummy EEA_{rs} and measuring the amount of additional trade EEA countries do with each other once discounted for other determinants of bilateral trade flows (distance, language, adjacency, past colonial ties). Such a parameter is an indicator of the trade-boosting effects of the EEA agreement and the Single Market and is the key to our counterfactual analysis.

As for population and GDP we borrow this data from the Eurostat Regio Database (for EEA regions) and the World Economic Outlook Database provided by the IMF (for non-EEA countries). Data on population and GDP refers to the year 2016. The 45 countries included in our analysis are all current members of the EEA plus BRIC and other OECD countries: Australia, Brazil, Canada, Chile, China, India, Israel, Japan, Korea, Mexico, New Zealand, Russia, Turkey and the US. In the first part of our analysis we quantity our model and do counterfactual analysis at the country-level for both EEA and non-EEA countries. In the second part of our analysis, we break down EEA countries into the corresponding NUTS-2 regions. We use the GDP of country/region r as a measure of Y_r, population as a measure of L_r and GDP per capita as a proxy for w_r.

2.3 Quantification procedure

We now explain the numerical procedure that we implement to calibrate the model to the initial equilibrium. The steps of our numerical procedure work as follows.

1. We specify trade costs as $\tau_{rs} \equiv d_{rs}^{\gamma} \mathrm{e}^{\theta_1 EEA_{rs}} \mathrm{e}^{\theta_2 Colony_{rs}} \mathrm{e}^{\theta_3 Lang_{rs}} \mathrm{e}^{\theta_4 Borger_{rs}}$. We are particularly interested in the coefficient corresponding to membership of the EEA: θ_1

2. Given our specification of trade costs τ_{rs}, the gravity equation (28) can be rewritten in stochastic form as follows:

$$X_{rs} = \frac{Y_r Y_s}{Y_W} \left(\frac{\tau_{rs}}{\Phi_r \Psi_s} \right)^{-k} \varepsilon_{rs}, \tag{32}$$

where ε_{rs} is an error term with the usual properties. We estimate (32) at the *country-level* for our group of EEA and non-EEA countries using the Poisson Pseudo Maximum Likelihood (PPML) method suggested in Santos Silva and

Tenreyro (2006). We do this in a way that is consistent with (28) by using origin and destination fixed effects to control for multilateral resistance terms Φ_r and Ψ_s as well as GDP Y_r and Y_s.[7] In Behrens et al. (2014), we also quantify the value of k. To this end, we compute the productivity advantage of US exporters from a random sample of firms drawn from the fitted productivity distributions of our model. We repeat this procedure for different values of k until our sample matches the 33% productivity advantage of US exporters in 1992, which is reported by Bernard et al. (2003). See Behrens et al. (2014) for details. Here, we use their value of $\widehat{k} = 8.5$ in the analysis.

3. Using estimates from the stochastic gravity regression (32), we construct trade costs. In the first part of our analysis we quantity our model and do counterfactual analysis at the country-level for both EEA and non-EEA countries and so compute trade costs across countries. In the second part of our analysis, we break down EEA countries into the corresponding NUTS-2 regions and thus compute trade costs across EEA regions and non-EEA countries.[8]

   Trade costs enter the gravity equation (28) as $\tau_{rs}^{-k} \equiv \phi_{rs} \in (0, 1)$ where ϕ_{rs} is an inverse measure of trade costs, i.e., the freeness of trade, and so we actually compute a measure of freeness of trade corresponding to the initial trading equilibrium. We do this separately for goods and services gravity regressions and then average the two sets of ϕ_{rs} by using world trade shares of trade in goods (75%) and services (25%).

4. We observe the initial values of regional/national populations L_r^0 and GDP $w_r^0 L_r^0$ from the data and so we can compute income shares σ_r^0. Since our trade costs are symmetric, we solve the system

$$\Phi_r^{-k} = \sum_v \sigma_v^0 \tau_{rv}^{-k} \Phi_v^k, \tag{33}$$

[7] We do not make use of the full general equilibrium system. Doing so makes actually little difference. See Behrens et al. (2014) for an estimation of the full system using US-Canada data.

[8] We assign trade costs between, for example, any UK NUTS-2 region and the US to be the same and equal to the trade costs between the UK and the US computed from (32). As for trade costs between the NUTS-2 regions of, for example, London and Rome we use country-level values for variables other than distance while for the latter we actually use the distance between London and Rome, along with our estimate of γ, to compute the distance-related component of trade costs.

for the Φ_r terms. Call that solution $\widehat{\Phi}_r^0$, where the hat stands for 'quantified' and where 0 is the initial iteration.

5. Using (26) and (27), we solve

$$\begin{aligned}(\widehat{\Phi}_s^0)^{-k} &= (\sigma_s^0)^{-k}\tau_{ss}^{-k-1}(m_s^d)^{-k-1}(L_s^0)^k \\ (\widehat{\Phi}_r^0)^{-k} &= (\sigma_r^0)^{k+1}\mu_r^{\max}(L_r^0)^{-k-1}\end{aligned}$$

for the cutoff $(\widehat{m}_s^d)^0$ and the unobserved upper bounds $\widehat{\mu}_r^{\max}$.

6. Using $(\widehat{m}_s^d)^0$, we use (20) to compute the indirect utility due to the consumption of the differentiated varieties:

$$\widehat{U}_r^0 = \frac{\alpha}{\tau_{rr}}\left[\frac{1}{(k+1)(\kappa_1+\kappa_2)} - 1\right]\frac{1}{(\widehat{m}_r^d)^0} \propto \frac{1}{\tau_{rr}}\frac{1}{(\widehat{m}_r^d)^0}. \tag{34}$$

We compute this up to a scaling that does not matter for the equilibrium (the level of utility is immaterial, and it cannot be meaningfully used).

7. Finally, we calibrate the model to replicate the initial distribution of population as an equilibrium. To this end, we use the initial populations and solve the logit equation system (21) as follows:

$$\frac{L_r^0}{\sum_s L_s^0} = \frac{\exp(D_r)}{\sum_s \exp(D_s)}, \tag{35}$$

for the D_r terms, using a linear random utility (LRU) as explained in Section 1.4. Using the quantified values of $\widehat{D}_r^0$ and $\widehat{U}_r^0$ we have $\widehat{A}_r = \widehat{D}_r^0 - \widehat{U}_r^0$. These are the (observed and unobserved) amenities that sustain the spatial equilibrium that we observe from the data. These amenities will be held fixed in the counterfactuals, just as the upper bounds $\widehat{\mu}_r^{\max}$ are held fixed. Note that we use equal weighting of utility and amenities in what follows. This has no strong implications for our results. We could use different weighting schemes, notably ones that are estimated using available amenity data and geological instruments to deal with potential problems of reverse causality (see Behrens et al., 2017).

The foregoing seven steps allow us to bring the model to the data and to replicate the observed regional population distribution as a spatial general equilibrium of the model.

2.4 Gravity estimation results

Table 1 below reports *country-level* gravity estimation results for goods (column 1) and services (column 2). As one can notice all coefficients (but colony in the trade in goods regression) are significant and have the usual sign and magnitude. In particular, the EEA dummy is positive and significant for both goods and service while being larger in the latter. Using estimates from Table 1 we construct trade costs corresponding to the initial trading equilibrium. In the first part of our analysis, we quantity our model and do counterfactual analysis at the country-level for both EEA and non-EEA countries and so we use trade costs at the country-level. In the second part of our analysis, we break down EEA countries into the corresponding NUTS-2 regions and so compute trade costs across EEA regions and non-EEA countries. In particular, we assign trade costs between, for example, any UK NUTS-2 region and the US to be the same and equal to the trade costs between the UK and the US computed from estimations of (32). As for trade costs between the NUTS-2 regions of, for example, London and Rome we use country-level values for variables other than distance while for the latter we actually use the distance between London and Rome, along with our estimate of the distance elasticity γ, to compute the distance-related component of trade costs.

We do this separately for goods and services gravity regressions and then average the two sets of ϕ_{rs} by using world trade shares of trade in goods (75%) and services (25%). Then, we also construct the counterfactual freeness $\widetilde{\phi}_{rs}$ that would prevail in the counterfactual scenario we consider. The main counterfactual we consider is one in which trade (in goods and services) between EEA countries would not be subject to the boosting effect of those trade facilitation policies put in place by the EEA agreement and the SM. To implement this, we update the dummy variable EEA_{rs} by imposing it is equal to 0 for all countries. With the counterfactual $\widetilde{EEA}_{rs}$ in our hands we then compute the counterfactual freeness $\widetilde{\phi}_{rs}$ as well as the related economic impacts.

3 Counterfactual analysis

We now run a counterfactual exercise to gauge the importance of the trade boosting effect of the EEA and the SM. To this end, we shock the initial equilibrium and let the system settle into a new equilibrium, taking into account all general equilibrium

Table 1: Gravity estimation results for goods (column 1) and services (column 2)

	Goods	Services
Distance	-0.694[a]	-0.742[a]
	(0.023)	(0.037)
EEA dummy	0.462[a]	0.906[a]
	(0.069)	(0.061)
Colony dummy	-0.046	0.218[a]
	(0.035)	(0.051)
Language dummy	0.152[a]	0.191[a]
	(0.038)	(0.056)
Border dummy	0.580[a]	0.214[a]
	(0.034)	(0.043)
Year dummies	Yes	Yes
Origin and Destination dummies	Yes	Yes
Observations	13,869	7,765
Pseudo R^2	0.936	0.924

Poisson Pseudo Maximum Likelihood estimations. Robust standard errors in parentheses. [abc] indicate the significance of the coefficient, [a] p<0.01, [b] p<0.05, [c] p<0.1.

effects and the mobility of people. Doing so allows us to simulate the impacts of such counterfactual on productivity, markups, product variety, welfare and the distribution of population across European countries and regions.

3.1 Numerical procedure for the counterfactuals

Formally, running our counterfactuals entails the following steps:

1. We shock trade costs changing from τ_{rs} to $\widetilde{\tau}_{rs}$. We first eliminate Ψ_v from (29) by substituting (30) to obtain:

$$\Phi_r^{-k} = \sum_v \frac{\sigma_v \widetilde{\tau}_{rv}^{-k}}{\sum_s \sigma_s \left(\frac{\widetilde{\tau}_{sv}}{\Phi_s}\right)^{-k}}. \tag{36}$$

Plugging (26) into both sides of (36) yields a system of equations that depends on the labor income shares σ_r only as follows:

$$(\sigma_r^t)^{k+1}\widehat{\mu}_r^{\max}(L_r^t)^{-k-1} = \sum_v \frac{\sigma_v^t \widetilde{\tau}_{rv}^{-k}}{\sum_s (\sigma_s^t)^{-k}\widetilde{\tau}_{sv}^{-k}(\widehat{\mu}_s^{\max})^{-1}(L_s^t)^{k+1}}, \tag{37}$$

where superscript t denotes the current iteration of the system ($t = 0$ at the beginning of the counterfactual). This system of equations holds exactly — since it has been calibrated in that way — at the initial shares σ_r^0 and populations L_r^0, given initial trade costs τ_{rs} and the upper bounds $\widehat{\mu}_r^{\max}$. However, it no longer holds for the counterfactual trade costs $\widetilde{\tau}_{rs}$. We hence solve that system for the new income shares σ_r^{t+1} that make it hold with equality. Since the system is not independent, we drop one of the equations and impose the constraint that the income shares sum to one: $\sum_v \sigma_v = 1$. The new labor income shares σ_r^{t+1} are those that would prevail after the shock and conditional on the *old* population distribution of the previous iteration t.

2. Using

$$\Phi_s^{-k} = (\sigma_r^{t+1})^{k+1}\widehat{\mu}_r^{\max}(L_r^t)^{-k-1}$$

we solve for the new multilateral resistance terms, $\widehat{\Phi}_r^{t+1}$, given the initial population distribution at iteration t and the new income shares at iteration $t+1$. Using those terms, we then solve

$$(\widehat{\Phi}_s^{t+1})^{-k} = (\sigma_s^{t+1})^{-k}\widetilde{\tau}_{ss}^{-k-1}(m_s^d)^{-k-1}(L_s^t)^k$$

for the new cutoffs $\widehat{m}_s^{d,t+1}$.

3. We construct the new utility

$$\widehat{U}_r^{t+1} \propto \frac{1}{\widetilde{\tau}_{rr}}\frac{1}{\widehat{m}_r^{d,t+1}}$$

associated with the new cutoffs. Given the trade shock, these utility levels will have changed from the initial equilibrium. Hence, the spatial allocation is no longer an equilibrium, i.e., some individuals have incentives to change location in order to take advantage of changes in prices and wages.

4. We hence solve

$$\frac{L_r}{\sum_s L_s} = \frac{\exp(\widehat{A}_r + \widehat{U}_r^{t+1})}{\sum_s \exp(\widehat{A}_s + \widehat{U}_s^{t+1})}, \tag{38}$$

for the new population distribution L_r^{t+1}. Since the system (38) is not independent, we drop one equation and recoup the final population by using the adding-up constraint $L = \sum_s \widehat{L}_s^{t+1}$ at all periods t (the total population of the system is held constant).

5. We go back to step 1 of the procedure. Since the populations have changed, the income shares need to adjust to solve (37). We solve for the new shares and iterate steps 1–4 of the above procedure until convergence is achieved. Letting $\mathbf{L}^t$ denote the vector of populations across regions at iteration t of the algorithm, we define convergence as $\|\mathbf{L}^{t+1} - \mathbf{L}^t\| \leq \varepsilon$, i.e., when the change in population between two consecutive iterations becomes sufficiently small.

In Behrens et al. (2017) we prove existence and uniqueness of the initial equilibrium, and we also show that any shock to the system leads (conditional on the initial equilibrium) to a unique counterfactual equilibrium. Hence, our framework is well-suited to investigate the implications of a trade shock on productivity, markups, product diversity, welfare and the regional distribution of population.

3.2 Computing changes in Welfare

One way of computing changes in welfare for the representative consumer is to consider changes in utility U_r. However, this approach has a number of known shortcoming and is particularly problematic in our setting given that the CARA preferences structure we use is non-homothetic. We thus consider here another approach based on the concept of equivalent variation. More specifically, we will compute the change in income that, given initial prices, would allow the representative consumer to reach the same utility level corresponding to the counterfactual equilibrium. Loosely speaking, this corresponds to the income reduction/increase equivalent of the counterfactual scenario.

In order to compute the equivalent variation we build on the results laid down in Arkolakis et al. (2018). Arkolakis et al. (2018) show that for a large class of models, that includes ours, there is a common formula to compute the equivalent variation.[9]

[9]Strictly speaking, their formula applies to small changes in trade costs and assumes constant population. However, Arkolakis et al. (2018) provide evidence that the formula does a very good job in matching large changes in trade costs too. Furthermore, simulations of our model with and without labour mobility deliver changes in utility U_r that are very close to each other and this implies

Denoting with W_r welfare and with dW_r the change in welfare (equivalent variation) the formula is:

$$dW_r = -(1 - \eta)\, d\, ln(\lambda_r)/k \tag{39}$$

where k is the shape parameter of the Pareto distribution of productivity (8.5 in our case), $d\, ln(\lambda_r)$ is the change in the (log of) the share of domestic expenditure on domestic goods caused by the change in trade costs, and η is a parameter that is specific to the preferences structured used and k (0.9551 in our case). Therefore, the only thing needed to compute counterfactual changes in welfare is to compute changes in the (log of) the share of domestic expenditure on domestic goods. To achieve this we compute the share of domestic expenditure on domestic goods (X_{rr}/Y_r) before and after changes in trade costs and compare them. Using (39), as well as the value of k and η, we are in turn able to compute changes in welfare.

that the equivalent variation figures we obtain with labour mobility are very close to those that we would obtain without labour mobility.

References

[1] Anderson, James E., and Eric van Wincoop. 2003. "Gravity with Gravitas: A Solution to the Border Puzzle." *American Economic Review* 93(1): 170–192.

[2] Arkolakis, C., Costinot, A., and Rodriguez-Clare, A. 2012. "New trade models, same old gains?" *American Economic Review*, 102(1): 94–130.

[3] Arkolakis, C., Costinot, A., Donaldson, D., and Rodriguez-Clare, A. 2018. "The elusive pro-competitive effects of trade." *Review of Economic Studies*, forthcoming.

[4] Behrens, Kristian, Giordano Mion, Yasusada Murata, and Jens Suedekum. 2014. "Trade, Wages, and Productivity." *International Economic Review* 55(4): 1305–1348.

[5] Behrens, Kristian, Giordano Mion, Yasusada Murata, and Jens Suedekum. 2017. "Spatial Frictions." *Journal of Urban Economics* 97: 40–70.

[6] Broda, Christian, and David E. Weinstein. 2006. "Globalization and the Gains from Variety." *Quarterly Journal of Economics* 121(2): 541–585.

[7] De Loecker, Jan. 2011. "Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity." *Econometrica* 79(5): 1407–1451.

[8] De Loecker, Jan, Pinilopi K. Goldberg, Amit K. Khandelwal, and Nina Pavcnik. 2016. "Prices, markups and trade reform." *Econometrica* 84(2): 445–510.

[9] Bernard, Andrew B., J. Bradford Jensem, Stephen J. Redding, and Peter K. Schott. 2007. "Firms in international trade." *Journal of Economic Perspectives* 21(3): 105–130.

[10] Faini, Riccardo. 1999. "European migrants: an endangered species?" In: *Market integration, regionalism and the global economy*. Cambridge University Press, Cambridge, 228–251.

[11] Faini, Riccardo, Giampaolo Galli, and Pietro Gennari. 1997. "An empirical puzzle: falling migration and growing unemployment differentials among Italian regions." *European Economic Review* 41(3): 571–579.

[12] Feenstra, Robert C., and David E. Weinstein. 2017. "Globalization, Markups and U.S. Welfare". *Review of Economic Studies*.

[13] Gaubert, Cecile. 2015. "Firm sorting and agglomeration." Mimeo, University of California, Berkeley.

[14] Haskel, Jonathan, and Holger Wolf. 2001. "The Law of One Price—A Case Study." *Scandinavian Journal of Economics* 103(4): 545–558.

[15] Helpman, Elhanan, Marc J. Melitz, and Yolanda Rubinstein. 2008. "Estimating Trade Flows: Trading Partners and Trading Volumes." *Quarterly Journal of Economics* 123(2): 441–487.

[16] McFadden, Daniel. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior." In: Zarembka, P. (Ed.), *Frontiers in Econometrics.* Academic Press, New York, 105–142.

[17] Melitz, Marc J. 2003. "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity." *Econometrica* 71(6): 1695–1725.

[18] Melitz, Marc J., and Gianmarco I.P. Ottaviano. 2008. "Market Size, Trade, and Productivity." *Review of Economic Studies* 75(1): 295–316.

[19] Murata, Yasusada. 2003. "Product diversity, taste heterogeneity, and geographic distribution of economic activities: market vs. non-market interactions." *Journal of Urban Economics* 53(1): 126–144.

[20] Santos Silva, J.M.C., Tenreyro, S. 2006. "The log of gravity". *The Review of Economics and statistics*, 88(4): 641–658.

[21] Simonovska, Ina. 2015. "Income differences and prices of tradables: Insights from an online retailer." *Review of Economic Studies* 82(4): 1612–1656.

[22] Tabuchi, Takatoshi, and Jacques-François Thisse. 2002. "Taste heterogeneity, labor mobility and economic geography." *Journal of Development Economics* 69(1): 155–177.

[23] Wolf, Holger C. 2000. "Intranational home bias in trade." *Review of Economics and Statistics* 82(4): 555–563.

[24] World Bank. 2016. *Doing Business 2016. Measuring Regulatory Quality and Efficiency.* International Bank for Reconstruction and Development/The World Bank: Washington, DC.

Appendix

Appendix A: Computational details

A.1. Derivation of the demand functions.

In this appendix, we derive expression (2). Let λ denote the Lagrange multiplier. The first-order condition for an interior solution to the maximization problem (1) satisfies

$$\alpha \mathrm{e}^{-\alpha q_{sr}(i)} = \lambda p_{sr}(i), \quad \forall i \in \Omega_{sr} \tag{A-1}$$

and the budget constraint $\sum_s \int_{\Omega_{sr}} p_{sr}(k) q_{sr}(k) \mathrm{d}k = E_r$. Taking the ratio of (A-1) for $i \in \Omega_{sr}$ and $j \in \Omega_{vr}$ yields

$$q_{sr}(i) = q_{vr}(j) + \frac{1}{\alpha} \ln \left[\frac{p_{vr}(j)}{p_{sr}(i)} \right] \quad \forall i \in \Omega_{sr}, \forall j \in \Omega_{vr}.$$

Multiplying this expression by $p_{vr}(j)$, integrating with respect to $j \in \Omega_{vr}$, and summing across all origin regions v we obtain

$$q_{sr}(i) \sum_v \int_{\Omega_{vr}} p_{vr}(j) \mathrm{d}j = \underbrace{\sum_v \int_{\Omega_{vr}} p_{vr}(j) q_{vr}(j) \mathrm{d}j}_{\equiv E_r} + \frac{1}{\alpha} \sum_v \int_{\Omega_{vr}} \ln \left[\frac{p_{vr}(j)}{p_{sr}(i)} \right] p_{vr}(j) \mathrm{d}j. \tag{A-2}$$

Using $\overline{p}_r \equiv (1/N_r^c) \sum_v \int_{\Omega_{vr}} p_{vr}(j) \mathrm{d}j$, expression (A-2) can be rewritten as follows:

$$\begin{aligned} q_{sr}(i) &= \frac{E_r}{N_r^c \bar{p}_r} - \frac{1}{\alpha} \ln p_{sr}(i) + \frac{1}{\alpha N_r^c \bar{p}_r} \sum_v \int_{\Omega_{vr}} \ln \left[p_{vr}(j) \right] p_{vr}(j) \mathrm{d}j \\ &= \frac{E_r}{N_r^c \bar{p}_r} - \frac{1}{\alpha} \ln \left[\frac{p_{sr}(i)}{N_r^c \bar{p}_r} \right] + \frac{1}{\alpha} \sum_v \int_{\Omega_{vr}} \ln \left[\frac{p_{vr}(j)}{N_r^c \bar{p}_r} \right] \frac{p_{vr}(j)}{N_r^c \bar{p}_r} \mathrm{d}j, \end{aligned}$$

which, given the definition of h_r, yields (2).

A.2. Derivation of the firm-level variables and properties of W

Since firms in country r differ only by their marginal labor requirements, we can express all firm-level variables in terms of m. Solving the first-order conditions (6) for profit maximization, the profit-maximizing prices and quantities, as well as operating profits, are given by:

$$p_{rs}(m) = \frac{\tau_{rs} m w_r}{W}, \quad q_{rs}(m) = \frac{1}{\alpha}(1 - W), \quad \pi_{rs} = \frac{L_s \tau_{rs} m w_r}{\alpha}(W^{-1} + W - 2), \quad \text{(A-3)}$$

where W denotes the Lambert W function with argument $\mathrm{e}m/m^x_{rs}$ that we suppress to alleviate notation (see Corless *et al.*, 1996, for a survey). To derive expressions (A-3) use $p^d_s = m^x_{rs}\tau_{rs}w_r$ so that the first-order condition (6) can be rewritten as

$$\ln\left[\frac{m^x_{rs}\tau_{rs}w_r}{p_{rs}(m)}\right] = 1 - \frac{\tau_{rs} m w_r}{p_{rs}(m)}.$$

Taking the exponential of both sides and rearranging terms, we have

$$\mathrm{e}\frac{m}{m^x_{rs}} = \frac{\tau_{rs} m w_r}{p_{rs}(m)} \mathrm{e}^{\frac{\tau_{rs} m w_r}{p_{rs}(m)}}.$$

Noting that the Lambert W function is defined as $\varphi = W(\varphi)\mathrm{e}^{W(\varphi)}$ and setting $\varphi = \mathrm{e}m/m^x_{rs}$, we obtain $W(\mathrm{e}m/m^x_{rs}) = \tau_{rs} m w_r / p_{rs}(m)$, which implies $p_{rs}(m)$ as given in (A-3). The expression for quantities $q_{rs}(m) = (1/\alpha)\left[1 - \tau_{rs} m w_r / p_{rs}(m)\right]$ and the expression for operating profits $\pi_{rs}(m) = L_s q_{rs}(m)\left[p_{rs}(m) - \tau_{rs} m w_r\right]$ are then straightforward to compute.

Turning to the properties of the Lambert W function, $\varphi = W(\varphi)\mathrm{e}^{W(\varphi)}$ implies that $W(\varphi) \geq 0$ for all $\varphi \geq 0$. Taking logarithms on both sides of the definition of W and differentiating yields

$$W'(\varphi) = \frac{W(\varphi)}{\varphi[W(\varphi) + 1]} > 0$$

for all $\varphi > 0$. Finally, we have $0 = W(0)\mathrm{e}^{W(0)}$, which implies $W(0) = 0$; and $\mathrm{e} = W(\mathrm{e})\mathrm{e}^{W(\mathrm{e})}$, which implies $W(\mathrm{e}) = 1$.

Since $W(0) = 0$, $W(\mathrm{e}) = 1$ and $W' > 0$ for all non-negative arguments, we have $0 \leq W \leq 1$ if $0 \leq m \leq m^x_{rs}$. The expressions in (A-3) show that a firm in r with a draw m^x_{rs} (equal to the cutoff labor requirement for selling to market s) charges a price equal to marginal cost, faces zero demand, and earns zero operating profits in market s. Furthermore, it follows that $\partial p_{rs}(m)/\partial m > 0$, $\partial q_{rs}(m)/\partial m < 0$, and $\partial \pi_{rs}(m)/\partial m < 0$. In words, firms with higher productivity (lower m) charge lower prices, sell larger quantities, and earn higher operating profits. These properties are similar to those of the Melitz (2003) model with CES preferences. Yet, our specification with variable demand elasticity also features higher markups for more productive firms (see, e.g., de Loecker, 2011; de Locker et al., 2016). Indeed, the *origin-destination markup* for a firm located in country r and selling to country s is given by

$$\Lambda_{rs}(m) \equiv \frac{p_{rs}(m)}{\tau_{rs} m w_r} = \frac{1}{W}, \tag{A-4}$$

thus implying that $\partial \Lambda_{rs}(m)/\partial m < 0$. Melitz and Ottaviano (2008) have a similar effect in their model, yet they use quasi-linear preferences which makes the model not really amenable to counterfactual analysis. We incorporate this feature of markups into a full-fledged general equilibrium model with income effects for varieties that can be taken neatly to the data.

A.3. Equilibrium conditions using the Lambert W function

In this appendix, we restate the equilibrium conditions (9)–(11) for the multicountry case using the Lambert W function.

First, plugging (A-3) into (9), zero expected profits can be rewritten as

$$\frac{1}{\alpha} \sum_s L_s \tau_{rs} \int_0^{m_{rs}^x} m \left[W\left(\mathrm{e}\frac{m}{m_{rs}^x} \right)^{-1} + W\left(\mathrm{e}\frac{m}{m_{rs}^x} \right) - 2 \right] \mathrm{d}G_r(m) = F_r. \tag{A-5}$$

Observe that this condition depends solely on the cutoffs m_{rs}^x and that it is independent of the mass of entrants. Using (A-3), the labor market clearing condition (10) becomes

$$N_r^E \left\{ \frac{1}{\alpha} \sum_s L_s \tau_{rs} \int_0^{m_{rs}^x} m \left[1 - W\left(\mathrm{e}\frac{m}{m_{rs}^x} \right) \right] \mathrm{d}G_r(m) + F_r \right\} = L_r. \tag{A-6}$$

Finally, using (A-3) the trade balance condition (11) is given by

$$\begin{aligned} N_r^E w_r \sum_{s \neq r} L_s \tau_{rs} \int_0^{m_{rs}^x} m \left[W\left(\mathrm{e}\,\frac{m}{m_{rs}^x} \right)^{-1} - 1 \right] \mathrm{d}G_r(m) \\ = L_r \sum_{s \neq r} N_s^E \tau_{sr} w_s \int_0^{m_{sr}^x} m \left[W\left(\mathrm{e}\,\frac{m}{m_{sr}^x} \right)^{-1} - 1 \right] \mathrm{d}G_s(m). \end{aligned} \tag{A-7}$$

We next apply the region-specific Pareto distributions $G_r(m) = (m/m_r^{\max})^k$ to the system (A-5)–(A-7). We then have a number of integrals that involve the Lambert

W function. To compute closed-form expressions for those integrals, we use the change in variables suggested by Corless *et al.* (1996, p.341). Let

$$z \equiv W\left(\mathrm{e}\frac{m}{I}\right), \quad \text{so that} \quad \mathrm{e}\frac{m}{I} = z\mathrm{e}^{z}, \quad \text{where} \quad I \in \{m_r^d, m_{rs}^x\},$$

where we drop the subscript r to alleviate notation. The change in variables then yields $\mathrm{d}m = (1+z)\mathrm{e}^{z-1}I\mathrm{d}z$, with the new integration bounds given by 0 and 1. This allows us to rewrite all integrals in simplified form.

A.3.1. First, consider the following expression, which appears when integrating firms' outputs:

$$\int_0^I m\left[1 - W\left(\mathrm{e}\frac{m}{I}\right)\right]\mathrm{d}G_r(m) = \kappa_1\left(m_r^{\max}\right)^{-k} I^{k+1},$$

where $\kappa_1 \equiv k\mathrm{e}^{-(k+1)}\int_0^1 (1-z^2)\left(z\mathrm{e}^{z}\right)^k \mathrm{e}^{z}\mathrm{d}z > 0$ is a constant term which solely depends on the shape parameter k.

A.3.2. Second, the following expression appears when integrating firms' operating profits:

$$\int_0^I m\left[W\left(\mathrm{e}\frac{m}{I}\right)^{-1} + W\left(\mathrm{e}\frac{m}{I}\right) - 2\right]\mathrm{d}G_r(m) = \kappa_2\left(m_r^{\max}\right)^{-k} I^{k+1},$$

where $\kappa_2 \equiv k\mathrm{e}^{-(k+1)}\int_0^1 (1+z)\left(z^{-1}+z-2\right)\left(z\mathrm{e}^{z}\right)^k \mathrm{e}^{z}\mathrm{d}z > 0$ is a constant term which solely depends on the shape parameter k.

A.3.3. Third, the following expression appears when deriving the (expenditure share) weighted average of markups:

$$\int_0^I m\left[W\left(\mathrm{e}\frac{m}{I}\right)^{-2} - W\left(\mathrm{e}\frac{m}{I}\right)^{-1}\right]\mathrm{d}G_r(m) = \kappa_3\left(m_r^{\max}\right)^{-k} I^{k+1},$$

where $\kappa_3 \equiv k\mathrm{e}^{-(k+1)}\int_0^1 (z^{-2}-z^{-1})(1+z)(z\mathrm{e}^{z})^k\mathrm{e}^{z}\mathrm{d}z > 0$ is a constant term which solely depends on the shape parameter k.

A.3.4. Finally, the following expression appears when integrating firms' revenues:

$$\int_0^I m \left[W \left(\mathrm{e} \frac{m}{I} \right)^{-1} - 1 \right] \mathrm{d}G_r(m) = \kappa_4 \left(m_r^{\max} \right)^{-k} I^{k+1},$$

where $\kappa_4 \equiv k\mathrm{e}^{-(k+1)} \int_0^1 (z^{-1} - z) \left(z\mathrm{e}^z \right)^k \mathrm{e}^z \mathrm{d}z > 0$ is a constant term which solely depends on the shape parameter k. Using the expressions for κ_1 and κ_2, one can verify that $\kappa_4 = \kappa_1 + \kappa_2$.

Using the expressions (A-5)–(A-7) and the results in **A.3.1–A.3.4** yields, after some more tedious but standard algebra, the expressions (12)–(14) given in the main text.

A.4. Other equilibrium expressions

In this appendix, we derive additional expressions that are required to characterize the equilibrium and to quantify the consequences of changes in trade costs.

A.4.1. The mass of varieties consumed. Using N_r^c as defined in (8), the export cutoff and the mass of entrants as given by (7) and (15), and making use of the Pareto distribution, we obtain:

$$N_r^c = \frac{\kappa_2}{\kappa_1 + \kappa_2} (m_r^d)^k \sum_s \frac{L_s}{F_s (m_s^{\max})^k} \left(\frac{\tau_{rr}}{\tau_{sr}} \frac{w_r}{w_s} \right)^k = \frac{\alpha}{\kappa_1 + \kappa_2} \frac{(m_r^d)^k}{\tau_{rr}} \sum_s L_s \tau_{rr} \left(\frac{\tau_{rr}}{\tau_{sr}} \frac{w_r}{w_s} \right)^k \frac{\kappa_2}{\alpha F_s (m_s^{\max})^k}.$$

Using the definition of $\mu_s^{\max}$, and noting that the summation in the foregoing expression appears in the equilibrium relationship (16), we can then express the mass of varieties consumed in region r as given in (17).

A.4.2 The (expenditure share) weighted average markup. Plugging (A-3) and (A-4) into the definition (18), the (expenditure share) weighted average markup in the multi-region case can be rewritten as

$$\overline{\Lambda}_r^c = \frac{1}{\alpha E_r \sum_s N_s^E G_s(m_{sr}^x)} \sum_s N_s^E \tau_{sr} w_s \int_0^{m_{sr}^x} m \left(W^{-2} - W^{-1} \right) \mathrm{d}G_s(m),$$

where the argument $\mathrm{e}m/m_{sr}^x$ of the Lambert W function is suppressed to alleviate notation. As shown in Appendix A.3.3, the integral term in the above expression is given by $\kappa_3 (m_s^{\max})^{-k} (m_{sr}^x)^{k+1} = \kappa_3 G_s(m_{sr}^x) m_{sr}^x$. Using this together with (7) and $E_r = w_r$ yields the expression in (18).

A.4.3. Indirect utility. To derive the indirect utility, we first compute the (unweighted) average price across all varieties sold in each market. Multiplying both sides of (6) by $p_{rs}(i)$, integrating over Ω_{rs}, and summing the resulting expressions across r, we obtain:

$$\overline{p}_s \equiv \frac{1}{N_s^c}\sum_r \int_{\Omega_{rs}} p_{rs}(j)\mathrm{d}j = \frac{1}{N_s^c}\sum_r \tau_{rs} w_r \int_{\Omega_{rs}} m_r(j)\mathrm{d}j + \frac{\alpha E_s}{N_s^c},$$

where the first term is the average of marginal delivered costs. Under the Pareto distribution, $\int_{\Omega_{sr}} m_s(j)\mathrm{d}j = N_s^E \int_0^{m_{sr}^x} m\mathrm{d}G_s(m) = [k/(k+1)] m_{sr}^x N_s^E G_s(m_{sr}^x)$. Hence, the (unweighted) average price for region r can be rewritten as follows

$$\overline{p}_r = \frac{1}{N_r^c}\sum_s \tau_{sr} w_s \left(\frac{k}{k+1}\right) m_{sr}^x N_s^E G_s(m_{sr}^x) + \frac{\alpha E_r}{N_r^c} = \left(\frac{k}{k+1}\right) p_r^d + \frac{\alpha E_r}{N_r^c}, \qquad \text{(A-8)}$$

where we have used (8) and $p_r^d = \tau_{sr} w_s m_{sr}^x$. Plugging (A-8) into (4) and using (7), the indirect utility is then given by

$$U_r = \frac{N_r^c}{k+1} - \frac{\alpha}{\tau_{rr} m_r^d}, \qquad \text{(A-9)}$$

which, together with (17) and (18), yields (20).