THE ECONOMIC JOURNAL

The Economic Journal, **130** (*May*), 880–910 DOI: 10.1093/ej/ueaa034
Advance Access Publication Date: 2 April 2020

GRAVITY WITHOUT APOLOGY: THE SCIENCE OF ELASTICITIES, DISTANCE AND TRADE*

Céline Carrère, Monika Mrázová and J. Peter Neary

Gravity as both fact and theory is one of the great success stories of recent research on international trade, and has featured prominently in the policy debate over Brexit. We first review the facts, noting the overwhelming evidence that trade tends to fall with distance. We then introduce some expository tools for understanding constant-elasticity-of-substitution theories of gravity as a simple general-equilibrium system. Next, we point out some anomalies with the theory: mounting evidence against constant trade elasticities, and implausible predictions for bilateral trade balances. Finally, we sketch an approach based on subconvex gravity as a promising direction to resolving them.

Today, we stand on the verge of an unprecedented ability to liberate global trade for the benefit of our whole planet with technological advances dissolving away the barriers of time and distance. It is potentially the beginning of what I might call 'post geography trading world' where we are much less restricted in having to find partners who are physically close to us.

—Liam Fox (2016)

Recognition of the importance of gravity in international trade is one of the great successes of modern economics. To adapt a comment made about evolution by the late Harvard palaeontologist Stephen Jay Gould (1981), gravity in trade is both *fact* and *theory*. Countless empirical studies have shown a significantly negative effect of distance on trade volumes; and much theoretical work has shown that this pattern is consistent with almost all the major approaches to the theory of international trade, in the process opening the door to quantitative studies of the effects of trade barriers on trade flows and welfare. However, these relatively recent developments are not widely appreciated by economists who are not trade specialists. As for the general public, there is some

* Corresponding author: J. Peter Neary, Department of Economics, University of Oxford, Manor Road, Oxford OX1 3UQ, UK. Email: peter.neary@economics.ox.ac.uk

This paper was received on 16 August 2019 and accepted on 9 March 2020. The Editor was Estelle Cantillon.

The data not subject to exemption and the codes for this paper are available on the Journal website. They were checked for their ability to replicate the results presented in the paper. The authors requested an exemption on parts of the data on the grounds that access to these data is restricted but nevertheless provided access to these data to the journal for the sole purpose of replicating the results based on their codes.

This paper was presented as the Past President's Address at the Annual Conference of the Royal Economic Society in Warwick, 15 April 2019, and also at the 2019 CESifo Area Conference on the Global Economy in Munich, at the 7th Conference on Trade and Technology in Nankai University (Tianjin), at DEC25 in Dubrovnik, at the 3rd International Conference on Economic Research in Alanya, Turkey, and at seminars in BJUT (Beijing), Oxford, Pavia, Princeton, QMUL, and UCD Dublin. For helpful comments and discussions, we are very grateful to participants on these occasions, and to many friends and colleagues, including Treb Allen, Maria Balgova, Kirill Borysyak, Swati Dhingra, Carsten Eckel, Andrew Elliott, Gene Grossman, Stefanie Haller, Udo Kreickemeier, Eduardo Morales, Philip Neary, Marcelo Olarreaga, Steve Redding, Zuzanna Studnicka and Frank Windmeijer. We are also grateful to our editor, Estelle Cantillon, and an anonymous referee, for very helpful suggestions. Finally, special thanks are owed in particular to Jim Anderson, for stimulating exchanges on gravity over many years. Some of the research leading to the paper was supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013), ERC grant agreement no. 295669.

awareness of the role of gravity in trade.[1] But 'anti-gravity' continues to have popular appeal: witness the success of books such as *The Death of Distance* by Frances Cairncross (1997) and *The World Is Flat* by Thomas Friedman (2005), both embodying sentiments eloquently summarised in the opening quotation from Liam Fox, MP, some months after he became United Kingdom (UK) Minister for International Trade.

In this paper, we seek to introduce this literature for the benefit of non-specialists, and to suggest some directions it might profitably take for the benefit of insiders. We first review the evidence for gravity, illustrating in a novel way how it shapes the spatial pattern of UK exports. We then present the archetypal model of 'structural gravity', and introduce some new ways of understanding how it works as a simple general-equilibrium system. Next we note some counterfactual implications of the assumption of constant elasticity of substitution (CES) preferences that underlies almost all general-equilibrium gravity models. Finally, we sketch some alternative specifications of gravity models. Throughout, we note the relevance of gravity and trade to ongoing debates in the UK on the likely effects of 'Brexit': the process of the UK exiting the European Union (EU).[2]

Of course, Brexit is about much more than economics. Just how much more is suggested by an anonymous quote from a senior member of the UK's ruling Conservative Party:

> *I don't think we'll be poorer out [of the EU], but if you told me my family would have to eat grass I'd still have voted to leave.*
>
> (anon.; quoted by Robert Shrimsley, *Financial Times*, 14 Dec. 2018)

It is easy to mock this position. It is not clear if the family were informed. And it is very clear that the electorate were not: the 2016 UK referendum campaign featured an iconic red Brexit bus sporting the slogan 'We send the EU £350 million a week; let's fund our NHS [National Health Service] instead'; there are no reports of a bus proclaiming 'We may have to eat grass but we will be free'. Though perhaps there is too much mockery around these days, on both sides of the highly polarised Brexit debate. Perhaps it is kinder to take the second half of the quote as merely a rhetorical device, a passionate endorsement of sincere, strongly-held views on the desirability of cutting links between the UK and the EU in order to restore Britain's sovereignty, mirroring the sincere, strongly-held views of those who favour remaining in the EU on liberal internationalist grounds. By contrast, the first half of the quote makes a modest claim about the economic effects of Brexit. As we will show, the scientific evidence suggests overwhelmingly that this claim is false, though only modestly so.

In the rest of the paper, we focus on the economics of Brexit, and in particular its implications for international trade. There have been many studies of the trade effects of Brexit, using the gravity model. Examples include Dhingra *et al.* (2017), Sampson (2017), Brakman *et al.* (2018) and Mayer *et al.* (2019). There are also many other important aspects of Brexit on which economists

[1] A gravity equation featured on the front page of the *Financial Times* on 19 April 2016 in the context of discussions preceding the Brexit referendum, on which more below.

[2] The UK, officially the United Kingdom of Great Britain and Northern Ireland, is often referred to as just Britain. It joined the European Economic Community (EEC), the predecessor of the EU, on 1 January 1973. In a referendum held on 23 June 2016, the UK voted to leave the EU by 51.89% to 48.11%. Following much debate, including two general elections, the UK ceased to be a member of the EU in law on 31 January 2020, entering a transition period which the UK Parliament has legislated will end on 31 December 2020. During that time the UK will continue to be a full member of the EU Customs Union and Single Market while the future relationship between the UK and the EU is negotiated. At the time of writing (3 March 2020), there is no certainty about where the outcome of the Brexit process will fall on a spectrum between 'hard' (a 'no agreement' exit with the EU and UK trading on WTO terms), and 'soft' (including continued regulatory alignment and free movement of labour).

have already written, and no doubt there will be many more.[3] However, the predicted effects on trade and real incomes have been the focus of most popular discussion of the economics of Brexit. So it seems appropriate to concentrate on them in order to explain why academic economists are almost unanimous in warning of the economic costs of Brexit, and to document the role that gravity has played in moulding that professional consensus.[4] The purpose of this paper is not to add another calibration of these costs, but rather to explore why the existing ones give the results they do.

The overwhelming conclusions of the gravity studies cited above might be called the *three iron laws of the economics of Brexit*. To be clear, these conclusions refer to trade in goods only: services trade also follows gravity, but the available data are not as comprehensive as for merchandise trade. Moreover, these conclusions follow from studies using static micro-founded general-equilibrium models, so they ignore transitional problems; for example, they have little to say about the hard-to-forecast costs of a 'no-deal' Brexit. They also ignore macroeconomic policy responses: in what follows any change in real income is equal to the change in real wages; the models are silent on whether these would be effected through a deflationary fall in nominal wages, or through an accommodative monetary policy coupled with a depreciation of sterling, as happened in the wake of the 2016 referendum.

What then are the three iron laws? First, *the only good Brexit is a dead Brexit*: all realistic Brexit scenarios imply lower UK GDP than remaining in the EU. Second, *the harder the Brexit, the higher the economic costs*: for example, a 'hard' Brexit in which the UK completely withdraws from the EU Single Market and Customs Union will have higher costs than a 'soft' one that entails some continued participation in these deep trade agreements. Third, *even a hard Brexit will not have 'very' large costs*: the orders of magnitude from all the studies suggest a permanent but once-off loss of the order of 2% of GDP for a soft Brexit, and 6% or more of GDP for a hard one. These are significant economic costs, unprecedented for a deliberate policy choice by a peacetime government; to put them in context, at the height of the financial crisis in 2009 UK GDP fell by 5.0%, and in 2016–17 the UK spent 7.26% of its GDP on the NHS.[5] But this is not Armageddon, or a wartime scenario. Passionate leavers who value sovereignty above all else should be prepared for a major reduction in UK GDP relative to what it would otherwise have been, but, conditional on an orderly exit, need have no fears of a grass-only menu.

The plan of the paper follows the outline given above. Section 1 sketches the facts of gravity from the perspective of UK exports; like this introduction, it is intended to be accessible to non-specialists. Section 2 explains the structural gravity model and shows how it can be interpreted as a simple general-equilibrium system. Section 3 considers some anomalous implications of CES demands and CES gravity. Section 4 outlines an approach that may help overcome them. Finally, Section 5 summarises the paper, while the Appendix gives details on technical derivations.

[3] A short list would include Davies and Studnicka (2018) on the stock-market response to the unanticipated result of the Brexit referendum; McGrattan and Waddle (2018) on the impact of Brexit on foreign investment in a neoclassical growth model; Alabrese *et al.* (2019) on the determinants of voting patterns in the Brexit referendum; and O'Rourke (2019) on the historical context.

[4] Readers of some UK newspapers may be under the impression that the economics profession is deeply split on the issue. (It is tempting to draw parallels with other debates, such as on climate change, evolution or vaccination, where an overwhelming scientific consensus is sometimes depicted as only one view among many equally valid ones.) However, only a small minority of academic economists is in favour of Brexit on economic grounds.

[5] https://www.parliament.uk/documents/commons/lib/research/key_issues/Full-doc.pdf, p. 28, and https://www.nuffieldtrust.org.uk/chart/nhs-spending-as-a-percentage-of-gdp-1950-2020. Note that the gravity-based estimates of the economic costs of Brexit to the UK that we quote take account of savings on direct contributions to the EU budget.

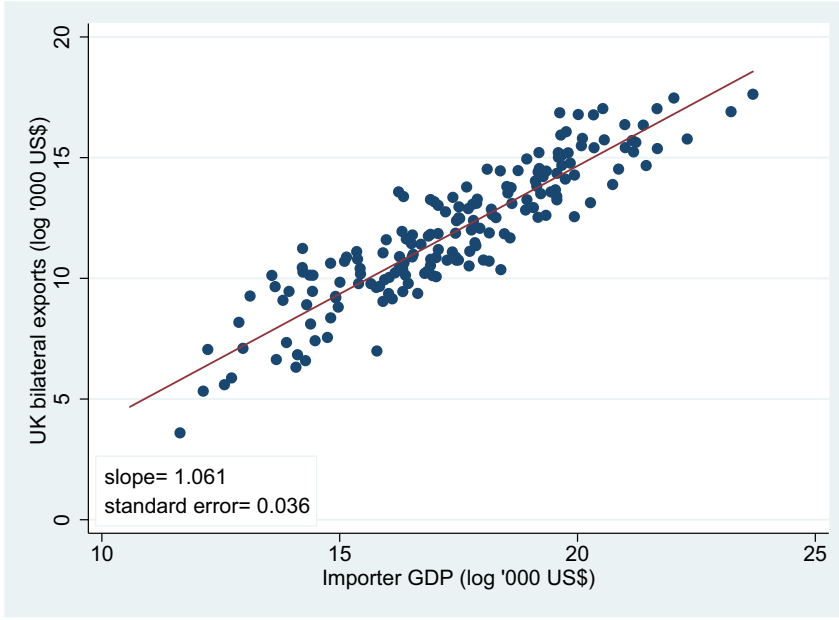


Fig. 1. *UK Bilateral Exports and Importer GDP, 2017.*
Notes: The data on trade flows come from the CEPII BACI database, and GDP data come from the World Bank's World Development Indicators and collected in the CEPII Gravity database.
Source. Authors' calculations.

1. Gravity as Fact

We begin this section by using some simple charts to illustrate the robustness of the gravity effect, both for geographic distance and for other distance variables such as membership of a common trade agreement and former colonial ties. The data are for UK merchandise exports to 181 countries in 2017 (the latest full year for which comparable data are available).[6] It goes without saying that this is not intended as a serious econometric exercise, though the tendencies we will point out are in line with the findings of almost all large-scale studies.

Figure 1 plots UK exports against importer GDP, both in logs of current dollars. The positive relationship between the two is apparent, and confirmed by the simple regression line. The estimated slope coefficient is 1.061 with a standard error of 0.036: significantly different from zero but not from one. As we will see in the next section, most theoretical foundations of the gravity equation assume that this coefficient equals one. Hence, to allow a visual exploration of the effect of distance, it makes sense to impose a value of one, which allows us to focus on the ratio of UK exports to importer GDP.

[6] Figures 1 and 2 follow Head and Mayer (2014) who illustrate similar patterns for French exports. Of the 206 countries and territories with at least one positive bilateral export value in 2017 in the initial database (the CEPII BACI database: see Gaulier and Zignago, 2010), the UK is recorded as trading with 203. We also exclude from the sample a further 22 partner countries for which GDP data (taken from the World Bank's World Development Indicators and collected in the CEPII Gravity database: see Head and Mayer, 2014) are not available.

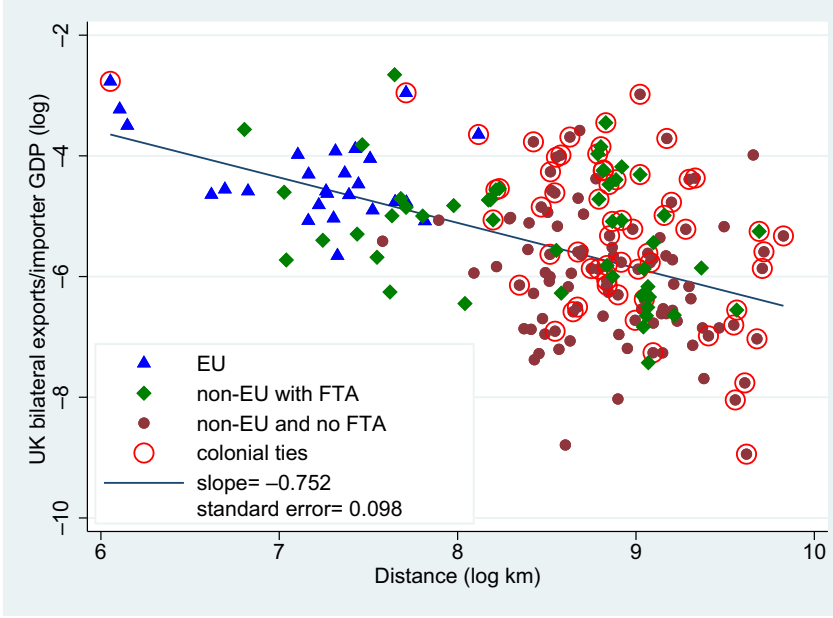


Fig. 2. *UK Bilateral Exports/Importer GDP and Distance, 2017.*
Notes: The data on trade flows come from the CEPII BACI database and data on other variables (GDP, distance, EU and FTA, colonial ties) come from the CEPII Gravity database. FTA variable refers to whether an FTA has been notified to the WTO as of 2016, 'colonial ties' to whether the UK has ever been in a colonial relationship with the importer country.
Source. Authors' calculations.

Figure 2 plots this ratio against bilateral distance, both once again in logs.[7] This time the simple regression line is downward-sloping. Its estimated slope coefficient is −0.752 with a standard error of 0.098: significantly different from zero. For each export market, the symbols indicate its trading relationship with the UK as of the end of 2016: a triangle if it is a member of the EU; a diamond if it has a free trade agreement (FTA) with the EU;[8] a circle otherwise; and a halo if it is a current or former UK colony. Figure 3 illustrates the same data as Figure 2 but this time with the size of each observation proportional to the share of UK exports to that country.

Figures 2 and 3 confirm that UK trade falls off with distance, when we control for the size of the importing country. Figure 3 also shows that the tendency to cluster around the best-fit line is even more pronounced for larger trading partners. Many of the extreme outliers in Figure 2 are barely visible when we scale by the share of exports as in Figure 3; while most of the largest export markets lie on or close to the best-fit line. (The best-fit line in Figure 3 is the same as that in Figure 2.)

[7] Distance is measured as a population-weighted average of distances between major cities. By this metric, Ireland is closer to the UK than either Belgium or the Netherlands, and all three are closer than France. See Mayer and Zignago (2011) for discussion.

[8] The EU has trade agreements with 76 countries in the data used in the figures, and new ones have been signed since then. The most recently concluded of these agreements, that with Japan, came into force on 1 February 2019. The UK benefits from these trade agreements as long as it remains an EU member, and is engaged in negotiations to roll them over post-Brexit. It is not clear if these will yield the full benefits of the current agreements.

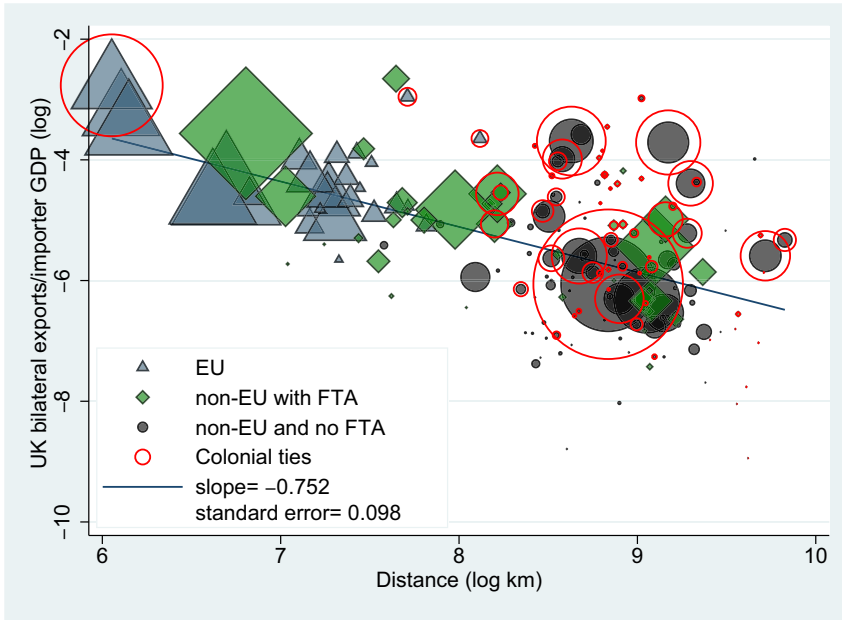


Fig. 3. *UK Bilateral Exports/Importer GDP and Distance, Scaled by Exports, 2017.*
Notes: Same data as in Figure 2. The weights are the share of UK exports to each of the 181 importing countries.
Source. Authors' calculations.

Is the relationship shown in these figures unique to the UK? The answer is definitely not: the same pattern can be seen in almost all empirical gravity studies. A comprehensive survey by Head and Mayer (2014) reviewed 159 papers that had estimated gravity equations. Focusing on papers using similar estimation methods to those we use in this paper, they found an average estimate of the distance elasticity of −1.10, with a standard deviation of 0.41 and a median of −1.14.

Moreover, the distance coefficient for goods trade has not fallen over time, contrary to suggestions in popular debates as discussed in the introduction. This has been called 'the mystery of the missing globalization', or 'the puzzling persistence of the distance effect' (Disdier and Head, 2008). However, it is not really a mystery, when we bear in mind that distance is relative in estimated gravity equations. Studies that use data on both domestic sales and exports to estimate a border dummy variable typically find that it has fallen over time. Thus international trade per se has become easier, but the relative attractiveness of nearby versus foreign markets has not changed much (see, for example, Anderson and Yotov, 2010; Yotov, 2012). Improvements in transport and communications technology have made it easier for UK firms to export to New Zealand, but also easier to export to Ireland.

We have focused so far on trade in goods only. Because standardised data on merchandise trade are much more widely available, the majority of gravity studies look only at this component of trade. However, it has been shown in many studies that distance also matters (though less so on average) for a whole range of international transactions. In rough order of distance coefficients that decrease in absolute value but remain significantly different from zero, negative effects of

distance have been found for: services trade (Kimura and Lee, 2006), foreign direct investment (Kleinert and Toubal, 2010; Keller and Yeaple, 2013), trade in equities (Portes and Rey, 2005), eBay transactions (Lendle *et al.*, 2016) and Google hits (Cowgill and Dorobantu, 2012).

It is not only geographical distance that matters in gravity equations. Distance in other senses also affects trade, with variables such as common language, common legal system, common colonial origins, membership of the same FTA invariably showing up as significant. Returning to Figures 2 and 3, some of these effects can be seen clearly for UK trade. In particular, recalling the tendency for the larger trading partners to cluster more closely around the best-fit line, it is noteworthy that most of the exceptions are former colonies.[9] Both figures show that the UK tends to export more to its former colonies than to other countries, relative to what geographical distance and the economic size of the importing country alone predicts. This is in line with an extensive literature which finds that former colonial ties tend to increase trade (see for example Head *et al.*, 2010). It is also relevant to the Brexit debate. A recurring theme in the arguments in favour of Brexit has been characterised, perhaps ironically, as 'Empire 2.0': the hope that new trade agreements with former UK colonies including the USA would more than compensate for the loss of preferential access to EU markets. But as Figure 3 shows, 'Empire 1.0' casts a long shadow: controlling for distance and importer GDP, the UK already trades much more with these countries (other than the USA) than it does on average. Almost all the countries with which there is a significant value of exports and which lie above the best-fit line are former colonies, from Australia and New Zealand at the far end of the globe, to the UAE, Hong Kong, Singapore and Malaysia in the Middle and Far East, to, perhaps most remarkably, all three former UK colonies that are EU members: Ireland, Malta and Cyprus.

What do these figures imply for Brexit? Gravity is not destiny. Yet it is hard not to look at Figure 3 without reflecting that the UK currently enjoys free and frictionless trade with the triangles, and preferential access to the diamonds; and without wondering at the wisdom of abandoning the first and risking the second in the hope of negotiating new trade agreements with the far-away circles. Of course, this argument is far from rigorous: there is no explicit counterfactual. For that we need a theory that is consistent with the data and that yields predictions of how changes in trade policy would affect trade patterns. We turn to this in the next section.

2. Gravity as Theory

> *[I] have explained the phenomena of the heavens and of our sea by the power of gravity, but have not yet assigned the cause of this power.*
>
> —Isaac Newton (1713)

> *The intent of this paper is to provide a theoretical explanation for the gravity equation applied to commodities.*
>
> —Jim Anderson (1979)

In later editions of his *Principia*, Isaac Newton conceded to his critics that his mathematical theory of gravity did not give a primitive explanation of the forces between bodies. Yet in the 1979 *American Economic Review*, Jim Anderson provided a micro-founded theoretical explanation for

[9] The only prominent exception with no colonial ties to the UK is Switzerland, though anecdotal evidence suggests that exports to it in 2017 were boosted above trend by flows of gold bullion, reflecting balance-of-payment adjustments rather than merchandise trade. See: 'Gold fingered for distorting Brexit Britain's trade balance', *Financial Times*, 24 February 2017; and 'How gold takes the shine off Britain's trade balance', *Sky News*, 18 April 2018.

the gravity equation of trade flows. The contrast between the two goes deeper than that. Newton gave an analytical solution for the force of gravity in the two-body problem only. Even today, while physicists can simulate the movements of planets and particles with extraordinary accuracy, there is no explicit expression for physical gravity in higher dimensions: the three- or n-body problem cannot be solved in closed form. Yet Jim Anderson (1979) and other economists since then have been able to provide closed-form gravity expressions for trade between any number of countries. Why has it proved easier to derive results of this kind in economics than in physics? The reason is simple: planets do not have CES preferences! Almost all the many theoretical rationales that have been provided for the gravity equation in international trade assume that consumers have CES preferences.

This assumption about preferences is very special. Yet it is a natural starting point for quantitative studies of trade; CES preferences are a standard benchmark for modelling consumer behaviour, they are widely used in many fields of economics other than trade, they are analytically very tractable, and they are easy to take to data, even if their predictions are not always fully confirmed, as we explore further in Section 3. The CES assumption also brings to centre stage the key feature of gravity models: that goods are imperfect substitutes. This avoids the tendency towards knife-edge specialisation and the prediction of dramatic shifts in production patterns in response to tiny price changes, that are implied by older trade models such as the textbook Ricardian model of comparative advantage. It also rationalises the data which show that, at every level of disaggregation, countries trade with more partners than is consistent with the hypothesis of perfect substitutability.

Moreover there is a compensating richness on the supply side. The gravity equation has been shown to be consistent with a wide range of canonical trade models, each with different assumptions about the structure of production: pure exchange, monopolistic competition with homogeneous or heterogeneous firms, and comparative advantage.[10] As highlighted in the synthesis of Arkolakis *et al.* (2012), all these frameworks yield the same 'structural gravity' model, and the same parsimonious expression for the gains from trade.

To fix ideas, we follow Anderson (1979) and focus on the simplest version of structural gravity, which assumes an exchange economy. In Subsection 2.1 we introduce notation and show how CES demands combined with market-clearing yield the structural gravity equations; this subsection can safely be skimmed by trade specialists. In Subsection 2.2 we present a new pedagogic approach to understanding the structural gravity model as a simple general-equilibrium system, while in Subsection 2.3 we show its usefulness with an application to Brexit.

2.1. *From CES Demands to Structural Gravity*

Consider a world of n countries. A typical country k is populated by a representative consumer who is endowed with a fixed supply of a unique good, denoted by Q_k. The domestic price of the good is p_k, so the value of national income is $Y_k = p_k Q_k$. This is not necessarily equal to the value of national expenditure E_k. However, gravity models do not attempt to explain divergences between national income and national expenditure, so one is assumed to be an exogenous multiple of the other: $E_k = \kappa_k Y_k$. Since the domestic good is not produced, we can identify the domestic

[10] Gravity equations have been derived for the Armington (1969) model of pure exchange by Anderson (1979) and Anderson and van Wincoop (2003); for the Krugman (1980) model of monopolistic competition by Bergstrand (1985) and Helpman (1987); for the Melitz (2003) model of heterogeneous firms by Chaney (2008); and for a multi-country Ricardian model by Eaton and Kortum (2002).

nominal wage rate with the domestic price: $w_k = p_k$; this assumption is relaxed in gravity models that allow for monopolistic competition.

Turning to consumer behaviour, we assume that all countries have the same CES preferences:[11]

$$U_k = \left(\overline{U}_k\right)^{\frac{\sigma-1}{\sigma}} = \sum_{j=1}^{n} \left(\eta_j x_{jk}\right)^{\frac{\sigma-1}{\sigma}}. \tag{1}$$

Provided trade costs are less than infinite, the representative consumer in country k derives utility from consuming some of all the goods in the world. We write the subscripts for exporting and importing countries in the same order as the direction of trade throughout, so x_{jk} denotes the quantity of exports from j to k. Utility depends on $n + 1$ parameters: σ denotes the elasticity of substitution between every pair of goods, while each η_j denotes a preference parameter for good j which depends only on the origin of that good, since all importing countries have the same tastes.

Utility (1) is maximised subject to country k's budget constraint:

$$\sum_{j=1}^{n} p_{jk} x_{jk} \leq E_k. \tag{2}$$

The total value of consumption, including both imports, $x_{jk}, j \neq k$, and consumption of the home good, x_{kk}, cannot exceed domestic expenditure E_k. Crucially, p_{jk} is the delivered price of j's export in k, which equals the origin or 'factory-gate' price p_j times an 'iceberg' trade cost, $t_{jk} \geq 1$: $p_{jk} = p_j t_{jk}$. Iceberg costs imply that t_{jk} units of country j's good must be shipped from j for one unit to arrive in country k; the other $t_{jk} - 1$ units 'melt' in transit.

Maximising (1) subject to (2) leads to the demand functions for country j's good in country k. Because reliable bilateral trade data are available only in value form, we write the demand function in terms of V_{jk}, the value of exports from j to k, which equals the price p_{jk} times the quantity x_{jk} of exports.

$$V_{jk} = p_{jk} x_{jk} = \left(\frac{\eta_j^{-1} p_{jk}}{P_k}\right)^{1-\sigma} E_k. \tag{3}$$

The determinants of demand on the right-hand side of (3) are standard for a CES function. Sales are proportional to total expenditure in the importing country, E_k, reflecting the fact that CES preferences are homothetic. Conditional on expenditure, demand for good j depends on the preference parameter η_j and on its price p_{jk} relative to the cost of living in the importing country, P_k. As for P_k, it takes the standard form of a CES price index:

$$P_k = \left(\sum_{h=1}^{n} (\eta_j^{-1} p_{hk})^{1-\sigma}\right)^{\frac{1}{1-\sigma}}. \tag{4}$$

Finally, the impact of relative prices on demand depends on the elasticity of substitution σ, which is also the elasticity of demand.

[11] We write utility in two different ways: U_k shows that the function is a special case of additively separable preferences, which we discuss further in Section 4 below; while $\overline{U}_k$ is a more familiar way of writing a CES function. These two ways of writing utility have identical implications for behaviour, since U_k is a monotonically increasing transformation of $\overline{U}_k$, and preferences are ordinal.

To go from CES demands to structural gravity, we add the conditions for goods-market equilibrium. For a typical country j, let V_j denote the total value of its sales to all countries, both exports, $V_{jk}, j \neq k$, and sales to the home consumer, V_{jj}. In equilibrium this must equal the value of its GDP, Y_j:

$$V_j \equiv \sum_{k=1}^{n} V_{jk} = Y_j. \tag{5}$$

Combining this with the demand function (3), we see that a term in the taste parameter and the exporting country's factory-gate price factors out:

$$Y_j = \sum_{k=1}^{n} V_{jk} = (\eta_j^{-1} p_j)^{1-\sigma} \sum_{k=1}^{n} \left(\frac{t_{jk}}{P_k}\right)^{1-\sigma} E_k. \tag{6}$$

Using this to eliminate the term $(\eta_j^{-1} p_j)^{1-\sigma}$ from V_{jk} in (3) and P_k in (4) yields the structural gravity equation:

$$V_{jk} = \underbrace{\left(\frac{t_{jk}}{\Pi_j P_k}\right)^{1-\sigma}}_{(1)} \underbrace{\frac{Y_j E_k}{Y_W}}_{(2)}, \tag{7}$$

where:

$$(\Pi_j)^{1-\sigma} = \sum_{h=1}^{n} \left(\frac{t_{jh}}{P_h}\right)^{1-\sigma} \frac{E_h}{Y_W}, \quad (P_k)^{1-\sigma} = \sum_{h=1}^{n} \left(\frac{t_{hk}}{\Pi_h}\right)^{1-\sigma} \frac{Y_h}{Y_W}. \tag{8}$$

To understand the implications of (7), consider the two numbered composite terms in reverse order. The second term represents the level of free and frictionless trade predicted by the model: if there are no trade costs (so all the t_{jk} equal one), then the value of exports from j to k equals the product of exporter GDP Y_j and importer expenditure E_k deflated by world income Y_W.[12] Putting this differently, when prices are the same everywhere, each country k spends a proportion of its total expenditure on imports from every other country j that is equal to the exporter country's share in world GDP: $V_{jk}/E_k = Y_j/Y_W$. The first term in (8) shows how trade costs modify this: exports from j to k are lower the greater is the elasticity of import demand, $\sigma - 1$, and the higher is the bilateral trade cost t_{jk} relative to the product of two indices of the average trade costs faced by the exporter and the importer respectively, Π_j and P_k.[13] Anderson and van Wincoop (2003) called these outward and inward 'multilateral resistance' respectively, and the fact that they are dual to one another underlines the elegance of the structural gravity system.

Where do we go from (7) and (8)? The first step is estimation. As Anderson and van Wincoop (2003) showed, this can be done structurally, using non-linear methods. However, this approach is seldom used since it requires that we take a stand on the supply side of the model summarised by the Y_j terms in (7) and (8). Hence the irony that the structural gravity model is rarely estimated structurally. In practice a different approach is taken. Irrespective of which model of

[12] Total output and expenditure must be equal for the world as a whole, so $Y_W = \Sigma_j Y_j = \Sigma_k E_k$.

[13] With zero trade costs, the terms Π_j and P_k do not in general reduce to one: after all, P_k continues to represent the true cost of living. However, with zero trade costs their product must equal one. As is easy to check from (8), when all t_{jk} equal one, Π_j and P_k are independent of j and k (as they must be, since the producer and consumer prices of each good are the same in all countries); and one is the reciprocal of the other, so $\Pi_j P_k = \Pi P = 1$. It is legitimate to set Π_j and P_k equal to one in the absence of trade costs by choice of numeraire. See Subsections 2.3 and 3.2 for further discussion of the choice of numeraire.

the production side of the economy is assumed, we can take logs of (7), add an error term u_{jk}, and write the result as:

$$\log V_{jk} = F_j + F_k + \mu \log t_{jk} + u_{jk}. \tag{9}$$

Thus the value of exports from j to k takes a simple log-linear form, depending on importer and exporter fixed effects, F_j and F_k, and on a term specific to the 'dyad' $\{j, k\}$. In practical applications, the latter term can be decomposed into a vector of trade cost measures such as geographical distance, contiguity, common language, colonial ties, membership of an FTA, etc. The coefficient of this term, μ, is (minus) the elasticity of trade, and its relationship to underlying structural parameters depends on the assumptions made about the supply side of the model.

Estimating (9) on the data used in Section 1, we obtain a distance coefficient of −1.452, with a clustered standard error of 0.019. We control for the full set of importing and exporting fixed effects and, as variables specific to the 'dyad' $\{j, k\}$, we include contiguity, common language, colonial ties, membership of a common trade agreement and/or currency area. All of these are available in the CEPII Gravity database (see Head and Mayer, 2014).[14]

Given estimates of (9), the next step usually taken is simulation. In particular, by considering changes to the generalised distance term t_{jk}, it is possible to simulate the effects of detailed changes in trade policy. This is the approach taken in the studies of the effects of Brexit mentioned in the introduction, for example. Applications of this kind are becoming increasingly common.[15] Given our current state of knowledge, they provide the best available quantitative answer to questions such as 'How will Brexit affect UK trade and GDP?'. However, given the complexity of the multilateral trade linkages considered, they run the risk of seeming like 'black boxes'. In the remainder of this section, we take a different, complementary approach. We ask what can be said about the qualitative properties of the model. Such a theoretical analysis yields few results when carried out in terms of levels. It is more insightful when done in terms of local changes, following the standard approach of comparative statics. This has the further advantage that the results are robust to relaxing the assumption of CES demands, an issue to which we return in Section 4.

2.2. *The Structure of Simple Structural Gravity Models*

Comparative statics for structural gravity have been explored by a number of authors, including Alvarez and Lucas (2007), Dekle *et al.* (2008) and Allen *et al.* (2020). The approach adopted here also has similarities to the framework for aggregating from micro to macro in a multisectoral economy developed by Baqaee and Farhi (2017). It is most closely related to the classic exposition of the two-sector Heckscher–Ohlin model in Jones (1965), and its multisectoral extension in Jones and Scheinkman (1977).[16]

[14] The complete trade matrix has 206 countries, and so 42,230 observations, whereas this OLS benchmark has 23,251 observations as it only includes strictly positive trade flows. If we assume that all the non-reported flows correspond to zero trade flows rather than missing values, and if we estimate the same gravity equation but for the complete trade matrix using Poisson pseudo maximum likelihood (PPML) as in Santos Silva and Tenreyro (2006) and Fally (2015), we obtain an estimated distance coefficient of −0.735 with a clustered standard error of 0.069. (We use the procedure developed by Larch *et al.*, 2019 specifically designed for the case of many fixed effects required by structural gravity models.)

[15] For an overview of the issues that arise in implementing them, see Yotov *et al.* (2016).

[16] Similar multisectoral results, using different notation, were obtained by Diewert and Woodland (1977).

Following Jones (1965) and Jones and Scheinkman (1977), we define the shares of bilateral trade in exporter GDP and importer expenditure as follows:

$$\lambda_{jk} = \frac{V_{jk}}{Y_j} = \frac{t_{jk}x_{jk}}{Q_j}, \quad \theta_{jk} = \frac{V_{jk}}{E_k}. \tag{10}$$

Here λ_{jk} denotes a 'real' share: the proportion of country j's output shipped to each market k; while θ_{jk} denotes a 'value' share: the proportion of country k's expenditure sourced from each supplying country j. With balanced trade ($E_j = Y_j, \forall j$), these shares are related to each other in exactly the same way as the analogous shares are in Jones (1965): $\lambda_{jk}\theta_j = \theta_{jk}\theta_k$, where $\theta_j \equiv Y_j/Y_W$ is country j's share in world GDP. An important special case is where country j is relatively 'small', so its share in world GDP is close to zero, $\theta_j \approx 0$, and all other countries are 'large'. In this case:

$$\lambda_{kj} = \frac{\theta_{kj}}{\theta_k}\theta_j \approx 0 \quad \text{and} \quad \theta_{jk} = \frac{\lambda_{jk}}{\theta_k}\theta_j \approx 0, \quad \forall k \neq j. \tag{11}$$

So, every other country exports only an infinitesimal proportion of its output to j, and devotes only an infinitesimal proportion of its expenditure to imports from j.

Now, express changes in terms of these shares, using 'hats' to denote local proportional changes, $\widehat{x} \equiv d\log x$.[17] It turns out to be most insightful to do this using the primitive equations of the model for consumer equilibrium and market-clearing, (4) and (5) respectively, rather than the structural gravity equations themselves, (7) and (8). First, we can express the requirement that the market for each good must clear as applying at the margin. Totally differentiating equation (5) and cancelling $\widehat{p}_j$ which appears on both sides, the change in each country's GDP must equal a λ-weighted average of the changes in its sales to each country (including home sales as well as exports):

$$\begin{aligned} Y_j = V_j \equiv \sum_{k=1}^{n} V_{jk} \quad &\Rightarrow \quad \widehat{Y}_j = \widehat{V}_j = \sum_{k=1}^{n} \lambda_{jk}\widehat{V}_{jk} \quad \Rightarrow \\ 0 = \sum_{k=1}^{n} \lambda_{jk}(\widehat{t}_{jk} + \widehat{x}_{jk}), \qquad & j = 1, \ldots, n. \end{aligned} \tag{12}$$

The middle expression in (12) holds for all versions of the structural gravity model. The final one specialises to the Armington version, where GDP is just the home price times the exogenously given stock of output, Q_j: $Y_j = p_j Q_j$. In this form it says that a λ-weighted average of changes in production for each market must sum to zero, keeping in mind that production includes a trade cost component. Second, from (4), the change in the price index in each country must equal a θ-weighted average of the changes in retail prices there, both of the home-produced good and of imports:

$$\widehat{P}_k = \sum_{j=1}^{n} \theta_{jk}\widehat{p}_{jk}, \qquad k = 1, \ldots, n. \tag{13}$$

(This holds for any true cost-of living index, not just the CES.)

Equations (12) and (13) are the basic building blocks of the general-equilibrium system. Next, we add the demand functions, from (3), totally differentiated to give changes in demand at the margin:

[17] A potential source of confusion with this notation is that these changes are sometimes partial and sometimes total, depending on what is held constant. Hopefully, the correct interpretation will be clear from the context.

$$\widehat{x}_{jk} = -\sigma \widehat{p}_{jk} + (\sigma - 1)\widehat{P}_k + \widehat{E}_k. \tag{14}$$

Using (13), we can write the own and cross-price derivatives of demand as follows:

$$\frac{\partial \log x_{jk}}{\partial \log p_{jk}} = -(\sigma(1-\theta_{jk}) + \theta_{jk}), \qquad \left.\frac{\partial \log x_{jk}}{\partial \log p_{hk}}\right|_{h \neq j} = (\sigma - 1)\theta_{hk}. \tag{15}$$

A central implication of CES preferences is that these derivatives exhibit the gross substitutes property: $-\frac{\partial \log x_{jk}}{\partial \log p_{jk}} > \frac{\partial \log x_{jk}}{\partial \log p_{hk}} > 0$. (See Alvarez and Lucas, 2007 in this context.) This property guarantees that the model is well behaved with intuitive properties. Finally, to complete the general-equilibrium system, we add to (12), (13) and (14) the three definitional equations already discussed in Subsection 2.1, re-expressed in terms of proportional changes.[18] First is the link between prices and trade costs:

$$p_{jk} = p_j t_{jk} \quad \Rightarrow \quad \widehat{p}_{jk} = \widehat{p}_j + \widehat{t}_{jk}. \tag{16}$$

Second is the link between national expenditure and GDP:

$$E_j = \kappa_j Y_j \quad \Rightarrow \quad \widehat{E}_j = \widehat{Y}_j. \tag{17}$$

As already stated, the parameter κ_j equals one when aggregate trade is balanced, and is assumed to be exogenous: the model does not attempt to explain macroeconomic imbalances. Third is the specification of the supply side, which takes the simple Armington form:

$$Y_j = p_j Q_j \quad \Rightarrow \quad \widehat{Y}_j = \widehat{p}_j. \tag{18}$$

Since the price of each country's output uniquely determines its income, we can also identify it with the wage: $w_j = p_j$ implying that $\widehat{w}_j = \widehat{p}_j$. Naturally, the models with more sophisticated supply sides discussed earlier have richer mechanisms for wage determination.

2.3. *An Application to Brexit*

To illustrate in a stylised way how the general model can be used to understand the trade effects of Brexit, we specialise to three countries that we label *A*, *B* and *E*; '*B*' for 'Britain', '*E*' for 'Europe' and '*A*' for the rest of the world, or, for concreteness, 'America'. This reduces the dimensionality of the general *n*-country model to three. Moreover, one market-clearing condition is redundant by Walras's law, and one country's domestic price can be set equal to unity by choice of numeraire, so we can explore the model's properties in two dimensions. To fix ideas, we concentrate on countries *B* and *E*. Hence we omit the market-clearing condition for country *A*'s good, and select the price of its good as the numeraire ($p_A = 1$), so all nominal variables are measured relative to prices in *A*. (This is a true numeraire or measuring rod: like the choice between Fahrenheit and Celsius, it does not affect the model's implications, though for framing reasons it may make us feel differently about them.) This allows us to understand the determination of equilibrium by illustrating in a simple diagram how the market-clearing conditions for outputs of *B* and *E* determine equilibrium wages, which are equivalent to prices: $w_B = p_B$ and $w_E = p_E$.

[18] With the further addition of the decomposition of changes in trade values into price and quantity components, $\widehat{V}_{jk} = \widehat{p}_{jk} + \widehat{x}_{jk}$, this gives seven equations in all, which determine the seven endogenous variables, $\widehat{Y}_j$, $\widehat{V}_{jk}$, $\widehat{x}_{jk}$, $\widehat{p}_{jk}$, $\widehat{E}_k$, $\widehat{P}_k$ and $\widehat{p}_j$.

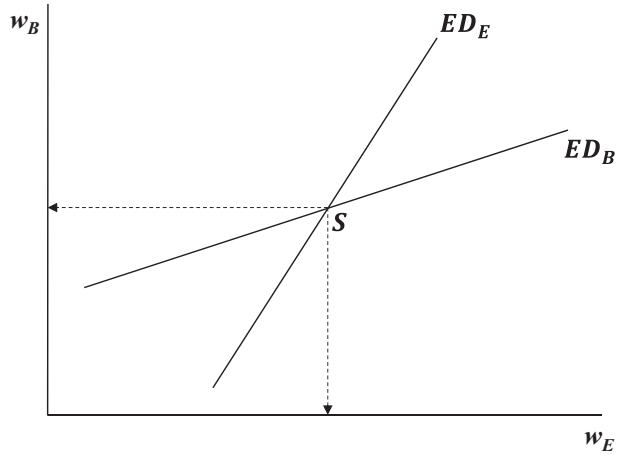


Fig. 4. *Determination of Equilibrium Wages in Countries B and E.*

Figure 4 illustrates in $\{w_E, w_B\}$ space. The curve labelled ED_B indicates the combinations of wages in B and E consistent with market-clearing for country B's output: total demand from all countries V_B equals total supply Y_B, so excess demand ED_B is zero. Its properties can be deduced from (12) and (14).[19] Starting at any point along the ED_B locus, a higher wage in B leads to excess supply, a higher wage in E leads to excess demand, while a uniform increase in both wages leads, from gross substitutability, to excess supply. Hence the ED_B locus must be upward-sloping but with a slope less than 45° as shown. A symmetric argument shows that the market-clearing locus for country E's output must also be upward-sloping, but with a slope greater than 45° as shown by the ED_E locus. The intersection of the two loci therefore determines the unique equilibrium wages w_B and w_E. Of course, the two loci need not have the same slope; for example, if B is a relatively small economy, then the ED_E locus is vertical.

Having illustrated the determination of equilibrium, we can now explore how it responds to shocks. This shows how easy it is to explore alternative trade policy scenarios in the gravity model.

As a first step, we decompose trade costs into those that are 'natural', denoted δ_{jk}, and those that are policy-induced, denoted τ_{jk}:[20]

$$t_{jk} = \delta_{jk}\tau_{jk}. \tag{19}$$

The former includes distance of course, as well as historically given factors that encourage or discourage trade, such as colonial ties or common language. For simplicity, we assume that all trade costs are bilaterally symmetric, and that policy costs yield no revenue. The first of these assumptions is perhaps less innocent than it seems: the distance from Britain to Europe is the same in each direction, but the costs of transporting goods need not be, if for example the mix of goods shipped in the two directions is very different. By contrast, the second assumption is perhaps more innocent than it seems: the majority of policy-induced trade barriers in the modern world economy are not tariffs but rather standards and technical barriers, that are not primarily,

[19] For detailed derivations, see the Appendix.
[20] See Maggi *et al.* (2018).

Table 1. *Alternative Trade Policy Scenarios.*

Scenario	δ_{BE}	τ_{BE}	δ_{BA}	τ_{BA}
Status quo	low	low	high	high
'Cake and eat'	low	low	high	low
'Global Britain'	low	high	high	low

Notes: (1) All trade costs are assumed to be bilaterally symmetric. (2) Revenue from policy costs is ignored.

if at all, revenue-raising. Finally, given our focus on Brexit, we assume that trade costs between America and Europe remain fixed throughout.

Next, we want to consider some alternative trade policy scenarios. For simplicity we consider only three, as summarised in Table 1. In all three, the natural trade costs are the same: Britain is always geographically closer to Europe than to America, so $\delta_{BE} < \delta_{BA}$. The differences between the scenarios relate to the policy-induced trade costs. First is the status quo, where membership of the EU's Customs Union and Single Market ensures that Britain faces lower policy-induced barriers with Europe than with America: τ_{BE} is low while τ_{BA} is high. Second is what we can call the 'cake and eat' scenario: lower trade costs with America plus unchanged trade costs with Europe ensure the best of both worlds for Britain.[21] Third is what we can call the 'global Britain' scenario: withdrawing from the Single Market and the EU Customs Union raises trade costs with Europe, so τ_{BE} rises, but leaves Britain free to negotiate an alternative trade agreement with America, so τ_{BA} falls. We focus throughout on the direct economic consequences of each of these two scenarios relative to the status quo, ignoring political-economy considerations, such as the clear incentive of EU countries to maintain the integrity of the Single Market, or the difficulties for Britain of negotiating new trade agreements on favourable terms with non-EU countries.

Consider first the 'cake and eat' scenario. The only change this implies relative to the status quo is a fall in the trade cost between Britain and America, τ_{BA}. Because trade costs fall in both directions, a lower τ_{BA} has an ambiguous effect on the world excess demand for British output at initial wages (see the Appendix for derivations):

$$\widehat{V}_B - \widehat{Y}_B = -(\sigma - 1)\{\underbrace{\lambda_{BA}(1 - \theta_{BA})}_{(1):\,+} \underbrace{-\lambda_{BB}\theta_{AB}}_{(2):\,-}\}\widehat{\tau}_{BA}. \tag{20}$$

The ambiguity arises because bilateral trade liberalisation has two opposing effects. The first effect reflects increased opportunities in the export market: a reduction in the cost of shipping good B to A raises B's exports. (This effect is dampened but cannot be reversed by the induced rise in the price index in A represented by θ_{BA}, which in any case can be ignored when B is small.) The second effect reflects increased competition in the home market: imports into B are cheaper, which diverts home demand away from home goods.[22] For the moment, assume that the first effect dominates: this case is easier to analyse diagrammatically, and as we will see the answer it gives applies more generally. Hence excess demand for country B's output rises at initial wages, and so the market-clearing locus for good B is shifted upwards as shown in Figure 5, tending to

[21] In practice, this outcome is infeasible: the UK could not align with both the EU and the USA as their rules on non-tariff barriers such as regulatory alignment are very different.

[22] The contrast between these two effects is reminiscent of the distinction between trade creation and trade diversion in the theory of customs unions. The analogy is useful though not perfect, as these effects would arise even if there were only two countries.

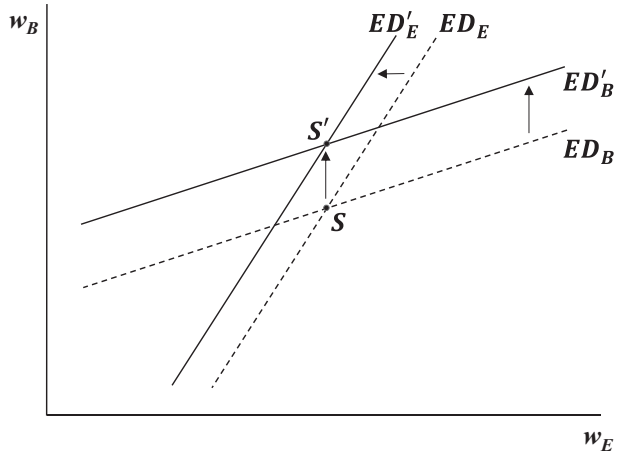


Fig. 5. *'Cake and Eat': the Case Where Wages Rise in Britain.*

raise the equilibrium wage w_B. (In this and subsequent figures the dashed lines are the loci under the status quo scenario.)

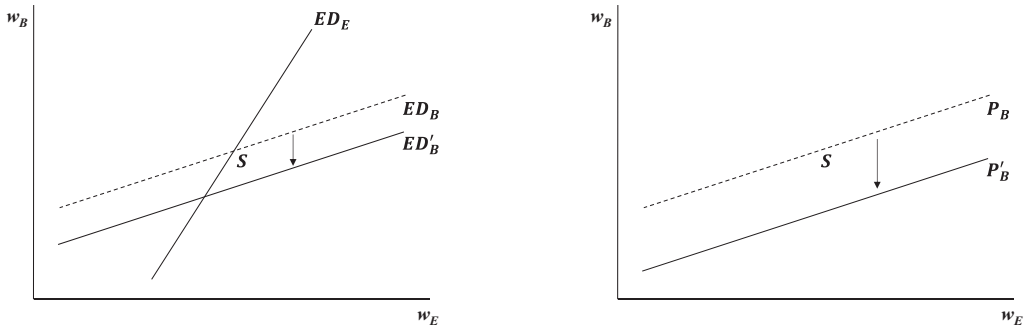
The market-clearing locus for good E is also shifted:

$$\widehat{V}_E - \widehat{Y}_E = (\sigma - 1)(\lambda_{EA}\underline{\theta_{BA}} + \underline{\lambda_{EB}}\theta_{AB})\widehat{\tau}_{BA}. \tag{21}$$

This effect is unambiguous, as European exports face tougher competition in both the British and American markets. Hence, the ED_E locus shifts to the left, as shown in Figure 5. However, these effects are not very strong if Britain is small, when the terms underlined tend to zero. Americans devote a relatively small proportion of their expenditure to British output, so θ_{BA} is small; and Europe exports a relatively small proportion of its GDP to Britain, so λ_{EB} is small. The net effect is therefore a rise in w_B and an ambiguous change in w_E, as illustrated by the move from the initial equilibrium S to the new equilibrium S' in Figure 5. What is unambiguous is that w_B rises relative to w_E; and since an absolute increase in w_B implies that it rises relative to the numeraire w_A, it follows that real wages in Britain must increase.

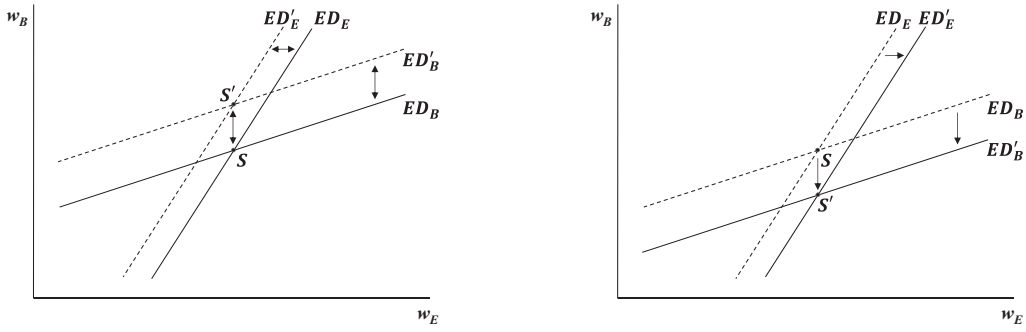
For completeness, recalling the ambiguity in (20), we must also consider the case where a lower τ_{BA} reduces demand for Y_B, as illustrated in panel (a) of Figure 6. This leads to a lower wage in Britain, but nonetheless the real wage there is likely to rise. This is because the same condition which implies a lower wage relative to American prices, namely where θ_{AB} is large enough that home demand for B falls, also implies that the price level in Britain falls a lot: the British consumer spends a lot on imported American goods, so benefits from the fall in the trade cost. As we show in the Appendix, when Britain is small these two effects exactly cancel, so the effect of higher exports dominates: the real wage in B definitely rises. (It can be shown that this result holds for any number of countries.) Hence the overall conclusion is an unsurprising one: having your cake and eating it too is good for you; in this stylised gravity model, Britain unambiguously gains from a bilateral reduction in trade costs with America coupled with unchanged trade costs with Europe.

What then of the second scenario, 'global Britain', where trade barriers with America come down but those with Europe go up? To understand this case it is helpful to begin with a hypothetical



(a) Lower wages (b) Compensated by a lower price level

Fig. 6. *'Cake and Eat': the Case Where Wages Fall in Britain.*



(a) The symmetric benchmark (b) Reality bites

Fig. 7. *'Global Britain'.*

symmetric benchmark in which America and Europe are identical from the perspective of Britain: equally distant, equally rich (so they have the same GDP), and equally restrictive (so their initial policy trade barriers are the same). Moreover, we assume that the proportionate reduction in τ_{BA} is small, and is exactly equal, except opposite in sign, to the proportionate increase in τ_{BE}. A moment's reflection should make it clear that in this case of complete symmetry between A and E the 'global Britain' scenario implies no net effect relative to the status quo. As panel (a) of Figure 7 illustrates, a small reduction in τ_{BA} shifts both loci, leading to a new equilibrium at S', but this is exactly offset by a small increase in τ_{BE}, so the equilibrium remains at the initial point S.

The point of this symmetric benchmark is of course to highlight the fact that, even at this level of abstraction, there are many reasons why the 'global Britain' scenario is likely to be asymmetric. The first and probably the most important reason for a departure from symmetry concerns the depth of economic integration with the two partner countries. The Single Market combined with the EU Customs Union is a very deep trade agreement: rather than merely abolishing tariffs, it also eliminates substantial non-tariff barriers to trade in both goods and services; moreover it

imposes regulatory alignment, underpinned by a framework of commercial law subject to the rulings of the European Court of Justice. It is highly unlikely that any future trade agreement between Britain and America could attain this degree of integration. And, stepping outside the three-country model for a moment, even if the UK and USA were to enter into such a deep trade agreement, it would not be matched by agreements with Britain's other non-EU partners. Thus, Britain can never attain with the rest of the world the same degree of economic integration that it currently enjoys with the EU. In terms of the model's parameters, all this implies that $\tau_{BE}\big|_S < \tau_{BA}\big|_{GB}$: the policy barriers to trade between Britain and Europe in the Single Market status quo are unambiguously lower than those between Britain and America in any plausible 'global Britain' scenario.

There are other reasons why the symmetric benchmark is misleading. A second dimension of asymmetry is size. Defenders of the economic case for Brexit often claim that the UK will gain by switching its trade away from the EU towards faster-growing countries. However, what matters is not absolute size, but size mediated by distance. As we have seen in Section 1, the EU27 accounts for 40% of 2017 UK trade, and countries that have trade agreements with the EU add another 15%. This dominance of EU countries in UK trade is partly a direct result of EU membership but mainly a consequence of geographic proximity; the latter will persist in any post-Brexit scenario, though on much less favourable terms in the case of a 'global Britain' scenario. A third source of asymmetries arises from the difference between increases in low policy costs and decreases in high ones. This is moot for infinitesimal changes in trade costs as in our comparative statics exercises, but it matters for discrete changes. Because τ_{BE} is initially much lower than τ_{BA}, the loss from a 10%-point increase in τ_{BE} is greater in absolute value than the gain from a 10%-point decrease in τ_{BA}. Finally, to the extent that distance imposes fixed costs on trade, the effects of changes in τ_{BE} and τ_{BA} will differ. To take a simple example, if the iceberg trade cost t_{jk} decomposes in an additive rather than a multiplicative way as in (19), then a reduction in policy costs with country j has a smaller effect the further away it is:

$$t_{jk} = \delta_{jk} + \tau_{jk} \quad \Rightarrow \quad \widehat{t}_{jk} = (1 - \omega_{jk})\widehat{\tau}_{jk}, \quad \omega_{jk} \equiv \frac{\delta_{jk}}{t_{jk}}. \tag{22}$$

For this reason too, the costs of raising trade barriers with nearby EU countries are likely to be higher than the benefits of lowering them against far-away trading partners.

All these reasons combined show that the benchmark case of symmetry between the two potential foreign partners is a poor reflection of the actual options facing the UK in choosing between alternative trade agreements. Moreover, all four departures from symmetry work in the same direction: against the neutral outcome of the symmetric benchmark, and in favour of an outcome such as that in panel (b) of Figure 7, where global Britain is poorer than in the status quo.

3. Gravity Anomalies

In this section, we turn to consider some anomalous features of the structural gravity model. In Subsection 3.1 we review the growing evidence against the model's key underlying assumption of CES preferences, while in Subsection 3.2, we show that, under plausible conditions, CES-based structural gravity imposes very strong counterfactual restrictions on bilateral trade balances. Note that there is no contradiction between the two parts of the paper: Sections 1 and 2 have presented the current consensus on the facts about gravity and the theory underlining them, while Sections 3

and 4 will discuss some problems with the theory and present some more speculative thoughts on how it might be improved. Science is provisional, so there is always room for improvement on current models and techniques, despite which, they provide the best answer we can currently give to applied questions.[23]

3.1. *Gravity Assumptions: CES Preferences*

It has been known for some time that CES preferences have very strong implications when embedded in models of monopolistic competition. In particular, they imply that markups should be the same across all firms, and that the pass-through from costs to prices should be always 100%. There is a substantial body of evidence from industrial organisation and international macroeconomics which estimates less than 100% rates of cost or exchange-rate pass-through. See, for example, Weyl and Fabinger (2013) and Gopinath and Itskhoki (2010) respectively. However, it is only relatively recently that credible micro evidence has become available that allows for joint tests of these two predictions. In particular, De Loecker *et al.* (2016), building on De Loecker and Warzynski (2012), study a large sample of Indian firms, and find both that markups are heterogeneous and that pass-through is less than 100%.[24] In the remainder of this section, we draw on Mrázová and Neary (2017) to illustrate just how strongly this evidence is inconsistent with the CES assumption.

The starting point of the approach in Mrázová and Neary (2017) is that, for many purposes, it is more insightful to consider demand functions not in quantity-price space as is standard, but in the space of the elasticity, $\varepsilon \equiv -p/xp'$, and convexity, $\rho \equiv -xp''/p'$, of demand, where $p(x)$ is the inverse demand function. Figure 8 illustrates. We can illustrate the first- and second-order conditions for profit maximisation in this space. They require that a monopolistically competitive firm can only be in equilibrium at a point that lies in an 'admissible region', defined by the conditions $\varepsilon > 1$ and $\rho < 2$: the demand function must be elastic and 'not too' convex. These restrictions are satisfied only in the area above and to the left of the solid dark lines in Figure 8.[25]

The next step is to establish which points in $\{\rho, \varepsilon\}$ space correspond to a given demand function in $\{x, p\}$ space. One special case is the CES: each CES demand function corresponds to a single value of the demand elasticity, $\varepsilon = \sigma$, as well as a single value of ρ, and so is represented by a single point in $\{\rho, \varepsilon\}$ space. The curve labelled 'CES' in Figure 8 is the locus of all such points.[26] This curve is also an important benchmark for comparative statics properties. As already noted, CES demands imply full proportional pass-through from marginal costs to price: $d \log p/d \log c = 1$. It is easy to check that points on a demand curve implying a greater degree of pass-through must correspond to points in this space that lie to the right of the CES locus, where demand is more convex than in the CES case. Mrázová and Neary (2019) call this the 'superconvex' region. Conversely, points in the 'subconvex' region to the left of the CES locus correspond to less than proportional pass-through. Subconvexity is also equivalent to Marshall's

[23] The analogy with evolution is helpful here too. Gould (1981) notes that creationists often interpret, wrongly, disagreements between evolutionary scientists over the detailed mechanisms of evolution as evidence of fundamental disagreement over the validity of evolution as fact.

[24] The findings of De Loecker *et al*. are particularly persuasive because the methods they use to estimate markups place no restrictions on technology or market structure.

[25] The first-order condition is that marginal revenue $p + xp'$ equal marginal cost c, assumed to be exogenous for each firm; the second-order condition is that marginal revenue is decreasing in output: $2p' + xp'' < 0$. It is easy to check that these imply the boundaries of the admissible region as shown.

[26] The CES demand function implies values for the elasticity and convexity of $\varepsilon = \sigma$ and $\rho = (\sigma + 1)/\sigma$ respectively. Eliminating σ yields the expression for the locus of CES point-manifolds: $\rho = (\varepsilon + 1)/\varepsilon$.

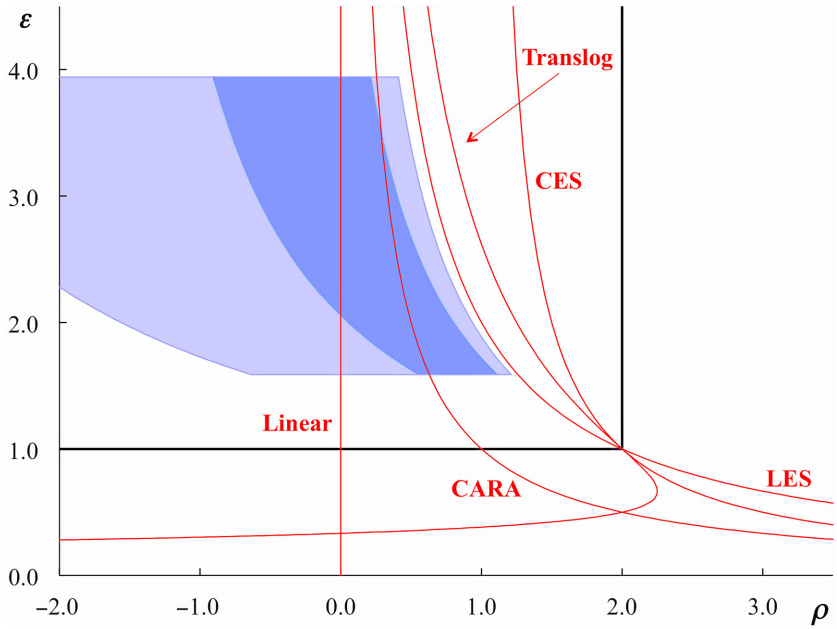


Fig. 8. *Evidence against CES Demands.*
Source. Mrázová and Neary (2017), based on data from De Loecker *et al*. (2016).

second law of demand, the hypothesis that the elasticity of demand increases with price, or, equivalently, decreases with sales.

As for non-CES demand functions, Mrázová and Neary (2017) show that, subject only to relatively mild technical restrictions, any such demand function can be represented by a smooth curve in $\{\rho, \varepsilon\}$ space. They call such a curve the 'demand manifold' corresponding to the original demand function. Figure 8 illustrates the demand manifolds for some widely used demand functions: linear, CARA (constant absolute risk-aversion), translog and LES (linear expenditure system, also known as Stone–Geary).[27] All of these lie in the subconvex portion of the admissible region. Moreover, they exhibit a property that Mrázová and Neary (2017) call 'manifold invariance': while the demand function $p(x; \phi)$ depends on a vector of parameters ϕ, many demand manifolds are invariant with respect to some or all elements of ϕ.[28] This makes it much easier to understand the implications of different assumptions about the form of demand.

How does this way of visualising demand curves relate to the findings of De Loecker *et al.* (2016)? We begin with the fact that their results give estimates, with confidence intervals, of the average markup, m, and pass-through coefficient, κ, for their sample of firms. Assuming that the market is monopolistically competitive, these expressions can be written as functions of the elasticity and convexity of demand:

$$(i)\ m \equiv \frac{p-c}{c} = \frac{1}{\varepsilon - 1}, \qquad (ii)\ \kappa \equiv \frac{d\log p}{d\log c} = \frac{\varepsilon - 1}{\varepsilon}\frac{1}{2-\rho}. \tag{23}$$

[27] The manifolds for these demand functions are given by $\rho = 0$, $\rho = \frac{1}{\varepsilon}$, $\rho = \frac{2}{\varepsilon}$, and $\rho = (3\varepsilon - 1)/\varepsilon^2$, respectively. From the perspective of a monopolistically competitive firm, the translog demand function is observationally equivalent to the almost ideal demand system of Deaton and Muellbauer (1980).

[28] Necessary and sufficient conditions for manifold invariance are given in Mrázová and Neary (2017).

These equations can then be solved to back out the values of the elasticity and convexity implied by the markup and pass-through estimates:

$$(i)\ \ \varepsilon = \frac{m+1}{m}, \qquad (ii)\ \ \rho = 2 - \frac{1}{\kappa}\frac{1}{m+1}. \tag{24}$$

The shaded regions in Figure 8, taken from Mrázová and Neary (2017), show the results of doing this, using the data from De Loecker *et al.* (2016). They give estimates of the pass-through coefficient using both ordinary least squares (OLS) and instrumental variables (IV): the dark-blue region in the figure represents the confidence region implied by the OLS estimate, while the light-blue region represents the confidence region implied by the IV estimate.[29]

It is clear from Figure 8 that all CES demands lie outside the implied confidence regions. (The data also reject LES and translog demands, though less strongly.) More tests of this kind are needed of course, but taken in conjunction with the evidence on pass-through mentioned earlier, it seems reasonable to conclude that there is substantial microeconomic evidence against the hypothesis of CES demands in monopolistic competition. All of this suggests that the assumption of CES preferences which underlies the structural gravity model is open to question.

3.2. *Gravity Predictions: Bilateral Trade Balances*

The previous subsection noted that the key assumption of CES preferences has implications that are inconsistent with a growing body of microeconometric evidence. In this subsection we turn to a prediction of the structural gravity model itself that is not confirmed by the data.

As we have seen, the structural gravity model predicts bilateral trade flows V_{jk}. Hence it also predicts their *ratios*, which equal the bilateral trade balances V_{jk}/V_{kj} in ratio form between each pair of countries.[30] However, under reasonable assumptions about bilateral trade costs, the model's predictions for bilateral trade balances are very stark and are not borne out by the data. This was first pointed out in the frictionless trade case by Davis and Weinstein (2002), who called the anomalous prediction 'the mystery of the excess trade balances'.[31] The result is also derived for a very general structural gravity model by Allen *et al.* (2020), though they do not emphasise the implications for bilateral trade balances. Here we give a self-contained presentation of the result and its implications.

We begin with the simplest case. Assume that overall trade is balanced, so national income and expenditure are equal for all countries: $Y_j = E_j$, $\forall j$. As for trade costs, we assume that they are bilaterally symmetric: $t_{jk} = t_{kj}$, $\forall j, k$. Now recall the structural gravity equation from (7), and divide exports from j to k by exports from k to j:

$$\frac{V_{jk}}{V_{kj}} = \left(\frac{\Pi_j}{P_j} \Big/ \frac{\Pi_k}{P_k}\right)^{\sigma-1}. \tag{25}$$

Given the assumptions we have made, the terms in income, expenditure and bilateral trade costs cancel, leaving only the ratios of outbound to inbound multilateral resistance for each country. However, with bilaterally symmetric trade costs, these are proportional to each other: $P_j = \lambda \Pi_j$,

[29] For details on the estimates and the calculations, see Mrázová and Neary (2017), Online Appendix B17.

[30] There is a better-known precedent for manipulating the expressions for bilateral trade flows implied by structural gravity: the *products* of bilateral trade flows predicted by gravity models are widely used to infer trade costs and the elasticity of trade. See, for example, Head and Ries (2001), Jacks *et al.* (2008) and Caliendo and Parro (2015).

[31] Other empirical work on the result includes Badinger and Fichet de Clairfontaine (2019), Cuñat and Zymek (2018) and Felbermayr and Yotov (2019).

as first pointed out by Anderson and van Wincoop (2003).[32] Hence equation (25) reduces to one, so the model implies that all bilateral trade balances are zero! It hardly needs checking that this prediction is overwhelmingly rejected by the data: there is considerable variation in bilateral trade balances across countries, whence the 'mystery of the excess trade balances'.

Of course, the assumptions made are strong, but they can be relaxed. Taking trade costs first, we can replace the assumption of bilateral symmetry with what Allen and Arkolakis (2016) call *quasi*-symmetric bilateral trade costs:[33]

$$t_{jk} = t_j^X \bar{t}_{jk} t_k^M, \quad \bar{t}_{jk} = \bar{t}_{kj}. \tag{26}$$

Here the dyad-specific term $\bar{t}_{jk}$ is symmetric as before, and in addition each country has two idiosyncratic trade cost terms, one that applies to all its exports and the other to all its imports. This allows among other things for home bias and for border effects, so it represents a big increase in realism over bilaterally symmetric trade costs. However, it does not change the result: as Allen *et al.* (2020) show, the ratio of outbound and inbound multilateral resistances for a given country, though no longer the same across all countries, is now equal to the ratio of that country's export and import trade cost parameters: $\Pi_j / P_j = t_j^X / t_j^M$. As a result, quasi-symmetric bilateral trade costs do not affect the prediction that all bilateral trade balances are zero.

Second, we can relax the assumption of overall trade balance, allowing E_j and Y_j to differ. This does allow for non-zero bilateral trade balances, but only in a restricted way. Retaining the assumption of quasi-symmetric trade costs we now obtain:

$$\frac{V_{jk}}{V_{kj}} = \frac{Y_j}{E_j} \Big/ \frac{Y_k}{E_k}. \tag{27}$$

Thus, the bilateral trade balance between countries j and k equals the ratio of their overall trade balances. Qualitatively, the implications of this are not surprising: a country j is more likely to have a bilateral trade surplus with a partner country k ($V_{jk} > V_{kj}$) if it has an overall trade surplus ($Y_j > E_j$) and if the partner has an overall trade deficit ($Y_k < E_k$). What is surprising and implausible about (27) is that the predicted relationship is very precise. To see this, rewrite (27) in logs:

$$v_{jk} - v_{kj} = \rho_j - \rho_k \quad \text{where:} \quad \rho_j \equiv \log \frac{Y_j}{E_j}. \tag{28}$$

Thus, with quasi-symmetric trade costs, the $\frac{1}{2}n(n-1)$ bilateral trade balances, $v_{jk} - v_{kj}$, are uniquely determined by n country-specific aggregate trade balances ρ_j. This reduces the dimensionality of the bilateral trade balance terms by a factor of $2/(n-1)$. Putting this differently, with quasi-symmetric trade costs, the vector of bilateral trade balances between any country j and all other countries is independent of j, except for a factor of proportionality.

Structural gravity models based on CES preferences thus make the very stark prediction that bilateral asymmetries in trade costs are the only source of bilateral asymmetries in trade balances. Is this plausible? There is abundant empirical evidence that bilateral trade balances are highly asymmetric. As for the convenient assumption of symmetric or quasi-symmetric bilateral trade

[32] There is some potential confusion in the literature on this point. Anderson and van Wincoop (2003) go further than proportionality between P_j and Π_j and set $\lambda = 1$. They call this 'an implicit normalization'; it would be more conventional to call it a choice of numeraire. As such, it is perfectly valid, though it is not advisable if another nominal variable is also chosen as numeraire.

[33] This assumption can be found in Eaton and Kortum (2002) and Anderson and van Wincoop (2003).

costs, it is easy to think of cases where it might be expected to fail. (As already noted, composition effects are an obvious example: e.g., a resource-exporting country might be expected to incur very different transport costs on its exports than on its imports.) Yet it is difficult to believe that trade cost asymmetries alone can save the CES gravity model from its singular inability to allow for the observed diversity in bilateral trade balances. This suggests that relaxing the CES assumption itself may provide a route to a better explanation of bilateral trade balances. In the next section we turn to explore an approach to doing this.

4. Subconvex Gravity

As we saw in the last section, the constant elasticities of demand and of trade that are a central feature of CES-based structural gravity models have anomalous implications once we move away from aggregate trade flows. This suggests that it is worth exploring alternative approaches to modelling trade flows, notwithstanding the fact that departing from the assumption of CES preferences requires a sacrifice of theoretical tractability and ease of estimation. As a compromise that relaxes the CES assumption but does not lead to an intractable specification, we explore in this section the case of demands that are generated by additively separable preferences:

$$U_k = \sum_{j=1}^{n} u\left(\eta_j x_{jk}\right). \tag{29}$$

Here η_j is a taste shifter for country j's good, written in a way that is consistent with the CES specification in (3). These preferences have the advantage that, as in the CES special case, all cross-price effects are summarised in terms of a single aggregate. At the same time they allow for subconvexity, and so are consistent with the empirical evidence in Subsection 3.1. Moreover, they nest not just CES itself, but also both sub- and superconvex cases, so we can test for subconvexity.

Additively separable preferences imply a simple first-order condition: the marginal utility of each good j in each consuming country k depends only on the amount of it consumed, and equals its price times country k's marginal utility of income. This can then be solved for demands that, as in the CES case, depend only on exporter and importer terms and on the trade cost between j and k:

$$u'(\eta_j x_{jk}) = \lambda_k p_{jk} \quad \Rightarrow \quad x_{jk} = \eta_j^{-1} f\left(\lambda_k p_j t_{jk}\right). \tag{30}$$

Multiplying by price and taking a first-order approximation expresses changes in the value of trade as a function of changes in the origin-country taste shifter and price, the destination-country marginal utility of income, and the bilateral trade cost:

$$\hat{V}_{jk} = -\hat{\eta}_j - (\sigma_{jk} - 1)\hat{p}_j - \sigma_{jk}\hat{\lambda}_k - \left(\sigma_{jk} - 1\right)\hat{t}_{jk}. \tag{31}$$

The only difference from the CES case is that the elasticity is variable. Of course, this is a major difference: the elasticity is not only variable but differs between each distinct pair of countries. The assumption of additive separability is helpful here, since it implies that the elasticity depends only on the volume of trade: $\sigma_{jk} \equiv \sigma(x_{jk})$.[34] In addition, subconvexity, for which there is substantial microeconometric evidence as we saw in the last section, implies that the elasticity is decreasing in the volume of trade: σ_{jk} is decreasing in x_{jk}.[35]

[34] See Goldman and Uzawa (1964).

[35] With additively separable preferences as in (29), the elasticity depends on per capita consumption, not total consumption. This does not affect our results, as importer population is subsumed into the importer fixed effect. The two

Taking (31) to data poses a challenge, as each σ_{jk} coefficient on the right-hand side depends on x_{jk}, which is a component of the dependent variable on the left-hand side. Allowing the coefficients to vary continuously with export volume is not feasible. However, we can allow them to vary discretely by using quantile regression.[36]

To implement this we first order the observations by V_{jk}, and divide them into quantiles. Let quantile $q \in (0, 1)$ denote the value of the dependent variable which splits the data into proportions q below and $1 - q$ above.[37] In practice, we use one hundred divisions, so we work with centiles. The specification ideally requires that we order the observations by the volume of trade x_{jk}, not the value, but this is not possible with the data available. We then estimate for each quantile q the following regression:

$$\log V_{q,jk} = F_{q,j} + F_{q,k} + \mu_q \log t_{jk} + u_{q,jk}. \tag{32}$$

This can be compared with the corresponding equation in the CES case, equation (9). We use the method of moments-quantile regression estimation procedure of Machado and Santos Silva (2019) to estimate the quantile coefficients; this approach is particularly attractive in a panel setting with a large number of individual fixed effects. Note that each quantile regression is estimated over the whole sample but with different penalties depending on the quantile we are interested in.

Figure 9 shows the estimated distance coefficients from the quantile regressions, with bootstrapped 95% confidence intervals.[38] For reference, Figure 9 also presents the OLS estimate.[39] The sample, as described in Section 1, includes 206 countries that report at least one strictly positive export value in 2017 with a partner. Of the potential 42,230 bilateral trade flows, only $n = 23{,}251$ report a positive trade in 2017. Table 2 presents the results of significance tests for differences between the decile and OLS estimates of the distance coefficient.

Overall, Figure 9 and Table 2 present persuasive evidence for subconvexity. The quantile estimates of the distance coefficient are significantly decreasing (in absolute value) in the value of trade, and the one-size-fits-all CES-based constant-coefficient hypothesis is rejected for both relatively high and relatively low trade flows.[40] This suggests a promising research agenda. Since

specifications (quantile regression on total or on per-capita consumption) yield identical results, except of course for the estimated importer fixed effect itself.

[36] For a previous application of quantile regression in a gravity context, see Baltagi and Egger (2016).

[37] The general idea is the following. Let $|e_i|$ denote the absolute deviations from the regression equation, $|e_i| = |y_i - x_i'\beta|$, where β denotes the coefficient vector $(F_j, F_k, \mu)'$. The quantile regression (QR) for quantile q selects the coefficient estimates to minimise a weighted sum of the $|e_i|$, where the weights assign asymmetric penalties $q\,|e_i|$ for underprediction and $(1 - q)\,|e_i|$ for overprediction. Thus the quantile regression estimator for quantile q minimises the loss function:

$$L(\beta_q) = \sum_{i:\, y_i \geq x_i'\beta_q}^{n} q\,|y_i - x_i'\beta_q| + \sum_{i:\, y_i < x_i'\beta_q}^{n} (1-q)\,|y_i - x_i'\beta_q|.$$

[38] We perform 100 replications, the sample drawn during each replication is a bootstrap sample of clusters.

[39] Recalling Subsection 2.1, this equals −1.452 with a clustered standard error of 0.019.

[40] Our findings are in line with those of Novy (2013) who uses a different technique to allow for variation in the distance coefficient: he uses an OLS specification, with slope dummies on the distance coefficient for each quantile of the predicted value of trade. Consistent with our results, he finds that the distance coefficient falls in absolute value with the level of trade. However, he assumes that other coefficients are invariant, whereas our theoretical specification implies that all coefficients (including fixed effects) vary with the level of trade. Chernozhukov *et al.* (2018) develop a different estimator for this class of problem, and give an application to trade data from 1986, where they find evidence that the distance coefficient rises in absolute value with the value of trade. However, their results are not comparable with ours: they include all zero trade flows, whereas our theoretical specification predicts that the coefficients vary with the level of trade along the intensive margin, but has nothing to say about the probability of trade along the extensive margin.

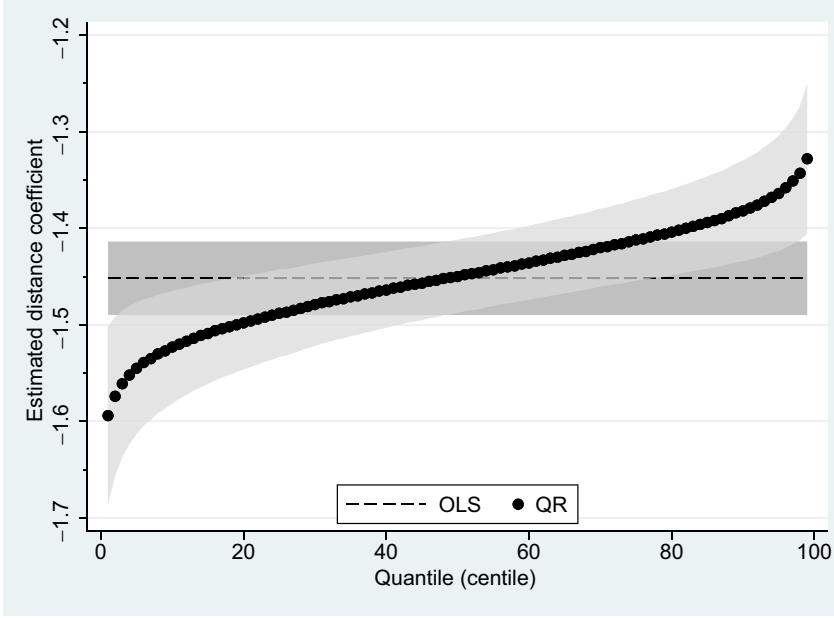


Fig. 9. *Quantile and OLS Estimates of the Distance Coefficient.*
Notes: We report the estimated distance coefficients from OLS—equation (9)—and QR—equation (32)—with bootstrapped 95% confidence intervals. We control for the full set of importer and exporter fixed effects and, in addition to distance, for contiguity, common language, colonial ties, membership of a common trade agreement and/or currency area. Data are from the CEPII BACI and CEPII Gravity databases.
Source. Authors' calculations.

Table 2. *t-Statistics Testing the Significance of Differences between the Quantile and OLS Estimates of the Distance Coefficient.*

	μ_{OLS}	μ_{10}	μ_{20}	μ_{30}	μ_{40}	μ_{50}	μ_{60}	μ_{70}	μ_{80}
μ_{10}	2.01*								
μ_{20}	1.47	0.65							
μ_{30}	0.94	1.19	0.56						
μ_{40}	0.44	1.65*	1.06	0.52					
μ_{50}	0.08	2.07*	1.54	1.02	0.52				
μ_{60}	0.58	2.46*	1.97*	1.49	1.01	0.51			
μ_{70}	1.11	2.84*	2.41*	1.96*	1.52	1.03	0.54		
μ_{80}	1.61	3.19*	2.80*	2.40*	1.99*	1.54	1.06	0.54	
μ_{90}	2.13*	3.55*	3.20*	2.84*	2.48*	2.07*	1.64*	1.15	0.63

Notes: * indicates values that are significantly different at the 5% level. Tests comparing an estimated quantile coefficient with the OLS estimate are two-sided, with a 5% threshold of 1.96. Tests comparing two estimated quantile coefficients are one-sided, with a 5% threshold of 1.64. Tests are based on coefficient values and bootstrapped standard errors reported in Figure 9.
Source. Authors' calculations.

moving from a CES-dominated view of gravity to one that allows for subconvexity provides a better fit, at least for this data set, it has the potential to help resolve both of the anomalies with CES gravity discussed in Section 3. By construction, subconvexity is consistent with the microeconometric evidence in favour of variable markups and less than 100% pass-through. It also allows for the elasticity of import demand to vary systematically across destinations, for

which there is some evidence (see Novy, 2013, for example). As for the trade balances puzzle, when demands are subconvex bilateral balances depend on distance, and by more the greater the imbalance. Subconvex demands provide an empirically better fit to trade volumes as we have seen, so it is likely that they will also help alleviate the mystery of the excess trade balances. In future work we hope to explore the potential for subconvex gravity in both these directions.

There is of course a cost to moving away from CES gravity: it is not possible to solve for a closed-form structural gravity equation in the general subconvex case. However, two other routes to working with gravity models of this kind are open. First, the local comparative statics results presented in Section 2 continue to apply: equations (A1)–(A4) in the Appendix show that, even in the CES case, all the own- and cross-price elasticities of import demand vary with the two countries involved. While it is true that the elasticity of substitution is the same in all cases, the elasticities also depend on the physical and value shares in world trade, λ_{jk} and θ_{jk}. Hence the qualitative conclusions drawn in Section 2, based on the derivations in the Appendix, require no modification whatsoever when we replace the CES assumption with subconvex preferences, which also allows the elasticity of demand to vary with the exporting and importing country. Second, computer modelling using parameterised forms of subconvex gravity poses no major problems in principle. In future work we hope to explore this approach.

Finally, we can ask what if any are the implications of subconvex gravity for predictions about the effects of trade policy changes such as Brexit? With subconvexity, elasticities of import demand are higher in smaller markets and lower in larger ones. Taking this into account explicitly is likely to lead to more nuanced predictions about the effects of trade policy changes in multi-country settings. However, there is no reason to expect that it will reverse the predictions of CES-based models. As shown in Section 2, the qualitative predictions of gravity models are robust to alternative specifications of demand, and there is no presumption that relaxing the CES assumption will change the quantitative predictions one way or another. In the 'cake and eat' scenario, the larger benefits of reducing trade costs in smaller non-EU markets are likely to be offset by the greater difficulty of raising exports to bigger markets. And in the more plausible 'global Britain' scenario, the same applies to the effects of increased access costs in EU markets: smaller reductions in exports to larger markets traded off against larger falls in smaller markets. The net outcome of these changes needs to be explored in fully specified empirically based models. But *a priori* it seems unlikely that the implications of subconvex gravity for the estimated effects of Brexit will be major. While, as we have seen, subconvex gravity is likely to imply first-order deviations from CES predictions for markups, pass-through and bilateral trade balances, it is likely that these will tend to cancel in aggregate yielding only second-order deviations for welfare.[41]

5. Conclusion

In this paper we have provided an overview of the role of gravity in international trade, developed some pedagogic tools to illustrate it, and discussed some ways in which the standard models could be extended. We have emphasised that gravity in trade is both fact and theory. At the level of fact, there is overwhelming evidence that trade tends to fall with distance, as even a superficial examination of UK export patterns confirms. At the level of theory, the structural gravity model is consistent with a range of theoretical underpinnings, including Ricardian comparative advantage and monopolistic competition with heterogeneous firms. We have emphasised that it is no more

[41] We are grateful to Estelle Cantillon for suggesting this interpretation.

and no less than a simple general-equilibrium system, and have presented new analytic tools for understanding it in small-scale applications. We have also noted some difficulties with the standard models, 'gravity anomalies', which arise from the underlying assumption of CES preferences, and which imply that models with a constant elasticity of trade cannot tell the whole story about trade patterns. Finally, we have sketched an approach based on subconvex gravity that provides a promising way forward. Relaxing the assumption of a constant elasticity of trade makes the model more consistent with microeconomic evidence on markups and pass-through, and also avoids the CES model's stark predictions about bilateral trade balances. However, it is unlikely to change the 'three iron laws' of the economics of Brexit found in recent gravity-based studies.

Appendix A. Solving the Three-Country Case

To confirm the properties of Figures 4 and 5, consider first the general market-clearing condition (12), specialised to the case of country B in the three-country model. Recall that w_A is constant by choice of numeraire; that natural trade costs do not change, so $\widehat{t}_{jk} = \widehat{\tau}_{jk}$ for all j, k; that trade costs between A and E are assumed to be constant: $\widehat{\tau}_{AE} = 0$; and that trade costs between A and B and between B and E are assumed to be symmetric: $\widehat{\tau}_{AB} = \widehat{\tau}_{BA} = \widehat{\tau}_A$ and $\widehat{\tau}_{BE} = \widehat{\tau}_{EB} = \widehat{\tau}_E$. Substituting from the changes in the price indices and in demands given by (13) and (14) respectively into the market-clearing condition (12) gives:

$$\widehat{V}_B - \widehat{Y}_B = \varepsilon_{BB}\widehat{w}_B + \varepsilon_{BE}\widehat{w}_E + \varepsilon_{Bt_A}\widehat{\tau}_A + \varepsilon_{Bt_E}\widehat{\tau}_E = 0, \tag{A1}$$

where the elasticities of excess demand for country B's output with respect to prices and trade costs can be written in terms of σ and share parameters as follows:

$$\begin{bmatrix} \varepsilon_{BB} \\ \varepsilon_{BE} \\ \varepsilon_{Bt_A} \\ \varepsilon_{Bt_E} \end{bmatrix} = \begin{bmatrix} -(\sigma-1)\lambda_{BB}(1-\theta_{BB}) - \lambda_{BE}(\sigma(1-\underline{\theta_{BE}}) + \underline{\theta_{BE}}) - \lambda_{BA}(\sigma(1-\underline{\theta_{BA}}) + \underline{\theta_{BA}}) \\ (\sigma-1)\lambda_{BB}\theta_{EB} + \lambda_{BE}((\sigma-1)\theta_{EE}+1) + (\sigma-1)\lambda_{BA}\theta_{EA} \\ -(\sigma-1)(\lambda_{BA}(1-\underline{\theta_{BA}}) - \lambda_{BB}\theta_{AB}) \\ -(\sigma-1)(\lambda_{BE}(1-\underline{\theta_{BE}}) - \lambda_{BB}\theta_{EB}) \end{bmatrix}. \tag{A2}$$

Terms underlined are zero when country B is infinitesimally small. The expressions for the elasticities in (A2) confirm the properties noted in the text: ε_{BB} and ε_{BE} are negative and positive respectively, while ε_{Bt_A} and ε_{Bt_E} are ambiguous in sign.

We can repeat the analogous substitutions for the excess demand for country E's good:

$$\widehat{V}_E - \widehat{Y}_E = \varepsilon_{EB}\widehat{w}_B + \varepsilon_{EE}\widehat{w}_E + \varepsilon_{Et_A}\widehat{\tau}_A + \varepsilon_{Et_E}\widehat{\tau}_E = 0, \tag{A3}$$

where:

$$\begin{bmatrix} \underline{\varepsilon_{EB}} \\ \varepsilon_{EE} \\ \underline{\varepsilon_{Et_A}} \\ \underline{\varepsilon_{Et_E}} \end{bmatrix} = \begin{bmatrix} \underline{\lambda_{EB}}((\sigma-1)\theta_{BB}+1) + (\sigma-1)\lambda_{EE}\underline{\theta_{BE}} + (\sigma-1)\lambda_{EA}\underline{\theta_{BA}} \\ -\underline{\lambda_{EB}}(\sigma(1-\theta_{EB}) + \theta_{EB}) - (\sigma-1)\lambda_{EE}(1-\theta_{EE}) - \lambda_{EA}(\sigma(1-\theta_{EA}) + \theta_{EA}) \\ (\sigma-1)(\lambda_{EA}\underline{\theta_{BA}} + \underline{\lambda_{EB}}\theta_{AB}) \\ -(\sigma-1)(\underline{\lambda_{EB}}(1-\theta_{EB}) - \lambda_{EE}\underline{\theta_{BE}}) \end{bmatrix}. \tag{A4}$$

These expressions have similar properties to those in (A2), with two exceptions. First, country E is not directly involved in an increase in trade costs between A and B, so demand for its good changes as a result only to the extent that the price indices in A and B rise; hence ε_{Et_A} is

unambiguously positive. Second, when country B is infinitesimally small, neither its home price nor either of the trade costs it faces have any effect on the demand for country E's output.

Combining (A1) and (A3):

$$\begin{bmatrix} -\varepsilon_{BB} & -\varepsilon_{BE} \\ -\varepsilon_{EB} & -\varepsilon_{EE} \end{bmatrix} \begin{bmatrix} \widehat{w}_B \\ \widehat{w}_E \end{bmatrix} = \begin{bmatrix} \varepsilon_{Bt_A} \\ \varepsilon_{Et_A} \end{bmatrix} \widehat{\tau}_A + \begin{bmatrix} \varepsilon_{Bt_E} \\ \varepsilon_{At_E} \end{bmatrix} \widehat{\tau}_E. \tag{A5}$$

Let $\Delta \equiv \varepsilon_{BB}\varepsilon_{EE} - \varepsilon_{BE}\varepsilon_{EB}$ denote the determinant of the coefficient matrix on the left-hand side. Because demands exhibit gross substitutability, we know that $-\varepsilon_{BB} > \varepsilon_{BE} > 0$ and $-\varepsilon_{EE} > \varepsilon_{EB} \geq 0$. It follows that Δ is positive. Solving for the effect of a change in the bilateral trade cost with A, τ_A, on the wage in country B gives:

$$\widehat{w}_B = \frac{1}{\Delta} \begin{vmatrix} \varepsilon_{Bt_A} & -\varepsilon_{BE} \\ \varepsilon_{Et_A} & -\varepsilon_{EE} \end{vmatrix} \widehat{\tau}_A = \frac{1}{\Delta}(-\varepsilon_{Bt_A}\varepsilon_{EE} + \varepsilon_{BE}\varepsilon_{Et_A})\widehat{\tau}_A. \tag{A6}$$

Assume now that country B is small, so that $\varepsilon_{EB} = \varepsilon_{Et_A} = 0$. In Figures 4–7 this implies that the ED_E locus is vertical and is unaffected by changes in τ_A, so w_E is independent of τ_A. In algebraic terms, it implies that the change in w_B simplifies to:

$$\widehat{w}_B = -\frac{\varepsilon_{Bt_A}}{\varepsilon_{BB}}\widehat{\tau}_A. \tag{A7}$$

The denominator is negative, so the sign of the change in the wage rate depends on the numerator. Recalling (A2), this equals the change in the excess demand for country B's output at initial wages, as given by (20) in the text. As noted there, it is the sum of a positive 'trade creation' effect and a negative 'trade diversion' one.

Our main interest is in the change in the real wage in country B, which in general equals:

$$\widehat{w}_B - \widehat{P}_B = (1 - \theta_{BB})\,\widehat{w}_B - \theta_{EB}\,\widehat{w}_E - (\theta_{AB}\widehat{\tau}_A + \theta_{EB}\widehat{\tau}_E)\,. \tag{A8}$$

We consider the special case where $\widehat{w}_E = 0$ because B is small; and we focus on the effect of a fall in τ_A: $\widehat{\tau}_A < 0$. Substituting for $\widehat{w}_B$ from (A7):

$$\widehat{w}_B - \widehat{P}_B = -(1 - \theta_{BB})\frac{\varepsilon_{Bt_A}}{\varepsilon_{BB}}\widehat{\tau}_A - \theta_{AB}\widehat{\tau}_A = -\frac{1}{\varepsilon_{BB}}\left((1 - \theta_{BB})\varepsilon_{Bt_A} + \theta_{AB}\varepsilon_{BB}\right)\widehat{\tau}_A. \tag{A9}$$

Next we substitute for ε_{Bt_A} and ε_{BB} from (A2):

$$\begin{aligned} \widehat{w}_B - \widehat{P}_B = \frac{1}{\varepsilon_{BB}}\Big((1 - \theta_{BB})(\sigma - 1)(\lambda_{BA}\underline{-\lambda_{BB}\theta_{AB}}) \\ + \theta_{AB}((\sigma - 1)\lambda_{BB}(1 - \theta_{BB}) + \sigma(\lambda_{BE} + \lambda_{BA}))\Big)\widehat{\tau}_A. \end{aligned} \tag{A10}$$

As noted in the text, the source of ambiguity is the trade diversion effect represented by the underlined expression in the first set of parentheses: lowering the trade cost between A and B reduces the price level in B by more the greater is θ_{AB}; this in turn lowers domestic demand for the home good by more the greater is λ_{BB}. However, though this tends to reduce the wage in B, the lower price level tends to increase the real wage, which also rises by more the greater is θ_{AB}. In fact the two terms exactly cancel:

$$\widehat{w}_B - \widehat{P}_B = \frac{1}{\varepsilon_{BB}}((\sigma - 1)(1 - \theta_{BB})\lambda_{BA} + \sigma\theta_{AB}(1 - \lambda_{BB}))\widehat{\tau}_A. \tag{A11}$$

Thus (bearing in mind that ε_{BB} is negative), a reduction in τ_A unambiguously raises the real wage in B, as stated in the text. It can be shown that this result holds for any number of countries, provided B is small: the downward effect of the trade cost reduction on the home price index is always enough to counteract the negative trade diversion effect on wages of making imports more attractive and hence the home good less attractive to home consumers.

Geneva and CEPR
Geneva, CEPR and CESifo
Oxford, CEPR and CESifo

Additional Supporting Information may be found in the online version of this article:

Replication Package

References

Alabrese, E., Becker, S.O., Fetzer, T. and Novy, D. (2019). 'Who voted for Brexit? Individual and regional data combined', *European Journal of Political Economy*, vol. 56, pp. 132–50.
Allen, T. and Arkolakis, C. (2016). 'Elements of advanced international trade', http://www.econ.yale.edu/~ka265/teaching/GradTrade/notes/ClassNotes.pdf.
Allen, T., Arkolakis, C. and Takahashi, Y. (2020). 'Universal gravity', *Journal of Political Economy*, vol. 128(2), pp. 393–433.
Alvarez, F. and Lucas, R.E. (2007). 'General equilibrium analysis of the Eaton–Kortum model of international trade', *Journal of Monetary Economics*, vol. 54(6), pp. 1726–68.
Anderson, J.E. (1979). 'A theoretical foundation for the gravity equation', *American Economic Review*, vol. 69(1), pp. 106–16.
Anderson, J.E. and van Wincoop, E. (2003). 'Gravity with gravitas: a solution to the border puzzle', *American Economic Review*, vol. 93(1), pp. 170–92.
Anderson, J.E. and Yotov, Y.V. (2010). 'The changing incidence of geography', *American Economic Review*, vol. 100(5), pp. 2157–86.
Arkolakis, C., Costinot, A. and Rodríguez-Clare, A. (2012). 'New trade models, same old gains?', *American Economic Review*, vol. 102(1), pp. 94–130.
Armington, P.S. (1969). 'A theory of demand for products distinguished by place of production', *International Monetary Fund Staff Papers*, vol. 16(1), pp. 159–78.
Badinger, H. and Fichet de Clairfontaine, A. (2019). 'Trade balance dynamics and exchange rates: in search of the J-curve using a structural gravity approach', *Review of International Economics*, vol. 27, pp. 1268–93.
Baltagi, B.H. and Egger, P. (2016). 'Estimation of structural gravity quantile regression models', *Empirical Economics*, vol. 50(50), pp. 5–15.
Baqaee, D.R. and Farhi, E. (2017). 'Productivity and misallocation in general equilibrium', CEPR Discussion Paper No. 12447.
Bergstrand, J.H. (1985). 'The gravity equation in international trade: some microeconomic foundations and empirical evidence', *Review of Economics and Statistics*, vol. 67(3), pp. 474–81.
Brakman, S., Garretsen, H. and Kohl, T. (2018). 'Consequences of Brexit and options for a "global Britain"', *Papers in Regional Science*, vol. 97(1), pp. 55–72.
Cairncross, F. (1997). *The Death of Distance*, London: Texere Publishing.
Caliendo, L. and Parro, F. (2015). 'Estimates of the trade and welfare effects of NAFTA', *Review of Economic Studies*, vol. 82(1), pp. 1–44.
Chaney, T. (2008). 'Distorted gravity: the intensive and extensive margins of international trade', *American Economic Review*, vol. 98(4), pp. 1707–21.
Chernozhukov, V., Fernandez-Val, I. and Weidner, M. (2018). 'Network and panel quantile effects via distribution regression', CEMMAP Working Paper CWP21/18, UCL.
Cowgill, B. and Dorobantu, C. (2012). 'Gravity and borders in online commerce: results from Google', mimeo, Department of Economics, University of Oxford.
Cuñat, A. and Zymek, R. (2018). 'Bilateral trade imbalances', mimeo, University of Vienna.
Davies, R.B. and Studnicka, Z. (2018). 'The heterogeneous impact of Brexit: early indications from the FTSE', *European Economic Review*, vol. 110, pp. 1–17.
Davis, D.R. and Weinstein, D.E. (2002). 'The mystery of the excess trade (balances)', *American Economic Review*, vol. 92(2), pp. 170–4.

De Loecker, J., Goldberg, P.K., Khandelwal, A.K. and Pavcnik, N. (2016). 'Prices, markups and trade reform', *Econometrica*, vol. 84(2), pp. 445–510.
De Loecker, J. and Warzynski, F. (2012). 'Markups and firm-level export status', *American Economic Review*, vol. 102(6), pp. 2437–71.
Deaton, A. and Muellbauer, J. (1980). 'An almost ideal demand system', *American Economic Review*, vol. 70(3), pp. 312–26.
Dekle, R., Eaton, J. and Kortum, S. (2008). 'Global rebalancing with gravity: measuring the burden of adjustment', *IMF Staff Papers*, vol. 55(3), pp. 511–40.
Dhingra, S., Huang, H., Ottaviano, G., Paulo Pessoa, J., Sampson, T. and Van Reenen, J. (2017). 'The costs and benefits of leaving the EU: trade effects', *Economic Policy*, vol. 32(92), pp. 651–705.
Diewert, W.E. and Woodland, A.D. (1977). 'Frank Knight's theorem in linear programming revisited', *Econometrica*, vol. 45(2), pp. 375–98.
Disdier, A.C. and Head, K. (2008). 'The puzzling persistence of the distance effect on bilateral trade', *Review of Economics and Statistics*, vol. 90(1), pp. 37–48.
Eaton, J. and Kortum, S. (2002). 'Technology, geography, and trade', *Econometrica*, vol. 70(5), pp. 1741–79.
Fally, T. (2015). 'Structural gravity and fixed effects', *Journal of International Economics*, vol. 97(1), pp. 76–85.
Felbermayr, G. and Yotov, Y.V. (2019). 'From theory to policy with gravitas: a solution to the mystery of the excess trade balances', CESifo Working Paper Series 7825, CESifo Group Munich.
Friedman, T. (2005). *The World Is Flat*, New York: Farrar, Straus and Giroux.
Gaulier, G. and Zignago, S. (2010). 'BACI: international trade database at the product-level. The 1994–2007 version', CEPII Working Paper, No. 2010-23.
Goldman, S.M. and Uzawa, H. (1964). 'A note on separability in demand analysis', *Econometrica*, vol. 32(3), pp. 387–98.
Gopinath, G. and Itskhoki, O. (2010). 'Frequency of price adjustment and pass-through', *Quarterly Journal of Economics*, vol. 125(2), pp. 675–727.
Gould, S.J. (1981). 'Evolution as fact and theory', *Discover*, vol. 2, 34–7; reprinted in *Hen's Teeth and Horse's Toes*, (1994), pp. 253–62, New York: W.W. Norton.
Head, K. and Mayer, T. (2014). 'Gravity equations: workhorse, toolkit, and cookbook', in (G. Gopinath, E. Helpman and K. Rogoff, eds.), *Handbook of International Economics*, vol. 4, chap. 3, pp. 131–95, Elsevier.
Head, K., Mayer, T. and Ries, J. (2010). 'The erosion of colonial trade linkages after independence', *Journal of International Economics*, vol. 81(1), pp. 1–14.
Head, K. and Ries, J. (2001). 'Increasing returns versus national product differentiation as an explanation for the pattern of U.S.–Canada trade', *American Economic Review*, vol. 91(4), pp. 858–76.
Helpman, E. (1987). 'Imperfect competition and international trade: evidence from fourteen industrial countries', *Journal of the Japanese and International Economies*, vol. 1(1), pp. 62–81.
Jacks, D.S., Meissner, C.M. and Novy, D. (2008). 'Trade costs, 1870–2000', *American Economic Review*, vol. 98(2), pp. 529–34.
Jones, R.W. (1965). 'The structure of simple general equilibrium models', *Journal of Political Economy*, vol. 73(6), pp. 557–72.
Jones, R.W. and Scheinkman, J.A. (1977). 'The relevance of the two-sector production model in trade theory', *Journal of Political Economy*, vol. 85(5), pp. 909–35.
Keller, W. and Yeaple, S.R. (2013). 'The gravity of knowledge', *American Economic Review*, vol. 103(4), pp. 1414–44.
Kimura, F. and Lee, H.H. (2006). 'The gravity equation in international trade in services', *Review of World Economics*, vol. 142(1), pp. 92–121.
Kleinert, J. and Toubal, F. (2010). 'Gravity for FDI', *Review of International Economics*, vol. 18(1), pp. 1–13.
Krugman, P. (1980). 'Scale economies, product differentiation, and the pattern of trade', *American Economic Review*, vol. 70(5), pp. 950–9.
Larch, M., Wanner, J., Yotov, Y.V. and Zylkin, T. (2019). 'Currency unions and trade: a PPML re-assessment with high-dimensional fixed effects', *Oxford Bulletin of Economics and Statistics*, vol. 81(3), pp. 487–510.
Lendle, A., Olarreaga, M., Schropp, S. and Vézina, P.L. (2016). 'There goes gravity: eBay and the death of distance', Economic Journal, vol. 126(591), pp. 406–41.
McGrattan, E.R. and Waddle, A. (2018). 'The impact of Brexit on foreign investment and production', Federal Reserve Bank of Minneapolis, Research Department Staff Report 542.
Machado, J.A. and Santos Silva, J. (2019). 'Quantiles via moments', *Journal of Econometrics*, vol. 213(1), pp. 143–73.
Maggi, G., Mrázová, M. and Neary, J.P. (2018). 'Choked by red tape? The political economy of wasteful trade barriers', CEPR Discussion Paper No. 12985.
Mayer, T., Vicard, V. and Zignago, S. (2019). 'The cost of non-Europe, revisited', *Economic Policy*, vol. 34(98), pp. 145–99.
Mayer, T. and Zignago, S. (2011). 'Notes on CEPII's distances measures: the Geodist database', CEPII Working Paper No. 2011-25.
Melitz, M.J. (2003). 'The impact of trade on intra-industry reallocations and aggregate industry productivity', *Econometrica*, vol. 71(6), pp. 1695–725.
Mrázová, M. and Neary, J.P. (2017). 'Not so demanding: demand structure and firm behavior', *American Economic Review*, vol. 107(12), pp. 3835–74.

Mrázová, M. and Neary, J.P. (2019). 'Selection effects with heterogeneous firms', *Journal of the European Economic Association*, vol. 17(4), pp. 1294–334.
Newton, I. (1713). *Philosophiae Naturalis Principia Mathematica [Mathematical Principles of Natural Philosophy]*, 2nd edition, Cambridge; English translation by Andrew Motte, London, 1729.
Novy, D. (2013). 'International trade without CES: estimating translog gravity', *Journal of International Economics*, vol. 89(2), pp. 271–82.
O'Rourke, K. (2019). *A Short History of Brexit: From Brentry to Backstop*, London: Pelican Books.
Portes, R. and Rey, H. (2005). 'The determinants of cross-border equity flows', *Journal of International Economics*, vol. 65(2), pp. 269–96.
Sampson, T. (2017). 'Brexit: the economics of international disintegration', *Journal of Economic Perspectives*, vol. 31(4), pp. 163–84.
Santos Silva, J. and Tenreyro, S. (2006). 'The log of gravity', *Review of Economics and Statistics*, vol. 88(4), pp. 641–58.
Weyl, E.G. and Fabinger, M. (2013). 'Pass-through as an economic tool: principles of incidence under imperfect competition', *Journal of Political Economy*, vol. 121(3), pp. 528–83.
Yotov, Y.V. (2012). 'A simple solution to the distance puzzle in international trade', *Economics Letters*, vol. 117(3), pp. 794–8.
Yotov, Y.V., Piermartini, R., Monteiro, J.A. and Larch, M. (2016). *An Advanced Guide to Trade Policy Analysis: The Structural Gravity Model*, Geneva: WTO.