The Gravity Model of Economic Interaction

James E. Anderson
Boston College and NBER

August 17, 2016

Newton's physical law of gravity inspired the original gravity model of economic interaction over space. The intuition was that economic flows might plausibly vary with the masses of economic activity at origin and destination and inversely with the distance between origin and destination. Newton's Law applied strictly predicts that the economic flow X_{ij} from origin i to destination j is

$$X_{ij} = G\frac{Y_i E_j}{D_{ij}^2} \tag{1}$$

where G is the gravitational constant, Y_i is the relevant economic activity mass at origin i, E_j is the relevant economic activity mass at destination j and D_{ij} is the distance between i and j. Matching the predicted value on the right hand side of equation (1) to observed economic flows on the left hand side, a first step replaces the gravitational constant with a constant relevant to the flows being studied. Even with a more relevant constant the prediction does not fit the data well, suggesting that Newton's value of 2 for the exponent of distance, based on physical principles, should be replaced by a value appropriate to the data and the exponents equal to 1 for the mass variables should be replaced by exponents that improve the fit of the prediction to the data.

Application of (1) with exponents and the constant term altered to fit economic data yields $\hat{X}_{ij} = aY_i^b E_j^c D_{ij}^\delta$ where $\hat{X}$ is the fitted value of X and parameters a, b, c, δ are estimated by best fit methods. The first application was to migration flows within the UK (Ravenstein, 1889) and the first application to trade flows was by Tinbergen (1962). The original form of the gravity model gives a close fit to observed flows: a scatter plot of $\hat{X}$ on observed X shows most points clustered close to a 45 degree line, much closer than most estimated economic relationships. Estimated b and c tend to be close to 1 and estimated δ close to -1 in many different applications. Despite this remarkable performance (most econometric relationships perform nowhere near as well), mainstream international economists were averse to a gravity model with no *economic* foundation.

Recent development of economic foundations for gravity modify the original form to be consistent with plausible economic structure. Structural gravity is now firmly embedded in the economic mainstream. An explosion of empirical research is the result, surveyed in Head and Mayer (2014). Structural gravity embedded in models of resource allocation across economic sectors has improved quantification of the consequences of globalization for trade patterns, the location of economic activity and the development of economies over time. The structural gravity model of economic interaction is useful due to a remarkably simple characterization of the distribution of economic activity across many origin and destination pairs, allowing a tractable model of the global interaction of many relatively large regions. The exposition below focuses on goods trade, but structural gravity models of migration and foreign direct investment are essentially the same. Other gravity models of economic interactions (portfolio investment, dissemination of ideas and culture) fit data well but still lack an economic foundation (Anderson, 2011).

The first bricks of the economic foundation of gravity are adding up constraints. Adding up constraints are not satisfied by the original form of gravity (1), seen as follows. The total of sales by origin i, Y_i must equal the sum of sales to each destination X_{ij}. Similarly, the total expenditure by destination j, E_j, must equal the sum of purchases from each origin i, X_{ij}. Formally, $Y_i = \sum_j X_{ij}$ and $E_j = \sum_i X_{ij}$. Finally, world sales must equal world expenditure, $Y = \sum_i Y_i = \sum_j E_j$. Equation (1) does not satisfy these elementary economic requirements without further restrictions. For example, remove the effect of distance in (1) by replacing its exponent 2 with the exponent 0, interpreted as moving goods in a frictionless world. Then the adding up constraints are satisfied if $G = 1/Y$. Using this new constraint in a frictionless world (1) becomes

$$X_{ij} = \frac{Y_i}{Y} E_j. \tag{2}$$

Equation (2) means that purchases in each destination j from each origin i are equal to total expenditure in j, E_j, times i's global share of expenditure Y_i/Y, a share that is common to all destinations j and equal to i's global sales share. In a frictionless world with perfect arbitrage of prices, each destination j would face the same price p_i for shipments from i, and this would be true for goods from every origin i. Then (2) is consistent with an economic model where expenditure shares are identical across destinations that face the same set of prices. This line of reasoning points to a theory of expenditure shares as the next layer of bricks in the foundation.

Expenditure share structure predicts how shares at each destination j vary with trade frictions from each origin i. A full development is needed because observed economic interactions are far from the frictionless benchmark (2). For example, if US national income is

25% of world income (approximately right) then the US should be spending about 75% of its national income on imported goods. This follows from using j as the US and i as the rest of the world, implying $Y_{ROW}/Y = 0.75 = X_{ROW,US}/E_{US}$. The US actually spends around 15% of income on imports. Frictions can explain why international trade is so low relative to the benchmark. A useful structural economic model of the effect of frictions follows from a suitably simple specification of the expenditure share structure.

The identical shares requirement is consistent with standard models of expenditure if (i) goods are different according to place of origin, (ii) consumer preferences or technologies in the case of intermediate goods are identical across destinations, and (iii) preferences or technologies are invariant to income and size of output respectively. The first economic foundation for the gravity model of trade (Anderson, 1979) assumed Constant Elasticity of Substitution (CES) expenditure structure. Expenditure shares X_{ij}/E_j are given by

$$\frac{X_{ij}}{E_j} = \beta_i \left(\frac{p_{ij}}{P_j} \right)^{1-\sigma} \tag{3}$$

where p_{ij}/P_j is the price of goods from i delivered to j relative to a price index of goods at j, $\beta_i > 0$ is a 'distribution' parameter (one for goods from each origin i and $\sum_i \beta_i = 1$ to ensure that shares sum to 1) and σ is the elasticity of substitution parameter. To accord with observed behavior, $\sigma > 1$, meaning that a rise in the relative price of good i in destination j will reduce i's expenditure share in j. Since expenditure on each origin's goods adds up to total expenditure, the sum over origins i of expenditure shares (3) is equal to 1, an equation solved for the CES price index

$$P_j = \left(\sum_i \beta_i p_{ij}^{1-\sigma} \right)^{1/(1-\sigma)} . \tag{4}$$

Trade frictions are assumed to raise the delivered price of good i in destination j by a constant 'iceberg melting' factor $t_{ij} > 1$, as if 1 unit departing the origin factory i yields $1/t_{ij} < 1$ units at destination j. Then the assumption of perfect arbitrage implies that $p_{ij} = p_i t_{ij}$, destination prices are raised by exactly enough to cover 'melting'.

The adding up condition for each seller i is $Y_i = \sum_j X_{ij} = \sum_j \beta_i \left(p_i t_{ij}/P_j \right)^{1-\sigma} E_j$. Solve the adding up condition for $\beta_i p_i^{1-\sigma} = Y_i / \sum_j (t_{ij}/P_j)^{1-\sigma} E_j$. Next, use the preceding equation to replace $\beta_i p_i^{1-\sigma}$ in (3). The result is the structural gravity model (5)-(7).

$$X_{ij} = \frac{Y_i E_j}{Y} \left(\frac{t_{ij}}{\Pi_i P_j} \right)^{1-\sigma} \tag{5}$$

where

$$\Pi_i^{1-\sigma} = \sum_j \left(\frac{t_{ij}}{P_j}\right)^{1-\sigma} \frac{E_j}{Y}. \tag{6}$$

Replace $\beta_i p_i^{1-\sigma}$ in the CES price index using (6) to yield.

$$P_j^{1-\sigma} = \sum_i \left(\frac{t_{ij}}{\Pi_i}\right)^{1-\sigma} \frac{Y_i}{Y}. \tag{7}$$

The new variable Π_i is an index of the outward trade frictions facing shippers from i. The price index P_j is rewritten using Π_i as expression (7), an index of inward trade frictions facing shipments to destination j. Anderson and van Wincoop (2003) coined the term multilateral resistance for these indexes of bilateral resistance. The dependence of multilateral resistances on frictions from i and to j on *all* links and *all* masses in the world economy is implicit in the system of equations (6)-(7).

The importance of multilateral resistance is illustrated by the large role multilateral resistance plays in driving Canadian provinces to trade so much more with each other than do US states (the border puzzle posed by McCallum, 1995). Take a pair of provinces and a pair of states chosen so that origin and destination size is the same ($Y_i^{US}/Y_i^{CA} = 1, E_j^{US}/E_j^{CA} = 1$) and they are the same distance apart. Then using (5) and canceling equal terms in numerator and denominator

$$\frac{X_{ij}^{CA}}{X_{ij}^{US}} = \left(\frac{\Pi_i^{CA} P_j^{CA}}{\Pi_i^{US} P_j^{US}}\right)^{\sigma-1}$$

Since the Canadian economy is 10% the size of the US economy, far more of the trade of Canadian provinces must cross the border and incur border frictions than is the case for US states. This increases Π_i^{CA} and P_j^{CA} for province pairs in Canada relative to Π_i^{US} and P_j^{US} for province pairs in the US (the solution to McCallum's border puzzle proposed by Anderson and van Wincoop, 2003).

The bilateral trade flow in equation (5) comprises two elements. $Y_i E_j/Y$ is the frictionless flow. $(t_{ij}/\Pi_i P_j)^{1-\sigma}$ is the systemic effect of trade frictions, the factor by which the bilateral flow differs from its hypothetical frictionless value. The bilateral friction of the original trade gravity model is replaced with a relative bilateral friction, where the denominator is the product of multilateral resistances. Intuition had suggested to some investigators that third party interactions must modify the simple bilateral form in (1) but no theory before Anderson (1979) was available to formalize this intuition. The initial borrowing from Newton offered no useful guidance since physics formula (1) is for the two body case (other masses are too distant to matter) and the knotty N-body problem yields nothing simple as a solution.

In this context (5) is an elegantly simple economic theory of the equilibrium distribution of given supplies of goods from many origins to buyers at many destinations who spread given expenditures across goods from many origins.

The assumption that products are differentiated by place of origin in the gravity model is itself given an economic foundation in the modern theory of monopolistic competition where firms offer unique varieties to buyers with CES demand structure for all varieties (the love of variety model). This point is developed by Bergstrand (1989). System (5)-(7) still describes the model but the explanation of the set of Y_is is enriched.

Two subsequent model building blocks generate CES-type expenditure shares and thus the same structure as (5)-(7). Heterogeneity of buyers or producers accounts for why potentially all origins serve all destinations. In the buyer heterogeneity model (Anderson, de Palma and Thisse, 1992) each individual has a single preferred variety with individuals differing according to a probability distribution with dispersion characterized by a parameter playing the role of $1-\sigma$. In the seller heterogeneity model (Eaton and Kortum, 2002) it is the dispersion parameter of the probability distribution of the labor productivity of sellers that plays the role of $1-\sigma$. An origin-specific location parameter of the productivity distribution plays the role of β_i. (The Eaton-Kortum model has seller heterogeneity at the level of countries, with identical atomistic competitive firms, so monopolistic competition is suppressed.) More general models still manage to approach the simplicity of (5)-(7) and extensions remain an area of active research.

Fitting the structural gravity model (5) to data is almost as straightforward as in the original gravity model. The multilateral resistance terms are commonly estimated using origin and destination fixed effects in standard regressions. [A full information alternative estimator using the full system (5)-(7) was applied by Anderson and van Wincoop (2003).] Two equivalent procedures are used. One is to divide both sides of (5) by $Y_i E_j/Y$ and estimate. The other is to estimate (5) as it is, and to recover the multilateral resistances (up to a normalization) from the estimated fixed effects χ_i and μ_j using $\Pi_i^{1-\sigma} = \chi_i/Y_i$ and $P_j^{1-\sigma} = \mu_j/E_j$. In the absence of data on Y_i and E_j only the combined χ_i and μ_j can be identified. The heart of the analysis is estimating bilateral frictions t_{ij}, done with loglinear functions of proxies. Proxies for friction such as bilateral distance are used along with measures for free trade agreements, common language, past colonial relationships, etc. More recent approaches have emphasized the importance of network structure variables [Rauch and Trindade (2002), Chaney (2014)]. Estimated gravity equations fit very well and yield precisely estimated parameters.

While structural gravity is a static model with parameters identified by cross section variation of bilateral trade flows, it is also applied on panel data. On the time dimension,

theory implies using origin and destination fixed effects that vary independently across time and these indeed have a lot of time variation, especially in the multilateral resistances (Anderson and Yotov, 2010). In contrast, there is little evidence for time variation in distance elasticities. Development of dynamic models of gravity is a challenging frontier of research.

A problem with the use of the model (5) is zeroes. All theoretical rationales discussed above imply that trade should be positive, even if very small. One explanation is that zeroes are due to observation error. When the flow is small, the observer may not see what 'should' be there. There is no doubt some truth to this explanation. It suggests econometric procedures to deal appropriately with the nature of the random error term (Santos-Silva and Tenreyro, 2006).

Two extensions of the gravity model give economic explanations for zeroes. One extension is fixed costs of serving a market (a portion of the trade iceberg breaks off and is lost before the berg begins melting on its trip to destination). A zero results when an origin is not productive enough to serve a destination. With heterogeneous firms in each sector and country, this model implies a selection effect when positive trade is observed: the most productive firms are able to pay the fixed export cost. A fall in bilateral trade costs acts on the volume of trade on the intensive margin as in (5) but also on a bilateral extensive margin through the entry of more origin firms into trade with the destination. Bilateral frictions inferred when controlling for selection appropriately are a combination of intensive and extensive margin changes. For a successful approach to estimating such models see Helpman, Melitz and Rubinstein (2008). A problem with application is the difficulty of finding proxies for fixed costs that are not also proxies for the t_{ij}s.

The alternative explanation for zeroes arises when the buyers at j have willingness to pay for even a minimal amount (a choke price) that is less than the potential delivered price $p_{ij} = p_i t_{ij}$. Demand would be zero whether fixed costs are negligible or not. Non-CES demand structure is required for finite choke prices because CES share (3) with $\sigma > 1$ is positive for finite prices. For a (translog) example of expenditure shares with finite choke prices that yields a very tractable gravity model see Novy (2013).

The model (5)-(7) refers to goods in a single sector, which could be at any level of aggregation that is to be treated. The gravity model of distribution can be nested in any general equilibrium model of the allocation of resources between sectors in each location. Thus it is consistent with most such models developed in mainstream economics.

Gravity nested within general equilibrium models permits a number of applications measuring the effects of policy changes, actual or proposed. Head and Mayer (2014) classify such policy change analyses in order of the degree of interaction considered. Partial Trade Impact evaluates the effect of a change in a t_{ij} on X_{ij} holding all else equal in (5). Modular

Trade Impact evaluates the effect of a change in t_{ij} on all the trade flows and multilateral resistances using (5)-(7) with all Y_i and E_j held constant. General Equilibrium Trade Impact includes the response of Y_i and E_j to changes flowing from MTI induced by the original change in t_{ij}. The last step requires a model to determine Y_i and E_j for all countries and sectors. Structural gravity is the trade distribution component of the full model.

The simplest full general equilibrium model (Anderson and van Wincoop, 2003) determines for each origin i the factory gate price p_i of i's physical endowment y_i, hence $Y_i = p_i y_i$. The model is closed using $Y_i = E_i$, the balance of payments constraint of each origin i with the rest of the world. Shifts in trade costs alter the factory gate prices p_i of all origins i and the full model is solved for the new p_is and multilateral resistances. (A multi-sector endowments model is developed and applied in Anderson and Yotov, 2016.)

An attractive and simple alternative full general equilibrium model is due to Eaton and Kortum (2002). Production is Ricardian (labor is the only factor of production) and countries draw labor productivities from a Frechet probability distribution. In equilibrium each country specializes in a range of sectors and earns and spends income equal to the wage bill $w_i L_i = Y_i = E_i$. Full general equilibrium analysis includes the effect of changes in t_{ij} on the equilibrium wage w_i, shifting $Y_i = E_i$. The shifts are the same as in the endowments model applied in Anderson and van Wincoop (2003), but the interpretation differs. The Eaton-Kortum model features a richer model of Y_i in the sense that the composition of output of i's good Y_i is explained and all substitution is on the extensive rather than intensive margin. Costinot, Donaldson and Komunjer (2012) extend the Eaton-Kortum model to multiple sectors in contrast to the effectively one good composite of the original model. The extended model features inter-sectoral resource allocation in a particularly simple form. Equilibrium specialization occurs in a range of sub-sectors within each sector. The extended version implies that change in trade costs induces changes in the proportions of labor devoted to each sector as well as changes in the range of sub-sectors within each sector. Much recent research has used the Ricardian full general equilibrium model to explore such important policy changes as NAFTA (Caliendo and Parro, 2015).

References

[1] Anderson, J. E. (1979), "A Theoretical Foundation for the Gravity Equation," *American Economic Review*, 69(1), 106-16.

[2] Anderson, J. E. and E. van Wincoop (2003), "Gravity with Gravitas: A Solution to the Border Puzzle," *American Economic Review*, 93, pp. 170-192.

[3] Anderson, J. E. and Y. V. Yotov. (2010). "The Changing Incidence of Geography," *American Economic Review*, 100, 2157-86.

[4] Anderson, J.E. (2011), "The Gravity Model", *Annual Review of Economics, 3, 133-60.*

[5] Anderson, J.E. and Y.V. Yotov (2016). "Terms of Trade and Global Efficiency Effects of Free Trade Agreements, 1990-2002", *Journal of International Economics*, 98, 279-98.

[6] , Anderson, S.P., A. de Palma and J-F. Thisse (1992), *Discrete Choice Theory of Product Differentiation*, Cambridge: MIT Press.

[7] Bergstrand, J. H. (1989), "The Generalized Gravity Equation, Monopolistic Competition, and the Factor-Proportions Theory in International Trade," *Review of Economics and Statistics*, 71(1), 143-153.

[8] Caliendo, L. and F. Parro (2015), "Estimates of the Trade and Welfare Effects of NAFTA", *Review of Economic Studies*, 82(1), 1-44.

[9] Chaney, T. (2014), "The Network Structure of International Trade", *American Economic Review*, 104(11), 3600-34.

[10] Costinot, A., D. Donaldson and I. Komunjer (2012), "What Goods Do Countries Trade? A Quantitative Exploration of Ricardo's Ideas", *Review of Economic Studies*, 79, 581-608.

[11] Eaton, J. and S. Kortum (2002), "Technology, Geography, and Trade," *Econometrica*, 70 (5), 1741-1779.

[12] Head, K. and T. Mayer (2014), "Gravity Equations: Workhorse, Toolkit, and Cookbook." Chapter 3 in the *Handbook of International Economics, Vol. 4*, eds. Gita Gopinath, Elhanan Helpman, and Kenneth S. Rogoff, Elsevier Ltd., Oxford.

[13] Helpman, E., M. J. Melitz and Y. Rubinstein (2008), "Estimating Trade Flows: Trading Partners and Trading Volumes", Harvard University, *Quarterly Journal of Economics*, 123: 441-487.

[14] McCallum, J. (1995), "National Borders Matter: Canada-U.S. Regional Trade Patterns," *American Economic Review*, 1995, 85(3), pp. 615-623.

[15] Novy, D. (2013) "International Trade and Monopolistic Competition without CES: Estimating Translog Gravity", *Journal of International Economics*, 89(2), 271-82.

[16] Rauch, J.E. and V. Trindade (2002), "Ethnic Chinese Networks in International Trade", *Review of Economics and Statistics*, 84(1), 116-30.

[17] Ravenstein, E. G., 1889, "The Laws of Migration." *Journal of the Royal Statistical Society*, Vol. 52, No. 2. (June, 1889), pp. 241305.

[18] Santos Silva, J.M.C. and S. Tenreyro (2006), "The Log of Gravity," *The Review of Economics and Statistics*, vol. 88(4), pages 641-658.

[19] Tinbergen, J. 1(962). *Shaping the World Economy: Suggestions for an International Economic Policy*. New York: The Twentieth Century Fund.