

Gravity for Beginners*

Keith Head[†]

February 5, 2003

Contents

1	The Basic Gravity Equation	2
1.1	Origins: Newton's Apple	2
1.2	Economists Discover Gravity	2
1.3	Economic Explanations for Gravity	3
2	Estimation of the Gravity Equation	4
2.1	Economic Mass	4
2.2	Distance	5
2.3	Remoteness	8
3	"Augmenting" the Gravity Equation	9
3.1	Income per Capita	9
3.2	Adjacency	9
3.3	Common Language and Colonial Links	9
3.4	Border Effects	10
4	Evaluating Trade-Creating Policies	11
4.1	Free Trade Agreements	11
4.2	Monetary Agreements	11

*Original Version: October, 2000. This version prepared for UBC Econ 590a students, January 2003. This is a work in progress and I welcome comments and suggestions. The most up-to-date version is available at economics.ca/keith/gravity.pdf

[†]Faculty of Commerce, University of British Columbia, 2053 Main Mall, Vancouver, BC, V6T1Z2, Canada. Tel: (604)822-8492, Fax: (604)822-8477, Email:keith.head@ubc.ca

1 The Basic Gravity Equation

The gravity equation is a popular formulation for statistical analyses of bilateral flows between different geographical entities. In this paper, I provide an overview of the development and use of this equation. I also include some practical tips for researchers who want to use the equation in their own work.

1.1 Origins: Newton’s Apple

In 1687, Newton proposed the “Law of Universal Gravitation.” It held that the attractive force between two objects i and j is given by

$$F_{ij} = G \frac{M_i M_j}{D_{ij}^2}, \quad (1)$$

where notation is defined as follows

- F_{ij} is the attractive force.
- M_i and M_j are the masses.
- D_{ij} is the distance between the two objects.
- G is a gravitational constant depending on the units of measurement for mass and force.

1.2 Economists Discover Gravity

In 1962 Jan Tinbergen proposed that roughly the same functional form could be applied to international trade flows. However, it has since been applied to a whole range of what we might call “social interactions” including migration, tourism, and foreign direct investment. This general gravity law for social interaction may be expressed in roughly the same notation:

$$F_{ij} = G \frac{M_i^\alpha M_j^\beta}{D_{ij}^\theta}, \quad (2)$$

where notation is defined as follows

- F_{ij} is the “flow” from origin i to destination j . Alternatively, let \tilde{F}_{ij} represents total volume of interactions between i and j (i.e. the sum of the flows in both directions: $\tilde{F}_{ij} = F_{ij} + F_{ji}$).
- M_i and M_j are the relevant economic sizes of the two locations.
 - If F is measured as a monetary flow (e.g. export values), then M is usually the gross domestic product (GDP) or gross national income (GNI, formerly GNP) of each location.
 - For flows of people, it is more natural to measure M with the populations.
- D_{ij} is the distance between the locations (usually measured center to center).

Note that we return to Newton’s Law (equation 1) if $\alpha = \beta = 1$ and $\theta = 2$.

1.3 Economic Explanations for Gravity

The gravity equation can be thought of as a kind of short-hand representation of supply and demand forces. If country i is the origin, then M_i represents the total amount it is willing to *supply* to all customers. Meanwhile M_j represents the total amount destination j *demands*. Distance acts as a sort of tax “wedge,” imposing trade costs, and resulting in lower equilibrium trade flows.

More recently (starting with Anderson, 1979) there have been several attempts to derive the gravity equation formally. Here I sketch a derivation.

Let M_j be the amount of income country j spends on all goods from an source i . Let s_{ij} be the share of M_j spent on goods from country i . Then $F_{ij} = s_{ij}M_j$. What do we know about s_{ij} ?

1. It must lie between 0 and 1.
2. It should increase if i produces a wide variety of goods (large n_i) and/or goods perceived to be of high quality (large μ_i).
3. It should decrease due to trade barriers such as distance, D_{ij} .

In light of these arguments we suggest

$$s_{ij} = \frac{g(\mu_i, n_i, D_{ij})}{\sum_{\ell} g(\mu_{\ell}, n_{\ell}, D_{\ell j})},$$

where the $g(\cdot)$ function should be increasing in its first two arguments and decreasing in distance for all $s_{ij} > 0$.

To move forward, we need a specific form for $g(\cdot)$. One approach (taken by Bergstrand) uses the Dixit and Stiglitz model of monopolistic competition between differentiated but symmetric firms. This model sets $\mu_i = 1$ and makes n_i proportional to M_i . A second approach (due to Anderson) assumes a single good from each country, $n_i = 1$, but allows the preference parameter μ_i to vary across countries subject to the constraint of market-clearing. μ_i differ in such a way as to also be proportional to the size of the economy, M_i . Both let trade costs be a power function of distance. I prefer the monopolistic competition approach because it seems more natural to endogenize the number of varieties, n_i , than to endogenize the preference parameter. Later we will see that there are some empirical specifications that are valid under either approach.

Allowing both n and μ to vary across countries, let $g(n_i, \mu_i) = \sum_{v=1}^{n_i} (p_{ijv}/\mu_{ijv})^{1-\sigma}$, where v indexes particular varieties that are substitutable with an elasticity of substitution given by σ . If the goods from the same country are differentiated but of the same average quality and subject to the same transport costs, then we can drop the v subscripts and set $g() = n_i(p_{ij}/\mu_{ij})^{1-\sigma}$.

The next step is to relate the delivered (quality-adjusted) price to the price in the origin country and transportation costs between origin and destination. We assume the following relationship:

$$p_{ij}/\mu_{ij} = (p_i/\mu_i)D_{ij}^{\delta}.$$

The origin price, p_i , is often referred to as the free-on-board or fob price. We will postpone a complete discussion of the justification for this equation but note that it allows for both the effect of distance-based freight charges on the delivered price and for distance effects on perceived quality (due to mundane causes such as damage in transit or, more speculatively, to culture-based biases that are correlated with distance).

In the basic gravity equation, we assume away price differences.¹ Note that this is not quite as unrealistic as it at first seems—we require only that fob prices vary proportionally to the quality of the export country’s products, i.e. that $p_i/\mu_i \approx k$.

The number of varieties in each country n_i is not something we can hope to observe directly. Rather we take advantage of a property of the Dixit-Stiglitz model: namely, all firms are the same size. In that case, $n_i = M_i/q$ where q is firm size. Imposing these last assumptions, defining $\theta \equiv \delta(\sigma - 1) \geq 0$, we obtain $g() = M_i D^{-\theta}/(qk^{\sigma-1})$. This implies market shares for exporter i in country j of

$$s_{ij} = M_i D_{ij}^{-\theta} R_j,$$

where $R_j = 1/(\sum_{\ell} M_{\ell} D_{\ell j}^{-\theta})$. After substituting and rearranging we obtain a result that is very close to what we had sought for:

$$F_{ij} = R_j \frac{M_i M_j}{D_{ij}^{\theta}}. \quad (3)$$

The main difference is that now the term R_j replaces the “gravitational constant,” G . We will discuss the interpretation of that term in the next section.

Before that note what happens in a “frictionless” world, i.e. one in which $\theta = 0$. Then $R_j = 1/\sum_{\ell} M_{\ell} = 1/M_w$ and $F_{ij}^* = M_i M_j/M_w$ (the w subscript stands for “world”).

2 Estimation of the Gravity Equation

The multiplicative nature of the gravity equation means that we can take natural logs and obtain a linear relationship between log trade flows and the logged economy sizes and distances:

$$\ln F_{ij} = \alpha \ln M_i + \beta \ln M_j - \theta \ln D_{ij} + \rho \ln R_j + \epsilon_{ij}. \quad (4)$$

The inclusion of the error term ϵ_{ij} delivers an equation that can be estimated by ordinary least squares regression. If our derivations in the earlier section are correct, we would expect to estimate $\alpha = \beta = \rho = 1$.

2.1 Economic Mass

The economic sizes of the exporting and importing countries, M_i and M_j , are usually measured with gross domestic product. The estimated coefficients are usually close to

¹Recently developed methods of analyzing bilateral trade do not require this assumption. See Feenstra (*Scottish Journal of Political Economy*, 2002).

the predicted value of one. However, it is not unusual to obtain values ranging anywhere between 0.7 and 1.1.

Note that the theory we used to derive the gravity equation predicts coefficients of one. Indeed, we lack an interpretation for coefficients different from one. There are further problems with including the $\ln M_i$ and $\ln M_j$ as regressors. First, they tend to inflate the R^2 of the regressions since it is hard to imagine a world in which big countries don't trade more in absolute terms. Second, since exports and imports are part of GDP, there is a built accounting relationship between the F_{ij} and M_i and M_j . Some studies have tried to deal with this simultaneity by using instrumental variables for GNP (such as population). A simpler solution is just to impose the theoretical prediction of unitary elasticities. This implies that we pass the income terms over to the left hand side. Subtracting $\ln M_i + \ln M_j - \ln M_w$ from both sides of (4), we obtain

$$\ln(F_{ij}/F_{ij}^*) = \ln M_w + \rho \ln R_j - \theta \ln D_{ij} + \epsilon_{ij}. \quad (5)$$

The dependent variable measures the deviation of actual trade flows from the “frictionless” ideal. The sum of the first two terms on the right-hand side will be estimated as the regression's constant; that is variation in R_j is shoved into the error term. There are two test statistics that one can examine to see if the data statistically reject the frictionless idea. One is the t-stat on the constant. The other is the t-stat on $\hat{\theta}$.

2.2 Distance

Distance is almost always measured using the “great circle” formula. This formula approximates the shape of the earth as a sphere and calculates the minimum distance along the surface.

Tip: To calculate great circle distances you need the longitude and latitude of the capitol or “economic center” of each economy in the study. The apply the following formula to obtain the distance measure in miles:

$$D_{ij} = 3962.6 \arccos([\sin(Y_i) \cdot \sin(Y_j)] + [\cos(Y_i) \cdot \cos(Y_j) \cdot \cos(X_i - X_j)]), \quad (6)$$

where X is longitude in degrees multiplied by 57.3 to convert it to radians and Y is latitude multiplied by -57.3 (assuming it is measured in degrees West).

Even for air travel, great circle distances probably underestimate true distances since they do not take into account that most flights avoid the North Pole. For maritime travel, they do not take into account indirect routes mandated by land and ice barriers. In addition, many air and sea routes are shaped by economic considerations such as “hub economies.” Furthermore international shipping cartels often set freight costs that bear little relationship to distance travelled. Also, the costs of packaging, loading and unloading, seem to be primarily fixed costs that do not vary with distance. Taken together, these considerations suggest that distance should matter very little for trade.

While he have many *ex-ante* reasons to expect little relationship between trade and distance, the facts say that distance dramatically impedes trade. Together with Anne-Celia Disdier of the University of Paris, I have been conducting a meta-analysis of gravity equation distance estimates from 595 regressions reported in about 35 papers. The samples ranged from 1928 to 1995. The trading partners were mainly nations though some results for the trade of Canada’s provinces were included as well. The average distance effect turns out to be $\hat{\theta} = 0.94$. This means that a doubling of distance will decrease trade by one half.

Leamer and Levinsohn’s (1994) survey of the empirical evidence on international trade offers the identification of distance effects on bilateral trade as one of the “clearest and most robust empirical findings in economics.”²

They asked “Why don’t trade economists ‘admit’ the effect of distance into their thinking? One [answer] is that human beings are not disposed toward processing numbers, and empirical results will remain unpersuasive if not accompanied by a graph.” They showed Germany’s trade but I will stay closer to home, showing trade by Canadian provinces and US states.

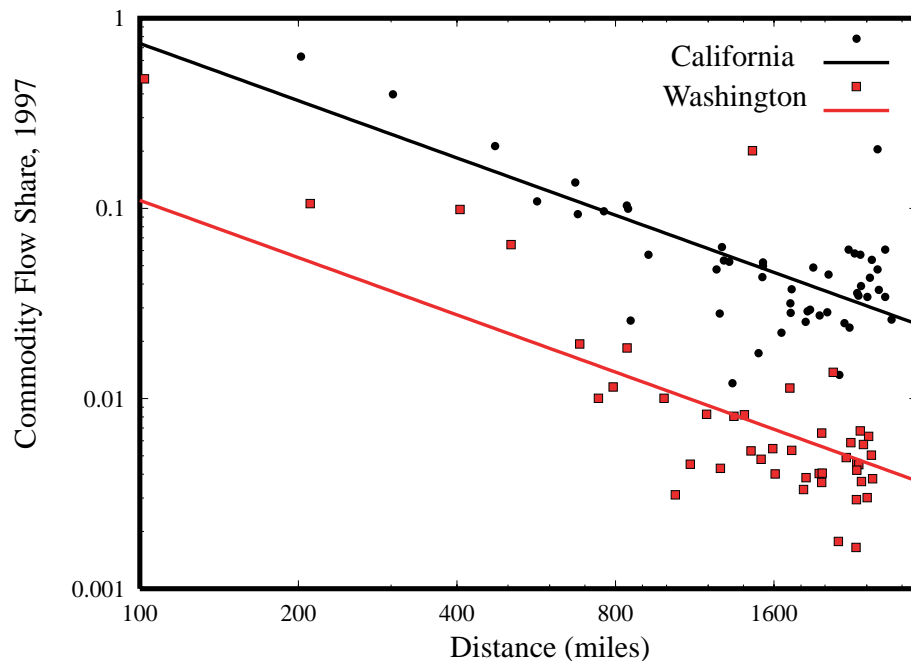
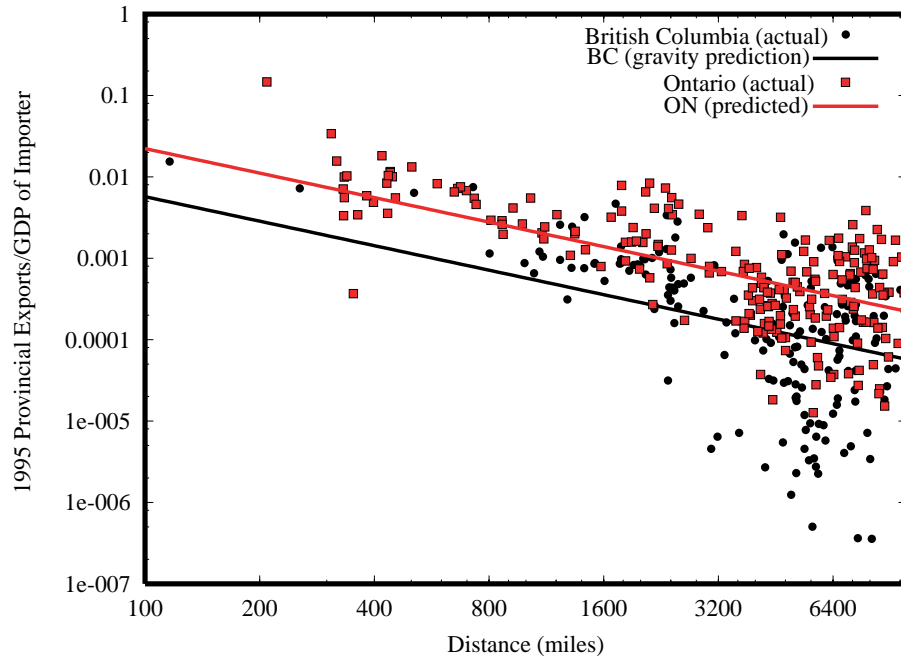
The graphical method is a scatterplot of F_{1j}/M_j (exporter 1’s share of market j) on the vertical axis against D_{1j} on the horizontal axis. Both axes are shown in “log” scale. Thus each space between tic marks raises the variable by some factor. For the vertical axis the factor is 10 while it is 2 for the horizontal axis. A line through the means of the data with a slope of -1 (in log scale) is also shown as a reference. It is often revealing to contrast exporter 1’s performance with that of some comparable economy ($i = 2$). The gap between the intercepts should be approximately equal to the relative sizes of the two exporters, i.e. M_1/M_2 . Alternatively, one might estimate (using ordinary least squares, for instance) and graph lines that best fit the data for each exporter.

Why does distance matter so much? Economists have offered four major explanations:

1. Distance is a proxy for transport costs. David Hummels has argued that shipping costs (freight charges and marine insurance) can go a long way towards explaining why distance matters.
2. Distance indicates the *time elapsed during shipment*. For perishable goods the probability of surviving intact is a decreasing function of time in transit. *Perishability* may be interpreted quite broadly to include the following risks:
 - (a) Damage or loss of the good due to weather or mishandling (e.g. ship sinks in a storm).
 - (b) Decomposition and spoiling of organic materials (e.g. maggot infestation).
 - (c) Loss of the market (the intended purchaser becomes unwilling or unable to make payment).

²They assert that the typical distance effect is 0.6.

Figure 1: Trade is Inversely Proportionate to Distance



3. Synchronization costs. When factories combine multiple inputs in the production process, they need those inputs to arrive in time or bottlenecks emerge. One possibility is to use warehouses to keep inventories of each input but this approach suffers from various drawbacks (land costs, technological obsolescence, fashion changes, and low pressures for quality control). Sourcing inputs from nearby lowers synchronization costs.
4. Communication costs. According to Paul Krugman, distance “proxies for the possibilities of personal contact between managers, customers, and so on; that much business depends on the ability to exchange more information, of a less formal kind, than can be sent over a wire.”
5. Transaction costs. Distance may also be correlated with the costs of searching for trading opportunities and the establishment of trust between potential trading partners.
6. “Cultural distance.” It may also be that greater geographic distances are correlated with larger cultural differences. Cultural differences can impeded trade in many ways such as inhibiting communication, generating misunderstandings, clashes in negotiation styles, etc.

2.3 Remoteness

Until recently, most papers implicitly assumed that R_j is constant across countries and therefore becomes the intercept in the regression equation. However, R_j is important because it measures each importer’s set of alternatives. Countries with many nearby sources of goods, i.e. those with low values of R_j , will import less from each particular source.

A few studies have included variables like R_j and referred to them as “remoteness.” However some of these measures differ from the theoretically correct R_j in ways that may be problematic. For instance, Helliwell (1998) measures remoteness as $REM_j = \sum_{\ell} D_{\ell j} / M_{\ell}$. This measure causes remoteness to be very large if it includes distant (high $D_{\ell j}$) but tiny (low M_{ℓ}) countries. Since the previous literature usually finds $\theta \approx 1$, a better measure of remoteness is $1 / (\sum_{\ell} M_{\ell} / D_{\ell j})$. In this measure the size of very distant countries becomes irrelevant.

The importance of remoteness in actual trade patterns can be illustrated by comparing trade between Australia and New Zealand with trade between Austria and Portugal. The distance between each pair’s major city is approximately the same: Lisbon–Vienna and Auckland–Canberra both happen to be 1430 miles apart. Furthermore the product of their GDP’s are similar (Australia–New Zealand is 20% smaller). Hence, omitting remoteness, the gravity equation would predict that Austria–Portugal trade would be slightly larger. In fact, however, in 1993 Australia–New Zealand trade was nine times greater than Austria–Portugal Trade.

Tip: The remoteness measure includes M_i/D_{ii} in its summation requiring us to specify a country's distance from itself, D_{ii} . For reasons provided in Head and Mayer (2000), I believe a good approximation for this “internal distance” is provided by the square root of the country's area multiplied by about 0.4.

3 “Augmenting” the Gravity Equation

Gravity equations do a pretty good job at explaining trade with just the size of the economies and their distances. However, there is a huge amount of variation in trade they cannot explain. Most authors add a few other variables with less theoretical justification, usually because past experience has shown that they “work.” In the next subsections I discuss the most commonly included of these variables.

3.1 Income per Capita

Many authors estimate gravity equations with the log of per-capita incomes ($\ln M/\text{POP}$) of the exporting and importing countries included as well as the log of aggregate incomes ($\ln M$).

The idea behind this appears to be that higher income countries trade more in general. One cause might be superior transportation infrastructure (roads to the interior, container ports, airports, etc.). High income countries probably have lower tariffs. A countervailing effect is that high income countries tend to be more service-oriented, leading to lower trade in merchandise for a given level of GDP.

Estimated coefficients on the log of per-capita GDP display considerable variation across studies, ranging as low as 0.2 and as high as 1.

3.2 Adjacency

Adjacent, or contiguous, countries share a border. Many studies include a dummy variable to identify such pairs.

The estimated coefficient usually lies in the vicinity of 0.5, suggesting that trade is about 65% higher as a result of sharing a border. It is not clear why adjacency should matter if one is already controlling for distance. Perhaps center-to-center distance overstates the effective distance because neighboring countries often engage in large volumes of border trade. Examples of this phenomenon include Windsor–Detroit, Tijuana–San Diego, and Hongkong–Shenzhen.

3.3 Common Language and Colonial Links

Recall that one explanation for the trade impeding effects of distance was transaction costs caused by inability to communicate and cultural differences. If so, we would expect that countries that speak the same language would trade more. The evidence strongly

confirms this proposition. Two countries that speak the same language will trade twice to three times as much as pairs that do not share a common language.

Part of the reason for this common language effect is probably the shared history that caused the two countries to share a language. Indeed, measures of colonial links also are positively correlated with trade. Including them as controls reduces the language effect somewhat but it remains quite strong.

3.4 Border Effects

A recent literature initiated by John McCallum's 1995 *American Economic Review* article investigates whether national borders still matter for trade.

In *The Borderless World*, Kenichi Ohmae of McKinsey asserted

“National borders have effectively disappeared and, along with them, the economic logic that made them useful lines of demarcation in the first place.”

McCallum's examination of the trade patterns of Canadian provinces countered that borders must matter very much because the typical Canadian province trades 20 times more with other provinces than with American states of a given size and distance.

Perhaps the best way to see how this sort of calculation would arise is from considering Ontario's shipments to British Columbia and Washington state. The distances involved are essentially the same but one case involves crossing a border and the other does not.

If borders were irrelevant, the gravity equation would predict that exports to BC should be 0.6 of exports to Washington because that is the ratio of the two states' economies. However, BC actually receives 12.6 times more goods from Ontario than does Washington. Thus the border effect, defined as the actual trade ratio divided by the predicted trade ratio, is $12.6/0.6 = 21!$

Since the Canada-US Free Trade Agreement was implemented, cross-border trade has grown dramatically (around 60%) and border effects have fallen to about 12 on average for Canadian trade.

Border effects can also be calculated without the “intra-national” trade flows that are only available for a few countries. This method, developed by Shang Jin Wei requires estimates of each country's distance to itself. Head and Mayer developed a way to measure internal and external distances in a consistent manner and applied it to European trade. They also found high border effects.

Why do borders matter? One approach is to question the methods and the measurements. Another approach is to accept the result and argue that it points to the great importance of national institutions (legal, monetary, social) that promote trade. The dust has not settled on this debate.

I believe that trade depends on networks of connected firms. These networks formed over time when borders and distance imposed higher costs because both tariffs and transport costs were higher. Members of networks focused on building local relationships. These strong local ties generate trade. Thus I think border and distance effects are large for the same reasons.

4 Evaluating Trade-Creating Policies

Countries often enter into agreements with intent of facilitating bilateral trade. Do such agreements work?

4.1 Free Trade Agreements

Regional trade liberalizing agreements like Europe's common market and North America's free trade agreements have proliferated in the last 20 years and one of the primary uses of gravity equations has been to evaluate them.

On average FTAs seem to raise trade by around 50%. However, a recent study by Frankel and Rose (National Bureau of Economic Research Working Paper 7857) finds that FTAs lead to a *tripling* of trade between partners.

4.2 Monetary Agreements

Studies of how exchange rate volatility affects trade have obtained mixed results. One recent study, by Frankel and Rose, finds that countries that share a common currency, such as the US and Panama, trade three times more with each other than one would expect.

This effect is surprisingly large and perhaps implausible as a general rule. For instance, I find it very unlikely that the adoption of the EURO by 11 countries in Europe will cause trade between them to triple!