Exploring differences in FJ estimates and AA estimates of distance costs

Farid Farrokhi and David Jinkins

June 2016

Our baseline estimates for distance costs vary substantially from those estimated using the same data and methodology developed in Allen and Arkolakis (2014). The authors of that study were kind enough to provide a main Matlab estimation file to us soon after we began working on this project. We downloaded and cleaned the input files ourselves from the same sources as the previous study, and wrote several small functions which were omitted from the original code provided to us. Thus, we were quite surprised to find that our estimates vary from the original not only quantitatively, but qualitatively as well. In particular, we estimate a much higher variable cost for water and air than do Allen and Arkolakis. The rank of our estimates for fixed costs are the same as Allen and Arkolakis. The rank of our variable costs are the same, except for water transport which we estimate to be the most expensive form of transport. See Table columns (1) and (2) below.

Allen and Arkolakis later released full replication code for their paper. Below we compare our estimates in more detail. While we were not able to make our results match theirs exactly, we can find several differences which explain part of the gap:

1. The input value of truck transport in Allen and Arkolakis is exactly twice what is reported in the 2007 Commodity Flow Survey data we downloaded. It appears this is a bug. Column (3) in Table shows that this does not drive the difference between our estimates. We doubled the value of truck transport in our data, and our estimates remained qualitatively the same. We speculate the estimates do not change much because road transport is already the dominant form of shipment in the domestic United States. Increasing the dominance does not have a qualitative effect on the estimates.
2. In the Commodity Flow Survey data, pure water transport and pure rail transport are separated from transport via water and truck and rail and truck.[1] Allen and Arkolakis use only pure water and rail transport in their input data, whereas we count both categories. In Column (4) we run our code using only pure water and pure rail figures. Our estimates of air and rail transport then move substantially closer to the numbers estimated in Allen Arkolakis, although there is still quite a large difference.
3. The maps we use to compute distances for road and rail are nearly exactly the same as those in Allen and Arkolakis. The water maps differ, however. We allow (cheap) water transport only along common shipping routes in the ocean. Allen and Arkolakis allow water transport along any part of the ocean. Because of different coordinate systems hardwired into the code, it is hard to directly analyze how much this factors in the analysis, although we argue below that map differences may be a substantial contributer to our differing estimates.[2]

[1] Explicit category definitions for CFS data can be found here: www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/commodity_flow_survey/def_terms/index.html

[2] In addition to these differences, the coordinates used by Allen and Arkolakis for CFS areas appear to be rounded as is typical when exporting data from Stata. Their coordinates

4. Allen and Arkolakis estimate the parameters for ther shippers' discrete choice of mode of transport minimizing the following loss function. Let $\varepsilon(\beta)^m_{od}$ be the difference between the predicted and observed fraction of shipments of mode m between origin o and destination d evaluated at parameter vector β. Let N be the total number of bilateral pairs:

$$\sum_m |\frac{1}{N} \sum_o \sum_d \varepsilon(\beta)^m_{od}|$$

Our algorithm minimizes the squared residual:

$$\sum_m \sum_o \sum_d (\varepsilon(\beta)^m_{od})^2$$

The two loss functions deliver qualitatively different solutions to the problem both in Allen and Arkolakis' code and in ours. We show how this affects our results by first running Allen and Arkolakis' baseline code using our data, and then running their code using our data as well as our minimization algorithm.[3] In Column (5) we see results that are much more similar to Allen and Arkolakis than in the baseline. Using our algorithm in Column (6), we move the results much closer to our baseline. Finally, in Column (7) we run Allen and Arkolakis' code with their data and with our minimization algorithm.[4] Here we see substantial convergence toward our baseline estimates. One caveat is that air transport variable costs become even smaller than those estimated in Allen and Arkolakis. We conclude that one of the main drivers of the difference in our estimates may be the loss function in the estimation algorithm for the mode of transport problem. Map differences may also be playing an important role, driving differences in our air cost estimates, along with the choice of input data for water and rail described above.

As a final comment, the results in our paper continue to be based on our baseline estimates. We believe that it is proper to count water and truck as a water shipment and rail and truck as a rail shipment since around 50% of the value of rail shipments in our data also involve trucking, and around 30% of the value of water shipments involve trucking. We also believe that forcing water shipments to be along trade routes to ports is also a realistic assumption, since loading and unloading cargo without a port is costly. Finally, we prefer the smoother least squares loss function for estimating the relative shipment costs by mode. In sum, although we have not been able to understand exactly what causes the difference between our estimates, we believe we have a few leads which could be investigated further. Since accounting for the differences is not the primary goal of our study, we leave the discussion here.

1. AA Baseline
2. FJ baseline[5]
3. FJ Double Truck
4. FJ Double Truck / Only water,only rail

---

range from -2.2 to 2.1 million on the x-axis and from -1.2 to 1.4 million on the y-axis. All coordinates with absolute value above one million have the final five digits rounded to zero. As we were unable to precisely link our data sets, the extent that this affects estimates is not clear.

[3]Because our input data on demographics was not in the same format as in Allen and Arkolakis, in our final gravity regressions we altered Allen and Arkolakis' code to omit demographic similarity between locations. This may be driving some of the results we report here

[4]Demographic similarity variables are included in the gravity regression here.

[5]There was a small bug which we fixed *after* running the AA comparisons below. In particular we were missing several small locations in our estimation code. Adding these locations did not change our estimates much, which can be seen by comparing column (8) and column (2).

Transport Type	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Road var	0.5636	0.4065	0.4414	0.3094	0.4235	0.3944	0.5445	0.4430
Rail var	0.1434	0.3661	0.3935	0.2965	0.0016	0.3414	0.4405	0.3986
Water var	0.0779	0.6265	0.6439	0.4556	0.0454	0.4597	0.7915	0.6806
Air var	0.0026	0.2233	0.3187	0.1050	0.0506	0.0000	0.0000	0.3157
Rail fixed	0.4219	0.2995	0.3274	0.3308	0.3687	0.4178	0.5246	0.3106
Water fixed	0.5407	0.3428	0.3600	0.3907	0.3887	0.5938	0.6037	0.3384
Air fixed	0.5734	0.5254	0.4963	0.4935	0.3630	0.6313	0.8323	0.4904

Table 1: Comparing distance estimates in several models

5. AA code FJ data
6. AA code FJ data FJ mode obj
7. AA code/data FJ mode obj
8. FJ baseline before bug fix (omitting several small locations)

References

Allen, T. and Arkolakis, C. (2014). Trade and the topography of the spatial economy. *The Quarterly Journal of Economics*, 129(3):1085–1140.