Comments on "Monte Carlo Appraisals
of Gravity Model Specifications"
by Anderson, Ferrantino, and Schaefer

Andrew K. Rose
UC Berkeley, CEPR and NBER

Summary

The authors use Monte Carlo methods to show that gravity models can generate unreliable results. For instance, a simple data generation process without, say, a currency union effect, delivers data that can be incorrectly estimated and shown to deliver an apparently significant currency union coefficient.

Analysis

What's Good? The idea of this paper seems worth pursuing, and I encourage the authors to pursue it further. I don't know of anything comparable in the literature, and there's certainly a lot of gravity estimation around, so that the topic seems worthwhile.

What's Bad? The execution of this project is way below what it should be. Some of it has to do with the exposition; I simply don't understand a lot of what was actually done. Some of it has to do with the actual execution of the statistics.

I suggest you pursue this idea, but in a much more careful, clear, focused way.

Major Points

The paper should be much more clear. By the time one gets to page 4, one should have a very clear idea of what the authors have done and what they've found. More generally, clear directions along the way ("In this section we do x and find y") would be helpful. A clearer and stronger conclusion would help.

The most important dimension of poor expositional quality is technical. You simply don't provide enough details to allow the reader to understand what you do; the material around p6 is inadequate. You estimate a model; do you then use the coefficient estimates and regenerate artificial regressand values using a suitably parameterized residual (generated from a random number generator)? If not, what do you do? If so, a) say so; b) elaborate more on the properties of the residuals. They should surely be dependent across time (the residual for Canada-US trade in 1990 has something to do with that for Canada-US trade in 1985), and countries (Canada-US trade in 1990 has something to do with Canada-Mexico trade in 1990). How do you handle such issues? Also, are the original residuals close to normal?

I'm not sure how compelling your results actually are. Certainly I can't tell easily from either the text or tables. You seem to discuss your results very casually, and I don't understand quite why. This should be very clear in the next version.

If you're finding false positives for variables that really aren't there, that means there must be some correlation between these variables and the others. It's almost surly got a simple correlation with the regressand, and probably with the other regressors too. This is surely worth investigation.

It seems like your major issue is that there's tons of data, and therefore lots of scope for false positives. Fair enough. But this is true of many areas of research (labor economics, asset-pricing, and empirical industrial organization are perhaps the most obvious examples, since there are many people, stock prices, and firms in the world – far more than country-pair dyads). Your paper should take advantage of these other literatures do worry about data-mining and excess parameterizations.

Aren't Bayesian techniques the obvious solution to dealing with large numbers of observations in a rigorous way?

Tables should be self-contained. I find your tables very hard to understand; a key problem, given that they're the critical message of the story.

It's bad form to quote a 2002 paper when you mean a 2000 paper. More seriously, there's no indication that you corrected the errors posted on my website. I have informally discouraged people from using the data set you use ever since I posted the data set for my paper with Glick.

Why did you choose the 5 models that you chose, and why should one be interested in those models? I would have thought that you'd chosen one of these as a benchmark (perhaps the model from Anderson and Van Wincoop, or maybe Deardorff) and use some from the empirical literature as sensitivity checks. It's better to have a single benchmark that's defensible on some ground. I also note that you haven't use anything that I'd consider particularly compelling: for instance, my default model for my 2000 paper isn't in there. Also, what makes your first model "the most common extant gravity model?" It sure likes like mine without the currency union dummy, no? Where, for instance, is the standard gravity model with just the product of GDPs and distance?

Minor Points

I post my output on the web, so that you can see exactly how it's done. If you can't replicate my work, the onus is surely on you to resolve the issue or point out my mistake (or whatever).

Most work in this analysis does a lot of sensitivity analysis to show that results are sensible and insensitive. It seems unfair to ignore that fact altogether.

250 replications seems like a really small number to assess size and power. Why not let the computer run all night? Most of your work is linear.

When you add a variable to your estimation that doesn't exist in the data generation process, how can I interpret the coefficients in your table? Can I tell, e.g., what fraction of them are significant? If not, why not?