1 Introduction

The following programs were written for 64-bit SAS 8.2. I am not a particularly sophisticated programmer so the code is clunky. It works though.

The online appendix contains most of the data you need, and all of the code, to replicate results in the paper. Furthermore, there are many results that did not make it into the paper that are needed to check robustness, alternative specifications, as so forth.

Both the data and the SAS code are given in .txt format. Copy and paste this text into your SAS compiler to run the programs. The data must be uploaded into a library titled, “wiser.” Also you need to create another library called “workdata” where important files are stored. Be sure to consult the “State Export Data Read Me” to get the correct variable names for each data set as well as the correct dataset names.

All of the data sets have a ending column "null" consisting entirely of ones. This is to protect data when importing in SAS.

2 Data Construction

This program cleans and assembles all the data into forms usable by other programs. It creates datasets by year and also lumped together.

This may be a difficult program to use because it requires you enter all the data with the proper headings, and to create the libraries and file names I use.

I put my raw data, most of which may be downloaded from the online appendix, into a library titled “wiser.” The OM data is titled ‘om2001” or whichever year needed. This is the raw OM data obtained from WISER. The columns are: “state”, “naics,” “country” which is the US census country code, “sales” which is the (f.a.s.) value of OM exports in current year US dollars (ones). The next columns give the value of exports leaving the US by air and ship, and the weight (in tons)
of these exports by air and ship.

The raw AR-1 files are titled “arlist2001” or whatever year used. There is a raw AR-1 data set for each year. The details of these files are given in the AR-1 data read me.

The library “wiser” also contains concordance files that give state names, FIPS, postcode etc. The files have the name “states.” Finally, important constructed files are put into a library titled, “workdata.”

The OM data from 1991 is different since it is not from WISER. The program cleans it. The 1991 is the raw data without the 0 observations added.

The 1987 OM is reported state totals only.

3 Comparing OM to AR-1

This program makes many of the tables found in the text. It creates the percent and mean differences, limits of agreement, and the map. Some of the tasks it performs, such as correlations, did not make it into the text.


4 Regression

This program is the fixed effects regression for determining individual subsector and state quality. The program not only produces the values reported in the paper, but checks the assumptions of the regression model and considers some alternate specifications.

To use it, first construct a SAS dataset titled “aromstnc”. The program “construct_data” achieves this. The original paper was written when only the years 1997 and 2001 were available. Now the years 2002 and 2003 are also available. Data from 2003 is included, but it may be best to remove it to see how the paper was constructed.

The program first shows the pure data in a scatterplot. Then it takes the log transformation
of both the AR-1 and OM data and runs the following regression.

\[
\log OM_{s,i,t} = \beta_0 + \beta_1 \log AR_{s,i,t} + \varepsilon_{s,i,t}
\]  

(1)

where \(s\) is the state index and \(i\) is the subsector index, and \(t\) is either 1997 or 2001 or 2003. The time index is over these years only. NAICS is not compatible with SIC so years earlier than 1997 are not included.

Next the program looks to verify the assumptions of the log transformed linear regression model in (1). It produces graphs to indicate non-normality of errors, spot outliers and influential points, and heteroskedasticity. The log transformation reduces these problems anyway, but should be further verified. A Goldfeld-Quandt stat is calculated to test for heteroskedasticity. A white test is calculated also. Results indicate some influential points (those above the horizontal black line at \(4/N = 0.0022\) and non-normality of errors and some (but not a lot of) heteroskedasticity. Thus robust (White’s) standard errors will be used.

Next I test to see if the intercepts and coefficient on \(\log AR\) differ over time. The models estimated:

\[
\log OM_{s,i,t} = \beta_0 + \beta_1 \log AR_{s,i,t} + \varepsilon_{s,i,t}
\]  

(2)

\[
\log OM_{s,i,t} = \beta_0 + \beta_1 \log AR_{s,i,t} + \varepsilon_{s,i,t}
\]  

(3)

Table 1 has the results.

I reject that either the intercept or time coefficients are constant over time. This agrees with expectations: inland freight costs and other unobserved characteristics are likely to change over time, not to mention the observations are measures in current year dollars which change over time. Thus a separate regression will be run for each cross-section (1997 and 2001 and 2003) rather than pooled over time.

Now I test (2) against

\[
\log OM_{s,i,t} = \beta_0 + \beta_1 \log AR_{s,i,t} + \varepsilon_{s,i,t}
\]  

(4)
where \( x \) indicates either \( s \) or \( i \) depending on the fixed effects needed. The year-by-year pooled regression is rejected for NAICS for both years.

The fixed effects dummy variable model is

\[
\log OM_{s,i,t} = \beta_{0,s,i} + \beta_{1,t} \log AR_t + \varepsilon_{s,i,t}.
\] (5)

(5) is also rejected by an F-test, indicating the appropriate model is (4). However, because of the small number of observations for some states in a year (such as Alaska 2001 with only one observation), I use (5) so that there are enough observations to make meaningful claims.

Finally, I test for group fixed effects (5) in each year against pooled yearly regression (2). I test it against the pooled regression by comparing the unadjusted \( R^2 \) in an F-test. The \( R^2 \) is for an intercept model where NAICS 339 is dropped. Note in the main text a no-intercept model is used. Since the F-tests for both years is significant at the 5% level, the data is in favor of using NAICS fixed effects.

Though all of these tests should be repeated by state, I assume if anything the need for fixed effects is even greater for states and hence use (5) without explicitly repeating all of the tests above. I do perform one test for fixed effects over a regression pooled across states.

I use (5) despite (4) being the better model because of the small number of observations for small states such as Alaska. There are no issues with the non-normality of residuals if one uses (4). There is minor heteroskedasticity with (5) so White se are used.

The percent of between group variation ranges from 1/3 to 1/5.

I do A Hausman-style test to see if random effects would be more appropriate. All tests have significant F-values (NC97 10.40, St97 27.89, NC01 5.72, St01 13.61) indicating we reject the null hypothesis of random effects.
<table>
<thead>
<tr>
<th>Test</th>
<th>Label</th>
<th>$R^2$</th>
<th>RSS</th>
<th>Numerator DF</th>
<th>Denominator DF</th>
<th>F stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$</td>
<td>(1)</td>
<td>0.803</td>
<td>1651.61</td>
<td>2</td>
<td>1785</td>
<td>3.403*</td>
</tr>
<tr>
<td>$H_1$</td>
<td>(2)</td>
<td></td>
<td>1645.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0$</td>
<td>(3)</td>
<td>0.803</td>
<td>1649.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_1$</td>
<td>(4)</td>
<td></td>
<td>1645.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0$</td>
<td>(2)</td>
<td>0.790</td>
<td>1049.31</td>
<td>40</td>
<td>894</td>
<td>3.297*</td>
</tr>
<tr>
<td>$H_1$</td>
<td>(3)</td>
<td>NAICS</td>
<td>1997</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0$</td>
<td>(2)</td>
<td>0.816</td>
<td>596.04</td>
<td>40</td>
<td>774</td>
<td>4.423*</td>
</tr>
<tr>
<td>$H_1$</td>
<td>(4)</td>
<td>NAICS</td>
<td>2001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0$</td>
<td>(5)</td>
<td>0.807</td>
<td>967.04</td>
<td>20</td>
<td>943</td>
<td>2.800*</td>
</tr>
<tr>
<td>$H_1$</td>
<td>(4)</td>
<td>NAICS</td>
<td>1997</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0$</td>
<td>(2)</td>
<td>0.936</td>
<td>529.85</td>
<td>20</td>
<td>823</td>
<td>3.792*</td>
</tr>
<tr>
<td>$H_1$</td>
<td>(5)</td>
<td>NAICS</td>
<td>2001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0$</td>
<td>(2)</td>
<td>0.790</td>
<td>1049.31</td>
<td>20</td>
<td>943</td>
<td>4.023*</td>
</tr>
<tr>
<td>$H_1$</td>
<td>(5)</td>
<td>NAICS</td>
<td>1997</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0$</td>
<td>(2)</td>
<td>0.816</td>
<td>596.04</td>
<td>20</td>
<td>823</td>
<td>5.144*</td>
</tr>
<tr>
<td>$H_1$</td>
<td>(4)</td>
<td>NAICS</td>
<td>2001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0$</td>
<td>(2)</td>
<td>0.790</td>
<td>1049.31</td>
<td>49</td>
<td>914</td>
<td>6.222*</td>
</tr>
<tr>
<td>$H_1$</td>
<td>(5)</td>
<td>State</td>
<td>1997</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0$</td>
<td>(2)</td>
<td>0.816</td>
<td>596.04</td>
<td>49</td>
<td>794</td>
<td>6.224*</td>
</tr>
<tr>
<td>$H_1$</td>
<td>(5)</td>
<td>State</td>
<td>2001</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(*) indicates significance at the 5% level.
RSS is the sum of squared residuals.