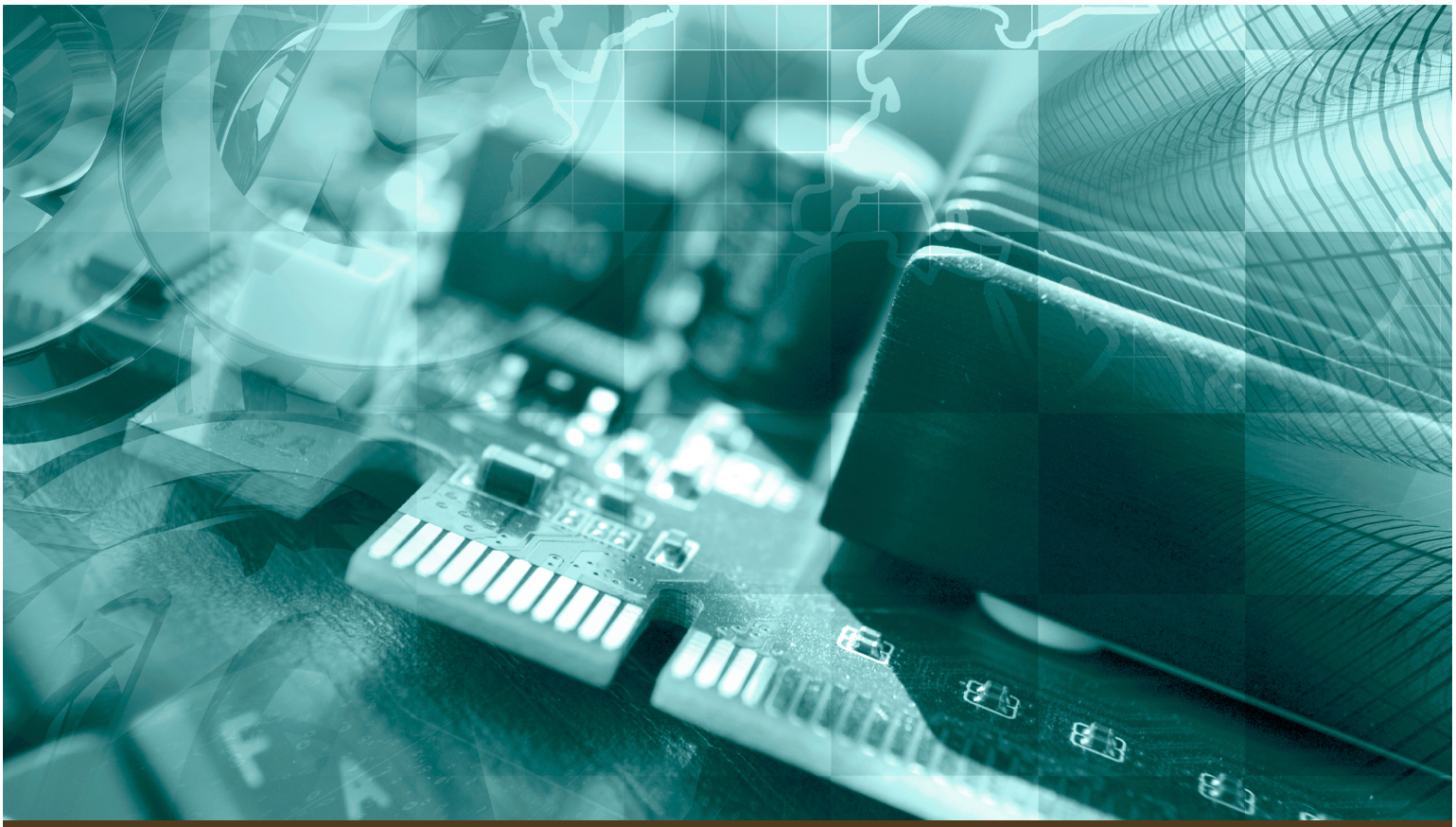
WASHINGTON STATE UNIVERSITY EXTENSION



A PRACTITIONER'S GUIDE TO TESTING REGIONAL INDUSTRIAL LOCALIZATION

By
Andrew J. Cassey, Associate Professor, School of Economic Sciences, Community and Economic Development Extension, WSU,
Ben O. Smith, Assistant Professor, College of Business Administration, University of Nebraska - Omaha

WSU PEER REVIEWED
TB05

A Practitioner's Guide to Testing Regional Industrial Localization

Abstract

The Ellison-Glaeser index is an unbiased measure of geographic industrial localization that improves upon simpler measures, such as the location quotient. We develop and describe software that allows for the Ellison-Glaeser index to be used in a statistical test to assess the chance that a particular industry is geographically localized. We give instructions on how to install the software, run the program, and interpret the results.

Introduction

Industrial localization is the geographic concentration of an industry's regional employment beyond that of the general economic activity or employment in the area. Thus an industry is localized in a country, state, or county if the employment in that industry is a large share of the overall employment. (The term "cluster" is often loosely used for "industrial localization" though these related concepts are, strictly speaking, distinct. Industrial clustering occurs when plants in the same industry are spatially correlated.) Industrial localization can occur because there are natural resources in the area as in the fishing, forestry, and mining industries. The gambling industry is localized in Las Vegas because of legislation preventing that kind of employment elsewhere. Alternatively, industrial localization may be the result of network effects, such as the computer industry in Silicon Valley or the automotive industry in Michigan in which there is pooling of skilled labor.

Identifying industrial localization is a common beginning step in regional and community development plans as analysts seek to build on the existing strengths of a community. (Regional development theorists emphasize the importance of economic base activities, which are the export industries of an area. For example, see Stimson et al. 2006.) Thus, given the area, identifying which industries are localized within it is important to regional development and planning. But doing so in a quantifiable way has not been possible, until now.

In this bulletin, we introduce a software tool that quantifies the degree an industry being studied is localized in a geographic space. Our software calculates the odds that the localization measured in the data could have been generated by chance alone, rather than from economic or natural advantages, such as those described above. Thus our software tool allows planners to identify which, if any, industries are localized and use that information when assessing regional development proposals.

A common measure of localization is the location quotient. A location quotient is the share of industrial employment to total employment in that region compared to the share of overall industrial employment to the overall total employment. When the location quotient is larger than one, that industry has a much larger share of employment in the area in question than that industry has in the overall economy. Thus a large location quotient suggests an industry concentration in the region in question. But as first noted by Ellison and Glaeser (1997), the location quotient is a poor measure of localization when the number of plants in the industry is small. They cite the U.S. vacuum cleaner industry as an example where it might look like localization because employment is in only four states. There are, however, only four plants in the vacuum cleaner manufacturing industry and so having one plant in each of four different states does not suggest localization of the type for one of those states to base an economic development plan around.

Because of examples such as the vacuum cleaner industry, Ellison and Glaeser developed a measure of industrial localization that corrects for the small numbers problem. As the geographic area being studied gets smaller, as from country to state to county to city, the small numbers problem becomes increasingly relevant, as it is likely the number of plants is also increasingly small.

Though more complicated and requiring more data than the location quotient, the Ellison-Glaeser (EG) index is a more accurate measure of industrial localization. The greater the value of the Ellison-Glaser statistic, the more the industry is localized in the geographic space. But though larger values indicate a stronger case for localization, there is not a simple interpretation of *how* large values of the index must be to truly indicate industrial localization. For example, the U.S. meat packing industry has an EG value of 0.042. Clearly that value is greater than zero, but it is not a large number. So is the meat packaging industry localized in the United States?

We developed software that improves the interpretation of the Ellison-Glaeser index by outputting the odds that the EG value from the data could have been generated by chance alone, rather than from economic or natural advantages. For example, after inputting the data, the program says there is a less than 5 percent chance that an industry with the characteristics of the U.S. meat packing industry would have an EG stat of 0.042 without economic forces for localization. This determination of whether something that looks like a localized industry is statistically localized because of natural or economic forces is important in regional planning because creating a development plan around an industry whose large employment share is likely due to randomness may differ from a plan around an industry whose large employment share is due to strong advantages of the area.

Our software is available for free download and can be personalized by a regional planner to test which industries, if any, in their area are statistically localized. The technical details and background theory of our software may be found in Cassey and Smith (2014). In addition to the Ellison and Gleaser (1997) statistic, our software has the option to output the test results for the related, but slightly different Maurel and Sédillot (1999) measure of industrial localization.

Data Requirements and an Example

Let's say regional planners have been asked to study the forestry industry in Whatever State, USA. Before using our program, they need to calculate the EG stat from the data. First they need to partition Whatever State into geographic subunits. One example would be counties. Researchers also need to assign how important each county is. The standard way to assign county weights is by the share of overall employment in those counties to the state employment. For example, let's say that there are three counties in Whatever State, with nonfarm employment of 12, 18, and 35. Then the weight for County One would be 12 / (12+18+35) = 0.1846. County Two would have weight 0.2769 and County Three would have weight 0.5385.

Next, the researcher would need to know the number of plants or establishments of the industry in the state. For this example, let's say there are seven plants in the forest industry in Whatever State, although there are 240 plants in the nation. Only the seven plants in the state are relevant. The researcher would also need to know the fraction of the industry employment in each of the three counties. Let's say that the industry employment in Whatever State is 14. The industry employment in each of the three counties is 8, 3, and 3. Thus in County One, of the 14 people employed, 8 of them are employed in the forestry industry. And thus County One has an industry employment share of 8 / 14 = 0.5714 whereas County Two and Country Three have industry employment shares of 0.21428 each.

Finally the researcher would need data on the plant Herfindahl index of the overall industry. The Herfindahl index is a measure of industry concentration used, among many applications, by the U.S. Department of Justice to determine the legality of corporate mergers. Secondary data on an industry's plant Herfindahl is usually available for free at the national level. The data may be publicly available for some industries at the state level. If the Herfindahl index is not available from a public agency, then the industry plant Herfindahl index needs to be calculated. To calculate a Herfindahl index, the researcher would need to know the employment at each plant or establishment in the industry as well as the total regional employment in that industry. The Herfindahl index is the sum of squared employment shares.

H = (plant 1 employment / industry employment)^2 + (plant 2 employment / industry employment)^2 + (plant 3 employment / industry employment)^2 + …..

In our example, there are 7 plants employing a total of 14 people. Let's say Plant One employs 6 people, Plant Two and Plant Three employ 2 people each, and Plants Four through Seven employ 1 person each. The sum of all the employment is 14. The Herfindahl is (6/14)^2+2*(2/14)^2+4*(1/14)^2 = 0.2449.

With those pieces of data in hand, the researcher can calculate the EG stat from their data and their specific application. The EG stat is:

$$EG = \frac{\sum_i (s_i - x_i)^2 - (1 - \sum_i x_i^2)H}{(1 - \sum_i x_i^2)(1 - H)}$$

where i is the index for subregions (in our example the 3 counties), x is the total employment share of the subregion, s is the industry employment share in the subregion, and H is the industry Herfindahl.

In our example,

$$\sum_i (s_i - x_i)^2 = (.5714 - .1846)^2 + (.2148 - .2769)^2 + (.2148 - .53846)^2 = 0.2586$$

and

$$\left(1 - \sum_i x_i^2\right) = 1 - (0.03408 + 0.07668 + 0.028994) = 0.5993.$$

Thus the example EG stat is (.2586–.5993*.2449)/(.5993*(1–.2449))=0.2472. This 0.247 value is the EG stat in our example data. Also, in our example, the location quotient in County One is 3.095, which is very large. Thus it *appears* as if the forestry industry is localized in Whatever State.

Our program simulates the regional economy many times to assess how likely it is that the EG = 0.247 we found in the data is from economic or natural forces rather than randomness.

How the Program Works

Our software calculates the odds that the EG value specified by the researcher given the data on plant employment in the area is generated from randomness and not from underlying economic or regulatory forces. The program asks the user to input some of the data that was collected above. It then uses that data to simulate 10,000 worlds. The program calculates an EG stat from each of those 10,000 worlds and compares the EG stat from the data to the distribution of 10,000 simulated EG stats.

If the EG stat from the data occurs outside the middle 95% of the simulated EG stats, then we know there is a less than 5% chance the EG stat from the data could have been generated by randomness. Thus that industrial localization is likely to have been generated by economic or natural advantages. A useful rule for interpretation is that there should be no more than a ten percent chance of the EG stat from the data being from randomness in order for the regional industry to be considered localized statistically.

In order to simulate an economy realistically, the user must input into the program the following data:

1. Geographic partition of the region with weights attached to the subregions
2. The number of plants in the industry in the overall region
3. The industry Herfindahl index for the region for interpretation of the output

Software Features and Customization

The program is customizable along the following features so that the user may get the most appropriate test for their region:

- The number of plants or establishments in the industry being tested, N.
- The number and importance of the geographic units partitioning the region, x. The default for the importance of the subregions should be the fraction of overall employment in those subregions. However there may be reasons to change this default weight. For example, the subregion weights may be modified to add importance for the location of natural resources (Ellison and Glaeser 1999).
- The rigor of the statistical test. The user can decide what statistical threshold is acceptable for declaring an industry to be localized. A standard cut-off is ten percent.
- The number of simulations (10,000 is the default).

Running, Understanding, and Personalizing the Application

Our software is different from the software used by most people in everyday life in that there is not a graphical interface. Rather our program is run by typing a line of code into a command line screen when prompted. The program then outputs files, which may be studied using spreadsheet software such as Excel. Instructions on downloading, installing, and opening the software are in the appendix.

Once the software has been downloaded and installed, use these instructions. From the included “command.txt” file, copy and paste the following line into the command prompt:

EGSimulation -t "data\tranche.txt" -c "data\criticalvalues.txt" –pvalues="data\pvalues.txt" -f "data\size.txt" -n "data\plants.txt" -s "data\sigma.txt" -i 10000 -d "data\out.csv"

This command executes the simulation using our predefined configuration files and saves the output to a CSV (named “out.csv”) in the data folder. Researchers will specify their own files (or replace our samples with text of their own) when using the program for testing the industrial clusters in their region. The number of simulations is set to 10,000 by default. But this can be changed by replacing the “10000” in the text line above. The more simulations, the more accurate our program will be and thus we do not recommend the number of simulations be less than 10,000. Note, however, that the program can take substantial time to run, especially if there are many plants and many subregions. Running the code that comes with the download takes a few hours on mid-range PCs. A full description of each customizable option is available from the command line by entering “EGSimulation –help”, but we will limit our discussion to the four necessary files: -t (“tranche.txt”), -c (“criticalvalues.txt”), -n (“plants.txt”) and -s (“sigma.txt”).

The lower left portion of Figure 1 shows the program ‘EGSimulation’ running on Windows using the sample configuration files. In the upper left region of Figure 1 we see the directory where the download resides. On the right we see the sample file “plants.txt”. The plants.txt file shown is a list with two items: 20 and 100. Running in this default mode will output a table with EG critical values for industries with 20 plants and industries with 100 plants from the same geographic weights. In our example of the forest industry in Whatever State, the user would delete the 20 and 100 from the default file and replace it with 7 since that is the number of plants in the forest industry in Whatever State.

The tranche text file defines the geographic weights to be used in the simulation. Each line indicates the size of the slice in percent terms and is cumulative from the first row so that the last row will always be one. Our sample is the United States partitioned into the 50 states. Each state is represented by the fraction of the nonfarm labor employed in that state. The order of the states represented by these employment shares does not matter. It also does not matter that the user be able to map each decimal representation to the state that decimal represents. All that matters is that every subregion is given *some* weight so that the sum of the weights is equal to one.

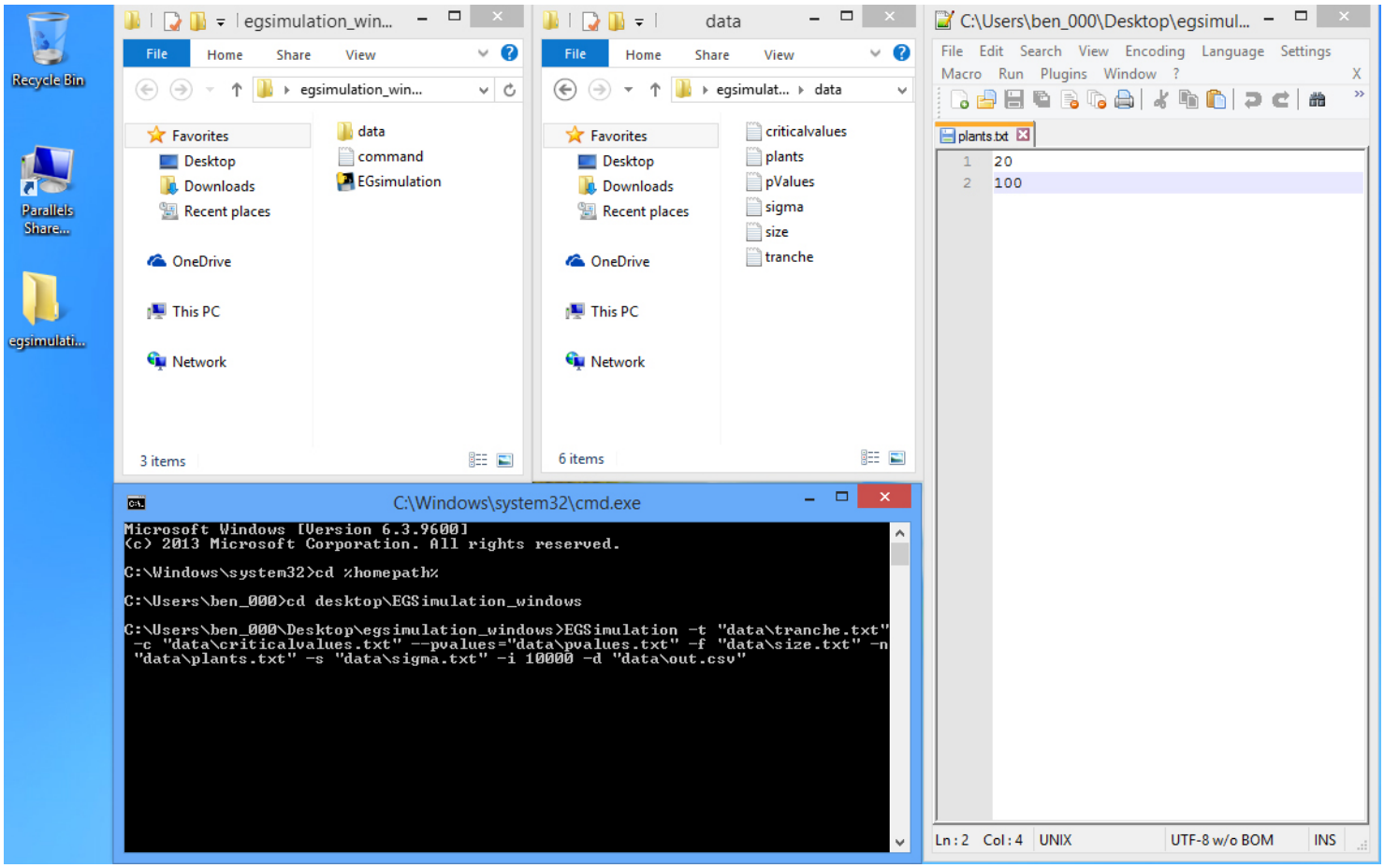


Figure 1: Entering program commands for 'EGSimulation' on the command line using the sample configuration files.

If users are interested in a particular state, they should adjust the tranche file to represent geographic subsections, like counties. For example, Washington has 39 counties and so there should be 39 lines in the tranche.txt file. King County is the largest county in Washington with about 1 million employed. There are about 3.2 million employed in Washington State. Thus the King County employment share is 0.3125. Spokane County's employment share is 214,700 / 3,200,000 = 0.0671. But the lines in the tranche.txt file will not be 0.3125 and 0.0671. Rather it will be 0.3125 then 0.3795 and so on, adding each country until 1 is reached.

The practitioner can do this by editing tranche.txt with a text editor. It is important to note that a text editor is an application like Notepad++ or Nano, and is capable of editing plain text files. It is not a word processor (like Microsoft Word). The user should always edit our configuration files with a text editor. The practitioner could also use weights other than the nonfarm employment weight in our sample. For example the weight used could treat each region equally regardless of size or could emphasize those regions with the most natural resources, depending on the application. In our example of Whatever State, there are three counties with overall employment shares of 0.1846, 0.2769, and 0.53846. These are *not* the values that should be entered into "tranche.txt". Rather replace the default file with 0.1846, 0.4615, and 1.

Criticalvalues.txt contains a list of critical values that the user wishes to include in the output. Critical values are numbers between 0 and 1 to indicate the level of statistical significance desired. The sample configuration file includes a fairly traditional list of critical values: 0.990, 0.975, 0.950 and 0.900. These correspond to 1%, 2.5%, 5%, and 10% levels of statistical significance. These critical values tell the program how wide to make the field goal posts to determine if the EG stat in the data is in fact generated by randomness. The greater the critical value, the more conservative the researcher wants to be to assess if the industrial localization being tested is caused by natural or economic causes instead of randomness. If the user wants to include or eliminate a critical value, they would simply add a new line or delete one from this sample.

The value in the "size.txt" file is a technical parameter that has absolutely no impact on the output files. But it needs to have *some* value. Thus keep the default value of 10.

Finally, our application works by simulating the world based on a specific distribution of employment to each of the plants in the industry. Studies indicate that often employment is log-normally distributed among plants, meaning there are a few very large plants and many plants with few employees. But that lognormal distribution is not a single distribution but rather a class of distributions.

A particular version of the lognormal distribution is specified with the sigma parameter. Sigma.txt is a list of standard deviations of the underlying normal distribution to simulate. There are two possible reasons to modify this file:

1. The user has fit their dataset to a lognormal distribution and knows the exact standard deviation of the lognormal distribution in the data.
2. The user has a belief about the possible range of standard deviations for a given industry and wants to limit the simulations to a subset of the values included in sigma.txt file.

In general, without information to the contrary, it is best to leave this to the default list.

Figure 2 shows the program running on the default settings on a Windows machine. The program will let the user know each time one standard deviation valued run is completed. As the default range for the standard deviation is 0.25 to 2, one can track the progress of the program. The program is computationally intensive and can take hours for a problem with many plants and many subunits on a midrange machine.

Interpreting the Software Output

When you run our software and after it finishes its simulation, it will generate a CSV file readable by any spreadsheet application such as Microsoft Excel. The output has many columns with lots of numbers, but most columns are internal checks for validity. (For example, the application reports the mean gamma value, which should always be very close to zero if everything is OK.). There are two sets of columns of particular interest: ones that start with “C” and ones that start with “H”.

Critical values are reported using a “C” followed by the corresponding confidence level. As an example, if you were performing a one-tailed test with five percent type I error, you would look at the column labeled “C095”. Let us return to our example of the forest industry from Whatever State. Using the data in that example, we calculated EG = 0.2472.

That value is greater than zero so there is evidence that forestry could be localized within Whatever State given the large share of forestry employment to overall employment in that county. We would like to statistically test if the chance of the data having an EG value of 0.2472 from randomness is less than 5% above and below zero. Thus we look to the “C0025” column and the “C0975” column because that puts 2.5% on the chance of the value being less than zero and 2.5% chance of the value being greater than zero for a total of 5%.

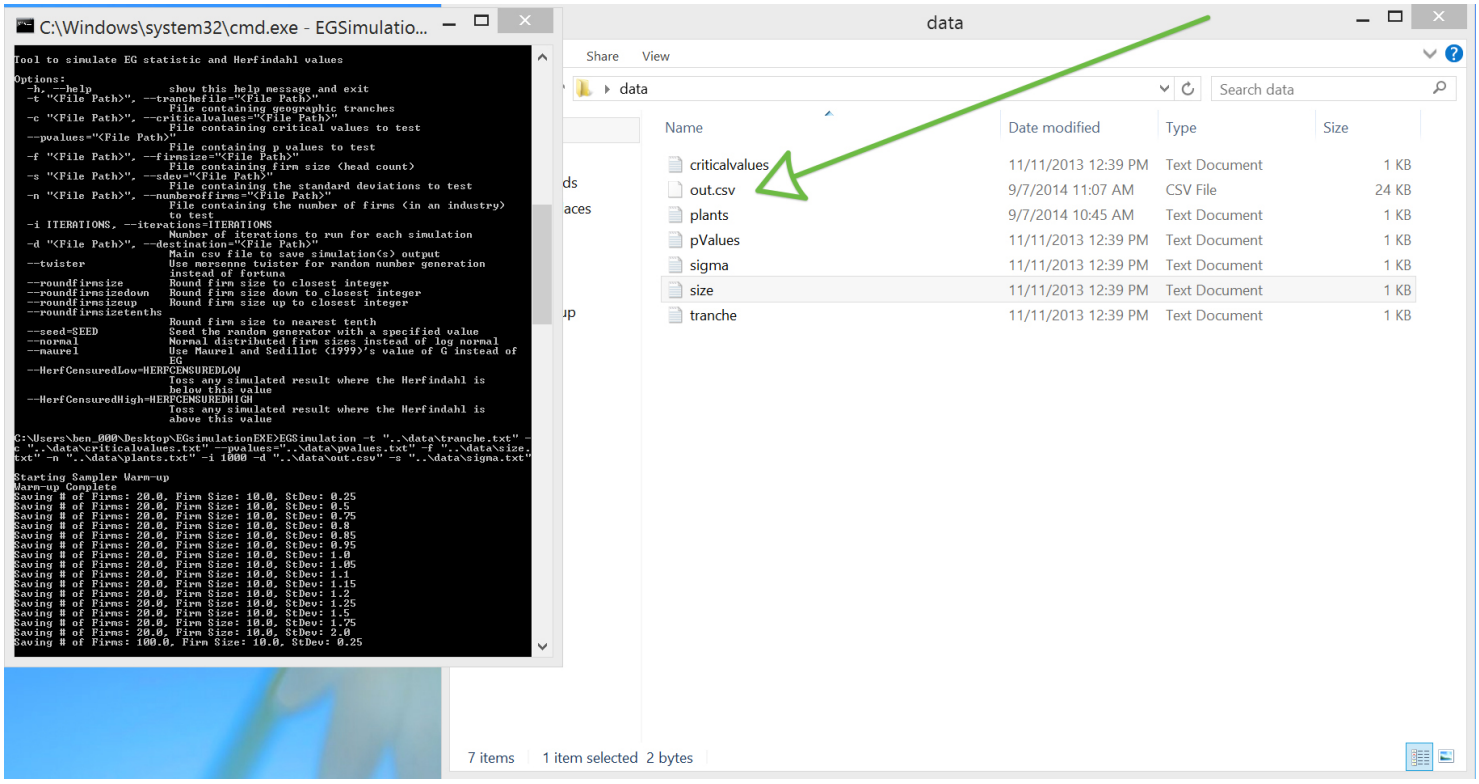


Figure 2: 'EGSimulation' running on Windows 8.1 using the sample configuration files.

Running our example with 7 plants in 3 countries on the code produces the following table, where we have eliminated many technical columns.

Plants	Stddev	Gamma Mean	p-values	H0025	H0975	C0025	C0975
7	0.25	0.0017	.6343	.144	.162	-0.1674	0.4600
7	0.5	0.0020	.6251	.148	.229	-0.1964	0.5210
7	0.75	-0.0022	.6203	.155	.349	-0.2515	0.6332
7	0.8	-0.0002	.6172	.157	.373	-0.2673	0.6605
7	0.85	-0.0017	.6185	.159	.400	-0.2853	0.6843
7	0.95	-0.0039	.6218	.162	.459	-0.3333	0.7839
7	1.00	0.0004	.6163	.165	.502	-0.3434	0.8200
7	1.05	-0.0021	.6190	.167	.528	-0.3765	0.8675
7	1.10	-0.0050	.6054	.169	.563	-0.3984	0.9007
7	1.15	-0.0007	.6063	.170	.583	-0.4396	0.9685
7	1.20	-0.0062	.6174	.172	.638	-0.4952	1.0712
7	1.25	0.0034	.6048	.173	.642	-0.5027	1.1244
7	1.50	0.0026	.6020	.184	.767	-0.7400	1.4706
7	1.75	0.0001	.5899	.200	.848	-1.0861	1.9678
7	2.00	0.0327	.5832	.207	.907	-1.5402	3.0951

The first column is the number of plants in the industry the user specified, 7. The stdev column is a list of the standard deviation values of the lognormal employment distribution from the "sigma.txt" input. Together these columns indicate what the odds of localization would be in a world with 7 plants and an employment distribution that is unknown but increasing incrementally in our output. The Gamma Mean column should have values near zero if the program is working correctly. The p-values column indicates the probability of obtaining a test statistic at least as large as in the data. Small p-values of .01 or .05 are needed to find evidence of industrial clustering statistically.

The next columns to consult are the "H" columns. Only keep those rows in which the Herfindahl data falls between the values in the columns for further consideration. In our example, the Herfindahl is 0.244. Thus we can eliminate the rows in which stdev is 0.25 or 0.5 because 0.244 falls outside of the "H0025" and "H0975" for those rows.

The final columns are the "C" columns, which give the critical values. All of the values in the "C0025" column are less than zero (as they will be in every case), and thus there is no evidence that the EG stat value of 0.247 from the example is more diffuse than randomness. But notice in the "C0975" column that none of the values are less than 0.247. That means that even though it *seems* like forestry is localized, there is a very large chance that it is due to randomness and not the economic structure of the industry in the region. If *all* of the columns had been less than 0.247 then there would have been strong statistical evidence that the forestry employment observed in county 1 would have been due to nonrandom economic structure.

If there were some rows that had a low p-value or the "C0975" column value had been less than 0.247, then the industry could be considered localized statistically for *some* structures of the industry but not others. But it is very hard, if not impossible, to know what the true value of "sigma" is. Thus the user has to decide how conservative or liberal they wish to be with the results. The most conservative researcher would say that the EG stat in the data must be greater than the "C0975" value for *all* rows in which the Herfindahl in the data falls within the "H0025" and "H0975" rows. A liberal researcher would say the industry is localized if there is *at least one* row for which the EG stat from the data exceeds the critical value.

Summary

Identifying regional industrial localizations is an important stage in developing an economic development plan. As the geographic space under study becomes smaller in size, going from the level of a country to a state, county, or city, the problem of small numbers of plants in the study area increases. The small numbers problem can cause the location quotient measure of industrial localization to identify an industry as localized due to randomness rather than natural or economic advantages in the region. To avoid basing an economic development plan of statistically insignificant industrial localizations, a planner should instead identify an industrial localization with the Ellison-Glaeser index. Though more complicated to calculate, the EG stat removes the problem of small numbers from the analysis and thus is more accurate.

Our program simulates the regional economy to statistically test if the EG stat calculated by the researcher from their data indicates industrial localization. The program is free and downloadable as a complete package for all operating systems at http://goo.gl/n1N06.

References

Cassey, A.J. and B.O. Smith 2014. "Simulating Confidence for the Ellison-Glaeser Index" *Journal of Urban Economics* 81(1): 85–103.

Ellison, G. and E.L. Glaeser 1997. "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach" *Journal of Political Economy* 105(5): 889–927.

Ellison, G. and E.L. Glaeser 1999. "Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration?" *American Economic Review* 89(2): 311–316.

Maurel, F. and B. Sédillot 1999. “A Measure of the Geographic Concentration in French Manufacturing Industries” *Regional Science and Urban Economics* 29(5): 575–604.

Stimson, R.J., R.R. Stough, and B.H. Roberts 2006. Regional Economic Development: Analysis and Planning Strategy, 2ed. Springer: Berlin.

Acknowledgements

Cassey acknowledges the financial support of Washington State University Extension and Agricultural Research Center project #0540 at Washington State University. We thank the comments of the editorial staff and two anonymous referees.

Appendix: Software Installation Instructions and Opening the Application

Source Code Installation: For Linux/Unix and Mac OS X Users with Python Installed

Using the source code directly is easy on Linux and Unix, including Mac OS X. Using our source code requires Python 2.7 (https://www.python.org/), SciPy (http://www.scipy.org) and PyCrypto (https://www.dlitz.net/software/pycrypto/). Because Python 2.7 is pre-installed on most Unix/Linux systems, using the source code directly is usually as simple as downloading it. The source code is available at http://goo.gl/n1N06 under the label "EGSimulation."

For Mac OS X users, our software is also included in the MacPorts repository, http://www.macports.org/. With MacPorts installed, a Mac OS X user should open the Terminal (found under Applications/Utilities) and type at the prompt:

```
sudo port -v selfupdate
```

(there will be a prompt for a password)

```
sudo port install EGSimulation
```

This will automatically download the latest version as well as all dependencies and automatic updating, which may take some time. If you do not have Macports installed already, you should skip ahead to the next subsection "Binary Code Installation."

You can install Macports by first going to the Macintosh app store and installing Xcode 4 or greater (as of this writing Xcode 6.1 is most recent). Next, open the Terminal and type at the prompt:

```
xcodebuild –license
```

Keep spacing down and type **agree** at the prompt to agree to the license. Next, go to http://www.macports.org/ and click on the "pkg" link for your version of Mac OS X. Once downloaded, open the .pkg file and run the installer. Then open the Terminal and run the commands above.

Unfortunately, installing on a machine running a Windows operating system is not as simple. However, we have greatly eased the installation process by providing pre-compiled binaries available for Windows and Macintosh operating systems.

Binary Code Installation: For Windows and Mac OS X Users

To make the process of using our application more seamless, we have created pre-compiled binaries for Windows and Mac OS X, meaning everything you need comes packaged together. Both versions may be obtained at http://goo.gl/n1N06 under the subheading "Binary Versions Available." These programs have no dependencies and thus can run on a system without having Python pre-installed. Click on the link for your operating system to begin the download of the binary package.

After you download the binary package, unzip the file to a folder or directory of your choice. The zip file will contain three items:

1. A binary file named "EGSimulation"
2. A directory/folder named "Data"
3. A text file name "command.txt"

"EGSimulation" is the program itself. You will execute it from the command line using configuration files, which are plain text, (.txt) files. We provide sample configuration files in the folder "Data."

Opening the Application

For Windows users, after you download and extract the Windows package on your desktop (in this example the folder is titled "EGSimulation_windows," but you can choose any name), find and open the software:

1. Click on the start menu and type in the search box "cmd"
2. Type "cd %homepath%"
3. Type "cd desktop\EGSimulation_Windows"

Step one opens the command prompt while steps two and three navigate to the location of the application. In this example, we told the command line interface to change the directory (cd) to the desktop folder named EGSimulation_Windows. You should be prompted.

For Mac users, open the Terminal by locating it in the Utilities subfolder of the Applications folder or by typing:

```
command-space
```

and searching for Terminal using Spotlight. Open the Terminal. You should be prompted. Type

```
cd path\egsimulation_mac:
```

where "path" is the folder and subfolders used to store the download. You should be prompted.

Issued by Washington State University Extension and the U.S. Department of Agriculture in furtherance of the Acts of May 8 and June 30, 1914. Extension programs and policies are consistent with federal and state laws and regulations on nondiscrimination regarding race, sex, religion, age, color, creed, and national or ethnic origin; physical, mental, or sensory disability; marital status or sexual orientation; and status as a Vietnam-era or disabled veteran. Evidence of noncompliance may be reported through your local WSU Extension office. Trade names have been used to simplify information; no endorsement is intended. Published November 2015.