The Economic Impact of Friends: Shift-Share Analysis With Regression

Abstract

This case study examines the research methods used to investigate the effects of social capital on economic development. It outlines the choices made to investigate the research question using a novel approach that combined shift-share analysis with traditional regression techniques. Using archival data to answer this question raises a number of issues that created both advantages and disadvantages for the research. The case concludes with three lessons that address (1) the use of shift-share analysis, (2) an evaluation of the strengths and weaknesses of archival research, and (3) a discussion of how differing units of measurement are addressed through aggregation at larger geographic levels.

Learning Outcomes

By the end of this case, students should be able to

- Explain how shift-share analysis addresses challenges of causation in economic research
- Describe the advantages and disadvantages of using archival data in conducting research
- Identify the challenges introduced by using multiple datasets within a single statistical model

Project Overview and Context

It has long been held that there is a connection between relationships and economic prosperity. Think about how many times you have heard someone say, "it's who you know that matters." This is question about what sociologist would refer to as social capital. Social capital is "the information, trust, and norms of reciprocity inherent in one's social networks" (Woolcock, 1998, p. 153). Social capital, like other forms of capital such as the physical capital used in production or the economic capital used to buy supplies, has an economic value for those who possess it.

One might ask a similar question as to whether communities where people are connected to each other in multiple ways are also able to take advantage of these relationships to help their communities grow economically. The question of "do communities with social capital produce greater economic outcomes" also speaks to a contemporary debate within regional economics. Some scholars have advocated that the key to economic growth lies in the idea of the creative class as advocated by Richard Florida (2005). This argument emphasizes increasing economic growth through the in-migration of new, talented, and creative individuals. This idea stands in contrast to the social capital perspective advanced by Putnam (1995) which argues that relationships are critical to thriving communities.

Given (1) the economic slump at the time of our research and (2) the billions of dollars that municipalities spend each year in direct subsidies or tax abatements, determining if relationships could be used as a means

of fostering economic development in the form of job creation and income growth was an important topic. However, this type of research is confounded by three factors. First, using statistics to demonstrate causation is an inherently difficult endeavor as any economic outcome is almost always the result of the confluence of many unmeasured social forces. Second, this research question is one of national scope and thus would generally require a lot of support to garner a nationally representative sample. Third, none of the faculty involved in this initiative had funding to support the research project and thus the research design needed to be one that required minimal economic expense and little infrastructure to administer. Because of these factors, we adopted a research design that balanced practicality and useful design.

Research Design

Archival Data in Research

Considering these constraints, we adopted a research design that would reply upon archival data. Each year, countless private and public organizations collect data for their own uses and then make these data available for others to use through data repositories such as the Inter-university Consortium for Political and Social Research (https://www.icpsr.umich.edu/icpsrweb/) or the American Census Bureau (https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml). The choice to use archival data introduces its own set of challenges, but dramatically lowers the cost of doing research as it limits the expenses required to collect, enter, and sometimes clean data. In the end, we would use nine different data sets to compile the variables needed to conduct a rigorous regression analysis.

In addition to cost-saving benefits associated with using archival data, using a wide range of data sources met another need of this study. The concept of social capital is one that has become theoretically and empirically diverse with countless approaches to measuring social capital (Engbers, Thompson, & Slaper, 2017). Given the lack of agreement on what social capital is and how it is measured, the study sought to include a wide range of constructs and measures. Using archival data allowed greater breadth of control variables as the study was not limited by the length of any one survey or the cost of administering a wider range of measures. Consequently, seven measures of social capital were ultimately included in the model in addition to 11economic control variables that were drawn from the economic development literature.

Selecting a Sample

However, because the study relied on data collected by others, this introduced a challenge associated with issues of aggregation. Each research organization collected the data with their own methods and own units of analysis. This meant that we would have to find a way to compare research collected on individuals with data reported at the county or Metropolitan Statistical Area (MSA) level. The decision was made that the study would focus on the 50 largest MSAs.

Focusing on MSAs makes sense for several reasons. First, economic outcomes in one local government jurisdiction often affect other nearby jurisdictions as job losses and gains spillover to nearby areas. Likewise, social capital relationships are rarely constrained to a political jurisdiction but rather relationships form across political boundaries. Second, many MSAs have regional economic development authorities who work to coordinate economic development within a region. Third, metropolitan areas are a unit of analysis at which policy makers may exert some level of control. As one moves toward higher levels of political authority such as states and national economies, the shaping of economic outcomes becomes less precise.

The choice to focus on the largest 50 metropolitan areas was a choice made largely out of necessity. It would have been valuable to include a larger number of metropolitan areas to foster greater external validity, but this was not possible using archival data. Because much of the data used in the study were based on individuals, the study needed to ensure that claims about metropolitan areas were accurate. This requires having a sufficient number of individuals in each jurisdiction to make community-level claims. In the smaller metropolitan areas, there were not enough survey responses to be able to make meaningful claims about the MSA.

However, the choice to focus on the largest 50 metropolitan areas was defensible. In 2000, the year used as the base year in the analysis, the top 50 MSAs represented 59% of the total U.S. population thus suggesting that the findings would be relevant in the communities in which most Americans live. Moreover, using the largest 50 metropolitan areas would ensure some geographic representation as these MSAs are found spread geographically throughout the United States. Finally, there is a statistical justification for using MSAs as a unit of analysis as they reduce statistical concerns related to spatial dependence in the error term which has been observed in county-level studies (Li, Cheng, & Haynes, 2011).

Inferring Causality

One of the biggest challenges in economic and social research is the difficulty with inferring causality. Social, economic, and political phenomena are rarely caused by one thing and reaching definitive conclusions from statistics is all but impossible. However, shift-share analysis presents one mathematical method to better infer the effects of local conditions on economic outcomes. Shift-share analysis has been used since the 1960s (Dunn, 1960) to separate job growth effects into the relative contribution from a national growth effect, an industrial growth effect, and a regional or competitive growth effect.[1] For example, a growth in jobs in Houston, Texas, in early 2008 can be explained by a number of factors. First, during that time, the cost of oil was at an all time high leading to a rapid inflow of jobs in communities with natural resource–based economies. Second, the U.S. economy was facing a period of expansion and almost all communities were experiencing job growth. Third, local policy makers may have been adopting policies that lead to economic expansion. Because of these simultaneous factors, it is difficult to ascribe economic growth to any local phenomena as it is confounded by industrial or national effects. Shift-share analysis solves this problem by taking the local job growth (loss) rate and subtracting from it the national trend and any effect anticipated from the local industrial mix. This leaves a measure of job growth that can more effectively be ascribed to local

conditions and thus better situated to infer causality.

What makes this study somewhat unique is that shift-share analysis is traditionally a descriptive method. A typical critique of shift-share analysis is that it only offers a descriptive accounting of job changes without providing inferential evidence for that change (Gabe, 2006). Our study extends shift-share analysis by paring the shift-share job growth rates for each MSA with inferential statistics. In this way, the locally attributed job growth rate becomes the dependant variable in a regression model with a full range of social and economic indicators to infer more clearly the effect of local factors such as social capital in explaining job growth.

Method Details

This study looked at job growth between 2000 and 2009 and because employment growth rates differ by industry (Gabe, 2006), this study included 19 industry-level dummy variables. Thus, our model most closely approximates a one-way fixed effects panel that varies by industry.

The regression model was estimated with the independent variables in per capita form to control for the effects of metropolitan size. The independent variables were represented as either a base-level measure for the year 2000 or a change measure over the study period, depending on the potential and actual impact of the specific variables. When change forms were used, the change was based on either the percentage change between 2000 and 2009 (for variables that represent a nominal number such as the per capita expenditure on education or the per capita number of nonprofits) or as the difference in the 2000 and 2009 data (for variables that represent a percent, like the percent of earnings or violent and property crime rates).

Because the study used data from both 2000 and 2009 and the largest 50 metropolitan areas changed during that time, the study ended up only focusing on 47 metropolitan areas. Consequently, the number of observations in our analysis was 893 (19 industries times 47 MSAs).

Practical Lessons

The results of the study demonstrated that social capital does have a positive effect on economic outcomes. At the metropolitan level, social capital as measured by the number of nonprofits and the percent change in bridging social capital (a type of social capital that indicates the number of connections among diverse people) demonstrated a substantive and statistically significant effect on job creation. Moreover, in addition to the empirical findings, this study offers three practical lessons for research: (1) the advantages of shift-share analysis, (2) the advantages and disadvantages of using archival data, and (3) insight into how aggregation helps overcome disparate data sources.

Advantages of Shift-Share

First, as suggested above, shift-share analysis offers a unique tool to determine the localized impact on job

creation. Because shift-share analysis decomposes job growth into three parts—a national growth effect, an industry growth or industry mix effect, and a regional or competitive advantage growth effect—researchers are able to judge the relative impact of local and national conditions. Without using shift-share analysis, any growth in the local economy might be attributed to either national trends or the industrial mix of the community.

Thus, the shift-share analysis accounts for the change in jobs that an industry's location in a given MSA has over the typical industry in other regions. This allows us to address the "great divide" between economic development theory and economic development practice by more appropriately considering the change in the number of jobs that regional planners can control through development measures (Currid-Halkett & Stolarick, 2011).

While shift-share analysis has traditionally been used to examine job creation, a similar technique could be used for measuring other social or economic conditions. All that is required is to have measures of the phenomena of interest at the national and local/regional level along with some indicator of composition of the community. For example, market researchers may wish to understand how thought leaders affect the views of a project in a particular jurisdiction; they could take local sales and subtract out national sales. This could then be adjusted by the demographic mix of the community to provide a true measure of the local effect of the thought leaders' preferences.

Evaluating Archival Data

Second, this study highlights the advantages and disadvantages of using archival data. For the purposes of this study, the advantages of archival data were significant. Given the constraints of time and money that the researchers faced, the study would not have been completed without the availability of archival data. National surveys routinely cost hundreds of thousands of dollars and in the case of longitudinal studies such as the General Social Survey and Panel Study of Income Dynamics, millions of dollars are spent. Added to this, the time and manpower required to develop research questions, identify a sample, collect data, and clean and analyze the data, such an undertaking would have been nearly impossible. That said, the cost and time savings are not the only advantages. For this study, the use of archival research presented two other advantages. For one, the desire to use a wide range of social capital measures and constructs was more easily facilitated by using existing data. The study was not constrained by many of the factors that researchers face when they collect data such as the acceptable length of the survey or the need to norm new measures. Rather, collecting data from a variety of sources limited survey fatigue and offered measures that had been subject to significant testing without the demands of doing the measurement validation ourselves. Similarly, the data used in this study were collected by institutions that are widely perceived as fostering high-quality research (e.g., The Bureau of Labor Statistics, The Census Bureau, the University of Michigan). This lent credibility to our data and the conclusions that were drawn.

Yet, despite all these advantages, there were significant disadvantages to doing archival research. Although all the data used in this study were free, they were not costless. The time required to combine the data sets

together and to clean the data was significant and required well over 100 hours of time to get the data to the point where it could be analyzed. This was particularly problematic because between 2000 and 2009, some of the jurisdictions of the metropolitan areas changed which required us to use county-level data and "rebuild" the MSA measures based on a consistent geography. Were we using our own data, we could have controlled the way in which the data were collected.

The use of archival data is also inherently problematic for a number of reasons (Elder, Pavalko, & Clipp, 1992). First and fundamentally, the data that one inherits are rarely the data that one wants. Archival data are collected by a researcher at a particular point in time to meet their needs. It reflects their original questions, data demands, preferences for analytic techniques, and personal biases. More often than not, this is not a perfect match for the preferences of later researchers. Second, given that archival data are rarely a perfect match for current research, this leaves the researcher with a need to justify its appropriateness. Although we justified the choices above based on practical considerations such as time and cost, it is critical for the legitimacy of the research that the archival measures can be justified on their inherent worth. In other words, researchers need to be able to defend the use of these data in terms of its theoretical relevance and empirical validity. Consider, for example, our use of crime data as a proxy for community trust and stability. As researchers, we struggled with whether this was an appropriate measure for this study. However, we were able to justify it to ourselves and to our reviewers based on (1) its theoretical consistency, (2) past empirical research demonstrating a relationship between crime rates and trust, and (3) a prerequisite of other scholars using the measure in the same way. Third, research results are the product of the time in which they were collected. Responses to survey results are biased by the social, economic, and political conditions of the time. As a researcher collecting data, one is aware of what confounding factors may be influencing the data collected. When one reaches into the past to get data, researchers are often less attentive to the social and political dynamics that may have influenced that data. Fourth, archival survey data of the kind used in this study are difficult to reinterpret. Unlike qualitative data that can be recoded based on the needs and preferences of the researcher, quantitative data are less subject to reanalysis. Rather, quantitative data are taken as having rather fixed meaning, a meaning that is often ambiguous without the presence of the research subjects to clarify (Elder et al., 1992). In these ways, archival data limit the ability to draw precise conclusions for contemporary questions.

Accounting for Time and Place

Third, this study highlights a method to use when data are collected over time from different respondents in different geographies at different units of analysis. One of the gold standards of statistical design is the longitudinal panel. The longitudinal panel design measures individual respondents over time allowing researchers to compare groups not only with each other but also with themselves in a past time. Consequently, it allows researchers to control for time-invariant characteristics such as sex or race (at the individual level) or typography or climate (at the municipal level). It is also a stronger design when there may be confusion about the temporal order of events. For example, in our research looking at relationships and economic development, we argue that tight knit connections lead to greater job creation. However, a

reasonable scholar might suggest that growing economics bring people together. In this sense, the time order effect is unclear. Consequently, a panel design would alleviate these concerns.

Unfortunately, to control costs and because we relied on archival data, we were unable to use a panel to address these issues. Rather, we constructed a pseudo-panel that had a similar effect without the requirement to track individuals over time or to conduct longitudinal data collection. We did this through aggregation. Much of the data used in this study were longitudinal. Rather than track individuals over time, these surveys would include an original set of survey respondents each year. Consequently, individuals could not be compared with their previous selves. As we were not concerned with individual results, but rather community-level outcomes, we aggregated individuals to their metropolitan level. For example, if 120 individuals answered the question about group membership in St. Louis in 2000 and 147 individuals answered the same question in St. Louis in 2009, we averaged those responses to determine a St. Louis metropolitan measure of group membership for both 2000 and 2009. Thus, the time-invariant aspects of the communities become controllable and we are able to determine time order effects with the pre- and post-MSA levels.

This solution is not without its limitations. First, any data aggregation must occur to the highest level of shared geography. When we had individual, county, and MSA indicators, we could not conduct analysis at the MSA level because although individuals could be aggregated up to the county level, there is no way to partition the survey responses for the MSA level. Second, this analysis requires geographic variables for all the individual data. For many publicly available data sets, geographic indicators have been stripped to ensure anonymity. We had to limit our use of archival data to surveys that included a county level or MSA geographic variable. Thus, in our research, we encountered the problem that the geographic indicators for MSA changed over time. In other words, the geographic definition of the MSA was not the same in 2000 and 2009. Fortunately for our study, we also had county-level indicators that allowed us to reconstruct the 2000 borders, but this added a level of work and complexity that we would not have faced with stable metropolitan boundaries.

Conclusion

Our research into the effects of social capital on job creation represents an interesting approach to studying economic development. The use of archival data allowed us to complete the research within a shorter time frame and with fewer economic resources. However, it also introduced a number of challenges. These included reliance on proxy measures and the need to develop a research strategy to deal with changes that occurred over time and in response to changing geographies. One of the strategies that we found particularly valuable was to merge multiple data sets together by aggregating to a higher level of geography. Once this occurred, we had a wide range of social and economic indicators that could be used to examine the question. We then paired traditional regression techniques with shift-share analysis to determine the local effect of social capital on economic development.

Exercises and Discussion Questions

1. What are the strengths and weaknesses of using archival data for quantitative research?
2. How do shift-share methods improve upon traditional estimates of economic and social phenomena?
3. What problems are introduced by using measures from different units of aggregation (e.g., individuals, counties, states)?
4. In what ways does combining regression with shift-share analysis improve descriptive shift-share techniques?
5. Were you to attempt to answer a similar research question as discussed in this study, what would you have done differently?

Further Reading

Elder, G. H., Jr., **Pavalko, E. K.**, & **Clipp, E. C.** (1992). *Working with archival data: Studying lives* (SAGE University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07088). Newbury Park, CA: SAGE.

Engbers, T., **Rubin, B. M.**, & **Aubuchon, C.** (2017). The currency of connections: An analysis of the urban economic impact of social capital. *Economic Development Quarterly*, 31(1), 37–49.

Jackson, R. W., & **Haynes, K. E.** (2009). Shift-share analysis. *International Encyclopedia of Human Geography*, 125–131.

Stimson, R. J., **Stough, R. R.**, & **Roberts, B. H.** (2006). *Regional economic development: Analysis and planning strategy*. Berlin, Germany: Springer Science & Business Media.

Wang, X., & **Hofe, R.** (2008). *Research methods in urban and regional planning*. Berlin, Germany: Springer Science & Business Media.

Web Resources

An explanation of shift-share techniques: https://kb.economicmodeling.com/all-about-shift-share/

An overview on archival data: http://ctb.ku.edu/en/table-of-contents/evaluate/evaluate-community-interventions/archival-data/main

A resource on data archives: https://guides.nyu.edu/c.php?g=276966&p=1848686

References

Currid-Halkett, E., & **Stolarick, K.** (2011). The great divide: Economic development theory versus

practice—A survey of the current landscape. *Economic Development Quarterly*, 25, 143–157.

Dunn, E. S. (1960). A statistical analytical technique for regional analysis. *Papers in Regional Science*, 6, 97–112.

Elder, G. H., Jr., **Pavalko, E. K.**, & **Clipp, E. C.** (1992). *Working with archival data: Studying lives* (SAGE University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07088). Newbury Park, CA: SAGE.

Engbers, T. A., **Thompson, M. F.**, & **Slaper, T. F.** (2017). Theory and measurement in social capital research. *Social Indicators Research*, 132, 537–558.

Florida, R. (2005). *The flight of the creative class: The new global competition for talent*. New York, NY: Harper Collins.

Gabe, T. M. (2006). Growth of creative occupations in U.S. metropolitan areas: A shift-share analysis. *Growth and Change*, 37, 396–415.

Li, H., **Cheng, S.**, & **Haynes, K. E.** (2011). The employment effects of new business formation: A regional perspective. *Economic Development Quarterly*, 25, 282–292.

Putnam, R. D. (1995). Bowling alone: America's declining social capital. *Journal of Democracy*, 6, 65–78.

Woolcock, M. (1998). Social capital and economic development: Toward a theoretical synthesis and policy framework. *Theory and Society*, 27, 151–208.