Chapter 2

Methodology

The research that contributes to this document involves mixed methods. Archival research is employed to construct the historical narratives in the text, particularly Chapters 3 and 8. This work relies primarily on contemporary sources and secondary reports of historical developments in the labor and technology of transportation. Quantitative analysis is incorporated to facilitate an understanding of the spatial implications of sociotechnical change. Data sets for these analyses includes existing government-produced data and data compiled from industry publications by the author.

2.1 Quantitative Analyses

The follow sections present the basic set up for the regression analyses and mapping that underlie and inform several of the following chapters, particularly Chapters 4, 6, and 7.

2.1.1 Computing and software

The data were processed through PostgreSQL 8.3 and 8.4, R 2.9.2 and 2.11.1, and QGIS 1.0.1 on a computer with an x86_64 Ubuntu Jaunty 9.04 and subsequently Lucid 10.04 operating system.

2.1.2 Data

The data for this study is drawn from three main sources. The first is *County Business Patterns* (U.S. Dept. of Commerce, Bureau of the Census 1986, 1987, 1992, 1998, 2003, 2004, 2007, 2009), which is prepared annually by the U.S. Bureau of the Census and contains information on total annual pay and total employment by firm size. Data sets have been chosen in five year increments from 1974 to 2004, which avoids all major recessions during the time period, coincides every second time with the collection of census data, and is a sufficiently long period to observe statistical change. Additionally, 1970 and 2007 data are appended. While these do not adhere to the five year increment, they are respectively the earliest and latest data available at the time and functional for the purposes of this analysis. The second data set consists of U.S. Census data generated by the National Historical Geographic Information System (Minnesota Population Center 2004) and linked to their historical geographical files, which allows for the calculation of distances between points (e.g., county centroids, which are the geographical center of a county, and international airports). Tax rate figures are collected from the Council of State Governments (1978, 1984, 1988, 1994, 1998, 2004), which lists state level corporate tax rates for each state. Two notes are in order with regard to the tax rate information. First, though there are often figures for small businesses and large business, the rates for large businesses are used, as these firms are the most likely to make decisions on the basis of tax rates. Second, these numbers do not take into account any incentives offered to particular industries and thus may underestimate the impact of these rates for those industries.

Data is all at the county level, as this is the only level at which sufficient and intertemporally consistent data on employment by sector is available.

Despite the stability at the county level, it is important to bear in mind that many industrial classifications are inconsistent over time as the government redefined them to account for contemporary economic change. This is particularly so for the 1997 reclassification from the Standard Industrial Classification (SIC) to the North American Industrial Classification System (NAICS) to unify the Canadian, Mexican, and American systems in the wake of the North American Free Trade Agreement (NAFTA). As a result of the reclassification, many industrial sectors cannot be compared across this gap. To gain the most insight despite this gap, the following analysis makes every effort to choose classifications that are consistent throughout the series. As a consequence, industrial sectors are often narrower than ideal.

2.1.3 Mapping

The maps in this section are all based on the *County Business Patterns* data linked to NHGIS map files via postGIS. Numbers for location quotients were generated in PostgreSQL by dividing the ratio of county level employment in the given industry (in this case, warehousing) by total employment in the county and then dividing this by the same ratio for the country as a whole. Thus values greater than one indicate an above average concentration of employment in the given sector, and values less than one indicate below average concentrations. In this manner, it is possible to trace the relative importance of a given industry to a county's wellbeing.

2.1.4 Regressions

The regressions add geographical variables to analyses that generally rely solely on socioeconomic indicators to predict industrial location. Most evaluations of industrial location include variables such as education, race, foreign-born population, and wage rates with politically determined tax rates as predictors. The regressions here add a measure of distance from three types of transportation infrastructure nodes to county centroids. The first of these is ports. Based on the latitude and longitude provided by the UN location code system (United Nations Economic Commission for Europe 2007) and augmented with information from World Port Source (2009), all major North American ports listed in Containerisation International (2006) were included. This necessarily excludes some small ports that handle containers, but none of those with foreign calls. The second measure of distance is from county centroids to airports with customs landing approval from the FAA as of 2002 (Federal Highway Administration (FHWA) 2007). This excludes many local airports, but ensures that all airports approved to serve as ports of entry for freight are included. The final geographical variable is distance from county centroid to intermodal terminals in 2002, based on data in Federal Highway Administration (FHWA) (2007). Intermodal terminals serve as points for the transfer between rail and either ships or trucks. Thus these terminals are located both on the coast and inland and capture the relative importance of rail to the location of economic activity. One caveat must be added, however. As far as the author has been able to determine, the only complete maps for these infrastructure nodes date from 2002. Therefore, all of these nodes are treated as if they existed throughout the time period studied. This is highly unlikely, but it may well be the case that most of these facilities existed in one form or the other throughout the period.

The operative hypothesis, as will be expanded in Chapter 4, is that containerization has reduced the likelihood that economic activity will locate near ports and will instead migrate toward the new break bulk points determined by concentrations of warehouses. Prior to containerization, most bulk was broken at the pier and then shipped off to factories and warehouses. With the reduction of time and labor required to move freight through ports, goods are now generally moved in containers to inland warehouses, where they are opened and the goods redistributed. Thus, if economic activity is coupled to break bulk points, whether for convenience or cost, then during the era of containerization, economic activity should be moving away from ports.

Variables used in the regressions are listed and described in Table 2.1.

Table 2.1: **Regression variables**

Variables	Descriptions
emp_log	Log of total employment in selected industry classification
pop_density_log	Log of population density, which reflects the relative urbanization of the county
inc_per_cap_log	Log of per capita income, which reflects wage levels and purchasing power
taxrate	Tax rate from previous period (except 1970 and 1974, which use rate from 1974) in percent
ba	Percent of population with bachelor's degrees or higher
hs	Percent of population with a high school diploma
nonwhite	Percent of nonwhite population
foreign	Percent of foreigne population
port	100km from county centroid to closest port
airport	100km from county centroid to closest international airport
intermodal	100km from county centroid to closest rail-based intermodal terminal (as recorded in 2000)

2.1.5 Notes for interpreting plots and regressions

A few tips on interpreting the plots and regressions are in order to facilitate the reader's understanding of the results of the analysis.

Since the regressions are logarithmic (based on clear trends indicated in preliminary pairwise scatterplots), a one unit change in the coefficients of non-logged independent variables can be interpreted roughly as the percent change in employment. The further away from zero the values get the more important it is to exponentiate the estimator, bearing in mind that the impact is multiplicative, not additive.

Second, the plots have vertical markers to represent reclassifications. The widest is the 1997 transition to NAICS. This should help the reader avoid making too many quick assumptions about the underlying numbers. There is also a horizontal grey line at $y = 0$ (for those plots that include it). This line is provided as a reference not just to the magnitude of effects but also to aid in evaluating the significance of an estimator. The vertical lines through the estimator points represent the 50th (thick) and 95th (thin) percentile confidence intervals. If these lines pass above or below zero, the estimator is not significant at that level. This will hopefully save the reader some unnecessary flipping back and forth between the plots and the regressions.

Finally, a good rule of thumb for interpreting the plots is that if the red trend line is increasing, employment is moving away from ports, airports, or intermodal terminals, as appropriate. And vice versa, if the line has a negative slope, the economic activity is getting closer. Negative values imply a concentration that trails away, while positive values indicate increasing employment the farther one gets. As values approach zero, they indicate an increasing geographical deconcentration (an evening out) of the economic activity with respect to the infrastructural node in question.

However, one should note that the confidence intervals, while often highly significant in any given year's regression, are wide enough that any interpretation of trend is indicative at best. Most estimators in most regressions have confidence intervals that overlap with each other, reducing the statistical possibility that the true value is actually moving. That said, some trends seem apparent, but should interpreted with great caution.

2.2 Historical analyses

The material for Chapters 3 and 8 is drawn from a mixture of primary and secondary sources. Secondary sources are cited throughout the text, but particularly important sources include Finlay (1988), Herod (2001), and Hobsbawm (1964). Primary sources have provided invaluable information on working conditions, labor processes, and technological trends, particularly for turn of the century dockers. Barnes's (1915) groundbreaking report for the Russell Sage Foundation, the first systematic study of dock work and dockers in the US, proved invaluable. The U.S. Bureau of Applied Economics (1920) and National Industrial Conference Board (1921, 26) supply information on standards of living at the time, from which the living conditions of longshoremen can be deduced. Copies of *The Waterfront Worker*, the radical newspaper by and for longshoremen on the West Coast in the 1930s, offer labor's perspective on the most vital episode in longshore unionization. And an International Labour Organisation working draft (International Labour Office 1973) helped identify the core components of mechanization and modernization agreements and their relevance and attractiveness around the globe.

Information on cargo handling efficiencies and port performance are drawn from Bureau of Labor Statistics bulletins, especially U.S. Bureau of Labor Statistics (1932) and Maritime Cargo Transportation Conference (1957), and an evaluation of Canadian ports by Picard (1967) for the Minister of Labour. Contemporary journals, including the *Journal of Commerce* and the compellingly titled *Cargo Handling*, also supply information on cargo handling technologies and processes and insights into the agreements and disagreements among employers on the future direction of those technologies.

Material for the Chapter 5 on port competition was drawn primarily from the *Journal of Commerce* and *The New York Times*. Efforts were made to contact actors involved in the struggle for Maersk's terminal, but they were universally rebuffed. The Maryland Ports Authority (MPA) was unable to "find anyone at the MPA who was close enough to the deal or has the institutional knowledge regarding the Maersk deal" (Miller 2009). Maersk representatives refused to return my phone calls. And the Port Authority of New York and New Jersey, after a preliminary interview, was unable to cooperate with me because of a sensitive, ongoing lawsuit related to the deal, which is touched upon in the chapter. However, the two sources provide sufficient numbers of direct quotes and relevant analyses that I believe provide sufficient grounds for the claims required by this document. Also, most of the information on PANYNJ revenues, expenses, and volumes were taken directly from the Authority's annual reports (Port Authority of New York & New Jersey 2006).

The minutes of the North Atlantic Ports Association (North Atlantic Ports Association 1999), housed at SUNY Maritime's library in Bronx, NY were also investigated in regard to the form and results of interport cooperation.

2.3 Reproducibility and literate programming

This project has adopted a foundational methodological approach by joining a contemporary movement that favors reproducible research and literate programming.

The ability to reproduce experimental results has long functioned as the cornerstone of the natural sciences. As Popper (1959, 23) writes:

> We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them. Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence', but with events which, on account of their regularity and reproducibility, are in principle inter-subjectively testable.

This ideal, of course, is not so simply implemented. In the natural sciences, the particular equipment in a particular laboratory or prohibitively expensive equipment can bar the reproduction of scientific results. In the social sciences, where researchers examine complex interactions incapable of being isolated effectively from their environments, reproducibility is an impossible goal. This is primarily because social interaction takes place in perpetually evolving social situations (Sayer 1992) and there is no way to turn back the clock on these changes (Prigogine 1997).

In one area of the social sciences, however, a limited form of reproducibility is possible: statistical analysis. By opening up the datasets and the programming code used to evaluate the data, other researchers can test the coding and the analysis by running it on their own machines. In this way, problematic coding and, to a lesser extent, errant data can be more quickly identified and rectified. Additionally, it permits other researchers more readily to consider alternative statistical approaches. While opening up data and coding in this way unravels individual monopolies on producing results (and may therefore contribute less to individual careers and encourage free riding), it is clearly superior as a way of contributing to knowledge.

As a consequence, the coding itself should take on the characteristics of a well written essay, so that it is readable to those unfamiliar with the coding itself. This concept of "literate programming" was introduced by Don Knuth in 1984 and has resulted in the use of clearly understandable variable names and copious commenting. This approach was facilitated by his development of the `WEB` system, which wove quality typesetting with actual programming code (Knuth 1984).

Fortunately, a contemporary approach for doing the same with statistical programming came to my attention as I began preparing this document. It is now possible to weave document coding with statistical coding through two open source programs: LaTeX, which provides sophisticated typesetting, and `R`, which is an open source statistical programming language in wide use. In a distant echo of Don Knuth's original excitement at developing the `WEB` system ("I enjoy the new methodology so much that it is hard for me to refrain from going back to every program that I've ever written and recasting it in 'literate' form." [Knuth 1984, 1]), this document has embraced the combination of LaTeXand `R` using the `pgfWeave` and `SWeave` packages for `R`.