2013

A Critical Spatial Econometric Counter Analysis of the Impact of New Toxic Facilities on Local Wages

Clinton E. Davis
West Virginia University

Follow this and additional works at: <https://researchrepository.wvu.edu/etd>

Recommended Citation

Davis, Clinton E., "A Critical Spatial Econometric Counter Analysis of the Impact of New Toxic Facilities on Local Wages" (2013). *Graduate Theses, Dissertations, and Problem Reports*. 363.
<https://researchrepository.wvu.edu/etd/363>

This Thesis is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Thesis in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/ or on the work itself. This Thesis has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact researchrepository@mail.wvu.edu.

A Critical Spatial Econometric Counter Analysis of the Impact of New Toxic Facilities on Local Wages

by

Clinton E. Davis

Thesis submitted to the Eberly College of Arts and Sciences
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Master of Arts
in
Geography

Approved by

Dr. Jamison Conley, Committee Chairperson

Dr. Jeremia Njeru

Dr. Gregory Elmes

Department of Geology and Geography

Morgantown, West Virginia

June 2013

Keywords: GIScience, Environmental Justice, Toxics, Hegemony, Spatial Econometrics, Counter Analysis

Abstract

Hegemonic assumptions and industry claims regarding the benefits of new toxic industrial facilities on local wages are tested and challenged. Analysis is performed by applying GIScience and Spatial Econometric methods within a critical framework. A call is made to develop more broadly the practice of Counter Analysis, which is presented as an analogue of Counter-Mapping that repurposes Spatial Analytic methods for the pursuit of critical research agendas.

Acknowledgments

Thank you, Dr. Jamison Conley, for your constant guidance and unmerited patience.

Thank you, Dr. Jeremia Njeru, for asking the right questions and encouraging me to answer them.

Thank you, Dr. Gianfranco Piras, for showing me a better way to design my analysis.

Thank you, Dr. Greg Elmes, for teaching me to make good maps and your willingness to help.

Finally, thank you, students and faculty of the WVU Geography Program, for your camaraderie and for facilitating a safe place to learn and take academic risks.

Contents

Acknowledgments	iii
Table of Contents	v
List of Figures	vi
List of Tables	vii
Introduction	1
1 Conceptual Background	3
1.1 Environmental Justice	3
1.2 Political Ecology	4
1.3 Geographic Information Science	5
1.4 Counter Analysis	6
2 Analysis	8
2.1 Data	9
2.1.1 EPA Toxic Release Inventory	9
2.1.2 Internal Revenue Service Statistics of Income	10
2.1.3 Other Data Resources Used	11
2.2 Methods	12
2.2.1 Difference-in-Differences Background	12

2.2.2	Treatment and Control Groups Selection	14
2.3	First Difference-in-Differences Analysis - All Counties	21
2.4	Second Difference-in-Differences Analysis - Filtered	21
2.5	Analysis of Percent Change	23
2.6	Third Difference-in-Differences Analysis - Wages Standardized by Number of Returns	23
2.7	Discussion of Results	24
3	Conclusion	30
3.1	Results of Analysis	30
3.2	Reflection on Integration of GISc, PE, and EJ	31
3.3	Implications and Directions for Future Work	32
	Bibliography	38
	Appendix	39
3.4	Python Code	39
3.5	R Code 1	48
3.6	R Code 2	53

List of Figures

2.1	Difference-in-Differences	13
2.2	Counties were classified into 6 clusters using hierarchical clustering techniques.	16
2.3	Cluster Attributes 1 of 2, Counties' Attributes	17
2.4	Cluster Attributes 2 of 2, Counties' Neighbors' Attributes	18
2.5	Treatment counties are those that received one or more new TRI facilities in 2000. Control counties did not receive any new TRI facilities in 2000.	19
2.6	Treatment and Control Counties by Cluster	20
2.7	First D-i-D: All Counties, Total Wages	26
2.8	Second D-i-D: Filtered, Total Wages	27
2.9	Percent Change in County Wages from 1999-2005	28
2.10	Third D-i-D: Wages Standardized by Number of Returns	29

List of Tables

2.1	Census data variables used for hierarchical clustering	15
2.2	Coefficients and P-values for First D-i-D Analysis: All Counties, Total Wages	22
2.3	2000 Population of Treatment and Control Counties by Cluster	22
2.4	Coefficients and P-values for Second D-i-D Analysis: Filtered Counties, Total Wages	23
2.5	Coefficients and P-values for Third D-i-D Analysis: All Counties, Wages standardized by Number of Returns	24

Introduction

The siting and construction of new toxic facilities can be highly contentious processes. Facilities are often expected to bring economic benefit, health and environmental risk, or both to the host communities. Proponents argue that the new facilities create jobs and generate wealth for the community (Wheeler 2012; Panda Energy, Inc. 2007; Shauk 2012; Frazier 2012) while opponents argue that such facilities put residents in jeopardy of adverse health effects and contribute to environmental degradation of the property and community (Shen 2012; Shipp 2012; Novak 1996). Often those being impacted by the decision are not fully aware of the likely costs and risks, but are left to form opinions based on the information provided to them by other stakeholders. The opinions they form will largely determine the nature and extent of their involvement in the planning process, deciding whether they mobilize for or against the construction, or if they do nothing at all. In order for community members to be fully involved in the process of determining the changes to their neighborhoods, it is important that they understand both the risks and benefits that they are likely to experience as a result of those changes.

There is a large body of research both within Geography and elsewhere in the academy focused on the health risks of new facilities, but there has not been the same level of critical examination of the assumed economic rewards. This thesis employs GIScience techniques to question the county level impacts of new facilities on local wages. It also serves as a case study of the integration of GISc*, Political Ecology, and Environmental Justice. A basic history of the three traditions and their interactions is presented, followed by a discussion of the route through which they are combined to achieve the work done in this thesis. Results

*The “GISc” notation will be employed throughout the remainder of this paper as a tacit reminder that this is a social scientific analysis of geographic information while not emphasizing too strongly the title of science and the cultural connotations of empirical authority claims associated with it.

of the analysis are then presented and are interpreted as a challenge to hegemonic notions associating local wage benefits with new facilities. The final section reflectively discusses the integration of GISc methodology with a Political Ecology research framework and the possible utility of such work to the Environmental Justice movement and activists.

Chapter 1

Conceptual Background

1.1 Environmental Justice

In 1982, resistance to the siting of a hazardous waste landfill in a low income, predominantly African American community in Warren County, NC brought awareness to what has since become known as the Environmental Justice movement (Cutter 1995). The movement, built on the civil rights movement and led largely by women of color (Shrader-Frechette 2002), seeks to resist the systematically racist environmental policy decisions that put communities of color and low income communities at disproportionate risk. Often the risks addressed by the Environmental Justice movement are those associated with exposure to toxins due to the situation of hazardous waste dumps and hazardous industrial facilities.

The movement was bolstered by the 1987 United Church of Christ Commission for Racial Justice (UCC-CRJ 1987) report entitled *Toxic Wastes and Race in the United States*, the report gave the results of a statistical study in which “race proved to be the most significant among variables tested in association with the location of commercial hazardous waste facilities” (UCC-CRJ 1987). The report went on to call this a “national pattern” and emphasized that “while socio-economic status appeared to play an important role ... race still proved to be more significant” (UCC-CRJ 1987).

The next major milestone for the movement was in 1991 when the UCC-CRJ-sponsored People of Color Environmental Leadership Summit resulted in the publication of *The Principles of Environmental Justice*.

The summit brought together leaders “to begin to build a national and international movement of all peoples of color to fight the destruction and taking of our lands and communities” (PCELS 1991).

Since the 1991 Summit, the Environmental Justice movement has grown both in prominence and in power. In 1994 President Clinton issued Executive Order 12898 tying together the Civil Rights Act and the National Environmental Policy Act to forbid federal money from being used in environmentally unjust circumstances (Bullard 2005). Robert Bullard said that in 2002 there were “a dozen environmental justice networks, four environmental justice centers, and a growing number of university-based legal clinics that emphasize environmental justice.” He also noted the development of graduate degrees in Environmental Justice at the University of Michigan (Bullard 2005).

Despite the growth of the Environmental Justice movement, the issues that it is resisting are still present. In its 2007 follow up to the 1987 report, the UCC-CRJ reported that “people of color make up the majority of those living in host neighborhoods within 3 kilometers of the nation’s hazardous waste facilities,” “host neighborhoods of commercial hazardous waste facilities are 56% people of color whereas non-host areas are 30% people of color,” and “Poverty rates in the host neighborhoods are 1.5 times greater than non-host areas” (Bullard et al. 2007).

1.2 Political Ecology

Some of the Environmental Justice research by geographers has been done through Political Ecology frameworks. Political Ecology is a multidisciplinary field made up of researchers working in many social and natural science disciplines who examine the intrinsically political nature of the interactions between human societies and their natural and built environments. Within geography, political ecology research is approached from many critical geography frameworks. As explained in the *Dictionary of Human Geography*, political ecology is “far from a coherent theory” and includes work from resource management, environmental history, science studies, actor-network theory, gender theory, discourse analysis, and Marxist Geography (Johnston et al. 2000).

Political Ecology work in Environmental Justice often works not only to demonstrate the existence of injustice, but also to advocate for the people and communities impacted by it (Njeru 2006; Heynen, Perkins,

and Roy 2006). In this way, political ecologists often have roles as activist researchers, doing work that both produces knowledge and works toward explicit political agendas. This advocacy is largely focused on resistance to environmental racism and support for environmental equity movements.

1.3 Geographic Information Science

In the late 1980s and early 1990s, Geographers began to distinguish the study of Geographic Information Science (GISc) from work on the use of GI Systems (Goodchild 1992). By differentiating the science from the software and systems, researchers were able to establish GISc as an arena of study in its own right as well as a set of methods and approaches for studying broader social and environmental research topics. The process of establishing GISc has also included the exploration of various ways of using and conceiving of geographic information as well as various ways of connecting GISc techniques and processes with other disciplines within Geography. Since the early days of the discipline, one challenge that has been recognized is the complexity and need for integrating GISc with human geography and social theory (Sheppard 2005) including Environmental Justice and Political Ecology (Harris and Weiner 1996).

The joining of GISc with Environmental Justice has largely been done by simply employing existing empirical GISc frameworks to Environmental Justice topics and situations. It has been used to demonstrate the existence of unjust environmental practices (Jerrett et al. 2001; Mennis 2002; Maantay 2007), to assess risk (Dolinoy and Miranda 2004; Maantay 2007), and to document health consequences of contamination (Maantay 2007; Conley 2011). By applying GISc methods to these questions, researchers have sought to use empirically produced knowledge to support Environmental Justice work by confirming the existence of structural injustices and exploring the realities of risk.

Meanwhile, other aspects of GISc have more thoroughly embraced critical geographic frameworks for study. Public Participatory GIS has introduced understandings of situated and informal knowledge to GISc as has work using GISc methods in feminist geography research (Kwan 2002). Likewise, studies of GIS and Society have employed social theory to examine GIS itself and the co-production that occurs between social development and the creation and evolution of Geographic Information Systems (Harris and Weiner 1998; Schuurman 2000).

1.4 Counter Analysis

Another way that GISc practitioners have integrated their work with critical geography and social justice is through creating maps that contradict the positions of the powerful in favor of advocating for oppressed and marginalized people and communities. Counter mapping has been applied to a diverse variety of situations including indigenous forest rights in Indonesia (Peluso 1995), the hazard of suburban commuters to young inner city pedestrians (Bunge 1988), and resistance to colonial place naming in the Arab Homeland (Culcasi 2012). This process of counter-mapping has been a bridge between Cartography and human geography by the reciprocity of using social theory to guide the way in which maps are made and by use of those maps to achieve the aims of the critical geographers. In this thesis work, I explore the use of a similar bridging of GISc and critical geography, but in this case rather than focusing on the cartographic functions of GIS, I employ spatial analysis techniques to pursue a critical agenda. I refer to this process as Counter Analysis to connect with the deep tradition of counter-mapping while emphasizing the focus on analysis rather than, or in addition to, mapping.

In contrast to the previously mentioned applications of GISc to Environmental Justice topics, this thesis does not aim to confirm the existence of environmental injustice or to describe its nature. Instead, this work accepts the previous evidence that toxic facilities are more likely to adversely impact low income communities and communities of color as a motivation to resist the addition of new facilities, but does not, itself, attempt to assess equity of harmful outcomes. Specifically, it follows Woods and Gordon's work employing GISc methods to challenge the hegemony of "coal means jobs" in West Virginia by challenging the more general hegemonic notion that new toxic facilities, while bringing ecological and health risks, can also be expected to bring increased wages to the counties that host them (Woods and Gordon 2011).

The idea of hegemony, developed by Antonio Gramsci, refers to a form of coercion employed in service of power in which ideas that benefit the powerful are accepted by the populace as common knowledge and therefore effect support for actions that benefit the elite even if the actions are in reality detrimental to the proletariat supporters themselves (Johnston et al. 2000). This work takes the jobs argument in support of new toxic facilities as a hegemony that obscures the ability of communities to make rational decisions about the placement of new facilities either by bringing about support or by simply discouraging opposition. This

work challenges that hegemony by testing an alternative hypothesis that new facilities have an impact on local wages against the null hypothesis that they have no significant impact. By challenging this hegemony I hope to encourage communities to examine individual proposals on a case by case basis and with a full understanding of the local impacts, both positive and negative, when considering allowing the construction of a new facility rather than accepting empty assurances that there will be economic benefits based solely on the common sense understanding that new facilities bring jobs and money. I also hope to encourage GIScientists to explore the possibility of using GISc methods to pursue critical geography research, that is progressive, social justice oriented research, by not only looking at the topics that critical geographers look at, but by asking the same questions.

Chapter 2

Analysis

Planning debates around the placement of new toxic facilities are often framed as debates between ecological and health risks versus economic rewards; pollution and sickness versus jobs and money (Panda Energy, Inc. 2007; Shauk 2012; Wheeler 2012; Frazier 2012). Previous work has assessed the risks (Dolinoy and Miranda 2004) and made the argument that they are so great that they outweigh the potential economic rewards (Templett and Farber 1994), but little work has been done to question the validity of the reward claims themselves. As we transition further into a post-industrial society, it is becoming more and more important to question whether new toxic facilities actually bring significant wage impacts to local economies. This research assesses those questions and provides a framework and analysis through which to pursue critical examinations of the “jobs and money” claims of toxic corporations and developers. Since planning decisions often happen at the county level, this work will look specifically at the impacts on wages at the county level. While it would be scientifically interesting to examine spill-over effects, the impact of facilities in one county on surrounding counties, they will not be pursued in this project because it is specifically meant to provide a resource to county planners considering the impacts on their own county of allowing the construction of a new facility within the limits of their county. It would be likewise interesting to examine the hyper-local impacts of new facilities on the neighborhoods in which they are placed, but again, since these decisions often happen at county planning boards, the examination will be limited to the county level. Both of these related questions have merit and deserve to be further investigated, but are not within the scope of this

project. This research focuses solely on questioning the validity of the common assumption that new toxic facilities will benefit the economy of the county by increasing wages by examining the nature and significance of the effect that new facilities have on wages in their counties. Therefore, the analysis will take as a null hypothesis that the addition of new toxic facilities does not affect the wages within counties and will test for support of an alternative hypothesis that it have an effect.

2.1 Data

This analysis relied mainly on two large nationwide, multi-year, and publicly available datasets. Both sets came as spreadsheets with companion shape (*.shp*) files of location data. The first is a comprehensive list of EPA sites of interest (EPA 2012b), from which I have extracted the subset of Toxic Release Inventory (TRI) reporting sites. The second is the Internal Revenue Services (IRS) Statistics of Income (SOI) which includes the total wages and salaries reported on tax returns each year for each county in the United States (IRS 2012).

2.1.1 EPA Toxic Release Inventory

In 1986, Section 313 of the Emergency Planning and Community Right-to-Know Act established the Toxic Release Inventory (TRI) as a means of increasing public knowledge of the risks they faced as a result of toxic chemical releases and transfers. The inventory is administered by the US Environmental Protection Agency (EPA) with data being self-reported by the facilities. Every facility in a “TRI-covered industry sector” that employs at least 10 full-time employees and “manufactures, processes, or otherwise uses a TRI-listed chemical in excess of the applicable threshold quantity for a given calendar year” is required to file a TRI Report with the EPA by July 1 of the following year (EPA 2012c). The reporting thresholds vary depending on the risk associated with the chemical from 0.1 grams for dioxins up to 25,000 pounds for chemicals not considered “persistent, bioaccumulative, [and] toxic” (EPA 2012a).

The EPA compiles those reports into the TRI which is in turn released to the public in December of the following year, and at the time of this research are available for reporting years 1987-2011. The reports list the compounds released, the amounts, the means of release, the locations of the facilities, and the facilities’

owners. Additionally, comprehensive spreadsheet files, consisting of each chemical reported by each facility in the country, are available for each reporting year.

The Facility Registry System (FRS) is a listing of all sites in which the EPA has an interest, including TRI facilities. Unlike the yearly TRI spreadsheets, this set is not concerned with which chemicals are used by sites, but rather serves as an index that links facility IDs with information about the facilities. It consists of several spreadsheets containing various pieces of information about the sites, including the National Environmental Interest File which lists the start and end years of reporting along with the facility interest type. The reporting years was used in this analysis as the record of the years of operation and the facility type was used to separate TRI facilities from other types included in the the spreadsheet, but not included in this analysis.

Since the original *.csv* file is too large to open in an ordinary spreadsheet editor and contains information on many sites of interest for reasons other than TRI reporting, TRI reporting sites were selected and exported with a custom Python script. The result was a list of 55,454 facilities that reported to the TRI between the years of 1987 and 2010.

The TRI spreadsheet data was jointed to a nationwide FRS site location shape file obtained from the EPA website which includes the 54,139 TRI sites. Using an EPA shapefile of facility locations addressed one of the major difficulties of spatial analysis of TRI data: being sure that the address used for geolocating is the facility rather than the corporate headquarters.

2.1.2 Internal Revenue Service Statistics of Income

Each year, the IRS releases personal income statistics aggregated to the county level (IRS 2009). The data are based on personal tax return 1040 forms. The data represent 97 percent of all tax returns and include information on Wages and Salaries, Adjusted Gross Income, Income from Interests, and details on the total number of returns and exemptions claimed which approximate households and individuals respectively. The data are provided in spreadsheet form for each year and county entries are labeled with state and county Federal Information Processing Standard (FIPS) codes, which enables them to be joined to county shapes from the US Census Bureau (U.S. Census 2012).

Data are available for the years of 1989 through 2009. Data files for 2004-2009 were obtained freely from the IRS website and the data for 1989-2003 were purchased directly from the IRS (IRS 2012). The purchased data arrived as individual csv files for each state for each year. These were processed into a single csv file with Python. After the files were combined, a small amount of manual post-processing was required to bring all data into a standardized format. This csv was then joined to the data for 2004-2009 to create one master file of wages for all years.

It should be acknowledged regarding the IRS data do not cover all income, only that which is reported on 1040 forms. In particular, the data only include wages from formal employment and do not account for wages in the informal economy. It is also important to note that counties with fewer than four tax returns are combined with other counties to protect individuals from identification.

2.1.3 Other Data Resources Used

In addition to the two main datasets, I also employed data from the 2000 Census (U.S. Census 2001a). County outline shapefiles (U.S. Census 2012) were obtained from the census and were utilized for the purpose of discerning neighbors and for display of results. Prior to use, the shape files were dissolved in *ArcGIS 10.1* by county name, so that each county was represented by a single polygon comprised of one or more parts. This was especially important for counties that include multiple islands because each island was originally represented as a single polygon.

Additionally, data from the 2000 Decennial Census itself was used for the purpose of building treatment and control groups for use in the difference-in-differences statistical analysis. Demographic identifiers were used from both the short (Summary File 1) and long (Summary File 3) forms. The short form data is a full summary of individual data collected by the Census Bureau for all residents (U.S. Census 2001a) while the long form data is a summary of results of survey data from a sample of residents (U.S. Census 2001b).

2.2 Methods

2.2.1 Difference-in-Differences Background

In order to best determine the impact of new toxic facilities on counties' incomes over time, it is important to isolate the effect of the new facility from other effects. I used the difference-in-differences (d-i-d) technique to accomplish this (Angrist and Pischke 2008; Shafrin 2006; Bertrand, Duflo, and Mullainathan 2004). D-i-d uses before and after measurements on treatment and control groups to isolate the impact of treatment events from non-treatment oriented time and group effect. The control group is chosen such that under normal circumstances, control subjects should be expected to behave similarly to subjects in the treatment group. Therefore, changes in the control group can be used to forecast expected changes in the treatment groups in the absence of treatment events. The forecasted changes in the treatment groups yield what is called the counterfactual values of the response variable. The difference between counterfactual and observed values for the treatment group at the end of the experimental period gives the effect of the treatment itself. Once control and treatment groups are established, the difference-in-differences analysis can be performed in two ways. The first compares average initial and final values for each group, while the second employs dummy variables in a regression. Both methods yield the same results, but the regression method adds the benefit of p-values for significance testing. This discussion will begin by explaining the simple version, then moving on to the regression-based method.

Simple D-i-D Method

In the simple d-i-d method, average values of the response variable, y_{igt} (where i is an individual observation, g is group, and t is time), are calculated for control group prior to the treatment ($\bar{y}_{c,t_0}$) and at time t_1 after the treatment has been effected ($\bar{y}_{c,t_1}$). Likewise, average values are calculated for the average response variable values for the treatment group before ($\bar{y}_{t,t_0}$) and after ($\bar{y}_{t,t_1}$) treatment. The initial difference between control and treatment group averages ($\alpha = \bar{y}_{t,t_0} - \bar{y}_{c,t_0}$) represents the non-treatment related group effects, since it is measured prior to the treatment taking place. The difference between the initial and final control group values ($\beta = \bar{y}_{c,t_1} - \bar{y}_{c,t_0}$) gives the non-treatment related time effect, since it is the effect of time on the untreated control group. The counterfactual forecast for expected value of the treatment group at t_1

in the absence of an actual treatment is calculated by adding the average of the initial value for the control ($\gamma = \bar{y}_{c,t_0}$) to non-treatment related group (α) and time (β) effects: $y_{cf,t_1} = \gamma + \alpha + \beta$. This counterfactual value is then subtracted from the average of the observed values of the response variable for the treatment group at t_1 to obtain the actual treatment effect δ : $\delta = \bar{y}_{t,t_1} - y_{cf,t_1}$. This value, δ , is the difference between change in the treatment group and the change in the control group, or the difference-in-differences (Angrist and Pischke 2008). Figure 2.1 gives a graphical representation of the analysis process.

Figure 2.1: Difference-in-Differences

Regression D-i-D Method

In the regression-based d-i-d, all values for individual units are used with dummy variables for each type of effect as shown in equation 2.1. As in the simple form, y_{igt} is the response variable for unit i in group g at time t . The intercept, γ , gives the initial value of the control group. α is the dummy variable for group effect and is assigned a value of 1 for treatment units or 0 for control units. The dummy variable for time effects, β , is assigned a value of 0 for measurements prior to treatment and a values of 1 for post-treatment measurements. δ is the interaction between group and time effects: 1 for treatment group after treatment, 0 for all other measurements. ϵ is the subject-level error term. The coefficients and accompanying p-values yielded by the regression demonstrate the nature and significance of each effect on the response variable. It

is the δ coefficient, and accompanying p-value, that is of interest in this research since it describes the effect of the treatment on wages.

$$y_{igt} = \gamma + \alpha A_g + \beta B_t + \delta D_{gt} + \epsilon_{igt} \quad (2.1)$$

2.2.2 Treatment and Control Groups Selection

Demographic data from the 2000 Census were used to classify counties into groups with similar characteristics which should, therefore, behave similarly under normal circumstances. Since county economies are strongly linked to the economies of neighboring counties, the demographics of neighbors was also incorporated into group selection. To accomplish this, it was necessary to first define a set of neighbors for each county. This was accomplished using the *PySAL* spatial weights matrix functionality (Rey and Anselin 2007). A Queen's contiguity conception of neighbors was used, meaning that any county sharing an edge or corner with the county of interest was considered a neighbor. Counties with no neighbors were excluded from the analysis due to their scarcity and unique natures. The *Python* code used to join demographic data to the county shape files and build spatial weights matrices is included in the Appendix.

The counties were then divided into clusters of counties with similar demographics and with similar neighbors using the *hclust* hierarchical clustering function in *R* (R Development Core Team 2012). The data used for the hierarchical clustering were the 2000 Census Data for the counties' population, percent employment, median household income, percent white not hispanic/latino, and the number of neighboring counties along with the aggregated values of these variables for the counties' neighbors (Table 2.1). Population and number of neighbors were summed for neighboring counties while the mean of employment, median household income, and percent white were used. The resulting cluster dendrogram was cut at all values between 2 and 10 clusters for comparison. After examining the distribution of attribute values for each number of clusters, it was determined that using six clusters maintained a good balance of keeping like counties together and unlike counties apart. Cluster locations are given in Figure 2.2 and the boxplots in Figures 2.3 and 2.4 show the distributions of each variable for each cluster.

Once clusters of demographically like counties were established, those clusters were further divided into

counties in which new TRI facilities began operation during the year 2000 (Treatment Counties) and counties with no new TRI facilities in 2000 (Control Counties). Figure 2.5 shows all treatment and control counties and Figure 2.6 shows the treatment and control counties divided out by cluster.

County Data	Aggregate of Neighboring Counties' Data
Population	Sum of Populations
Percent Employment	Mean Percent Empl.
Median Household Income	Mean MHI
Percent White, not Hispanic/Latino	Mean Percent White
Number of Neighbors	Sum of Neighbors

Table 2.1: Census data variables used for hierarchical clustering

Figure 2.2: Counties were classified into 6 clusters using hierarchical clustering techniques.

Figure 2.3: Cluster Attributes 1 of 2, Counties' Attributes

Figure 2.4: Cluster Attributes 2 of 2, Counties' Neighbors' Attributes

Figure 2.5: Treatment counties are those that received one or more new TRI facilities in 2000. Control counties did not receive any new TRI facilities in 2000.

Figure 2.6: Treatment and Control Counties by Cluster

2.3 First Difference-in-Differences Analysis - All Counties

The difference-in-differences analysis was applied to each cluster individually using the IRS County Wages values for 1999 (pre-treatment) and 2005 (post-treatment). 2005 was chosen as the post treatment year to allow time for the facilities to impact the wages in their counties. The treatment counties, those with new facilities in 2000, beat the counterfactual projections by a small amount, as shown in Figure 2.7. This offers some support for the claims that facilities can have impacts on wages, but the small differences cast doubt that they necessarily bring benefit and shows some interesting patterns worthy of further scrutiny.

The regression based d-i-d analysis (Table 2.2) demonstrates that the most significant factors in determining wages are the initial control values (intercept) and the group effect (γ) which are consistently significant across call clusters. The actual treatment effect, the interaction of the group and time effects, is not significant at the 0.05 level for any of the six clusters.

Also, the treatment groups consistently start with higher wages than the control groups. Since wages are a total value, rather than per capita, it is reasonable to suspect that this could be population related. Indeed, it turns out that despite including population in the hierarchical clustering process, the treatment counties still tend to have higher populations than the control groups. Table 2.3 shows the average populations of counties in the treatment and control groups for each cluster. The treatment counties are consistently 2-4 times more populous on average than the control counties for the same cluster. This aspect will be analyzed further later, but first another concern should be acknowledged.

2.4 Second Difference-in-Differences Analysis - Filtered

Since the addition of a new facility during 2000 is taken to be the treatment, but the effect is measured in 2005, concerns arise that facilities added between 2001 and 2005 might impact results as well. In the main analysis, this effect is relegated to the error term because removing counties that have facilities added during the 2001-2005 period dramatically reduces the number of counties available for analysis. It is important, however to acknowledge this critique and investigate its merit. Therefore, the original D-i-D analysis was repeated only using counties that did not receive new facilities during the latter period (for both control

	Cluster 1 Rural North	Cluster 2 Borders	Cluster 3 Appalachia	Cluster 4 Suburbs	Cluster 5 Southern	Cluster 6 Urban
Intercept: γ (initial control wages)	297,045 p<.0001***	343,593 p=0.004**	137,794 p<.0001***	890,569 p<.0001***	269,355 p<.0001***	5,434,306 p=0.013*
Group Effect: α (initial difference: treatment and control wages)	602,096 p<.0001***	1,413,576 p<.0001***	133,123 p<.0001***	1,689,937 p<.0001***	559,139 p<.0001***	8,017,740 p=0.001***
Time Effect: β (change in control wages)	58,455 p=0.0767	83,460 p=0.624	28,574 p=0.165	226,506 p=0.116	58,734 p=0.0852	1,414,367 p=0.647
Treatment Effect: δ (Group X Time) (Difference-in-Differences)	126,173 p=0.0799	276,777 p=0.493	21,457 p=0.650	321,155 p=0.185	117,522 p=0.148	1,424,456 p=0.686
Significance notation: · p < 0.1, * p< 0.05, ** p<0.01, *** p< 0.001						

Table 2.2: Coefficients and P-values for First D-i-D Analysis: All Counties, Total Wages

	Cluster 1 Rural North	Cluster 2 Borders	Cluster 3 Appalachia	Cluster 4 Suburbs	Cluster 5 Southern	Cluster 6 Urban
Control	74480.89	132039.17	32456.74	173709.72	83274.27	878414.58
Treatment	27624.53	31506.26	18311.16	61767.86	29325.82	311137.09
Ratio	0.3708941	0.2386130	0.5641711	0.3555809	0.3521594	0.3542030

Table 2.3: 2000 Population of Treatment and Control Counties by Cluster

and treatment groups). The results were largely similar to those of the previous analysis. Again, the treatment effect was not significant for any cluster and the factors determining wages were largely the initial control values (intercept) and the group effects (γ), although, the time effect (β) had a significant impact for some clusters (Table 2.4). Also, after filtering, the counterfactual projection for cluster 2 outperformed the treatment group, as can be seen in Figure 2.8. In Cluster 2, of the counties that added no new facilities in 2001-2005, those that had not added a new facility in 2000 had a greater increase in total wages than those that added a new facility in 2000, but the treatment effect was still not significant.

	Cluster 1 Rural North	Cluster 2 Borders	Cluster 3 Appalachia	Cluster 4 Suburbs	Cluster 5 Southern	Cluster 6 Urban
Intercept: γ (initial control wages)	156,867 p<.0001***	201,984 p<.0001***	107,562 p<.0001***	381,657 p<.0001***	121,664 p<.0001***	965,760 p=0.425
Group Effect: α (initial difference: treatment and control wages)	113,447 p<.0001***	-17,241 p=0.901	154,942 p<.0001***	433,272 p=0.0016**	168,106 p<.0001***	4,292,523 p=0.054
Time Effect: β (change in control wages)	31,009 p=0.0149*	49,141 p=0.391	22,444 p=0.296	111,895 p=0.138	24,307 p=0.031*	330,969 p=0.845
Treatment Effect: δ (Group X Time) (Difference-in-Differences)	12,199 p=0.738	-12,640 p=0.948	26,324 p=0.623	132,048 p=0.494	31,245 p=0.388	974,639 p=0.74

Table 2.4: Coefficients and P-values for Second D-i-D Analysis: Filtered Counties, Total Wages

2.5 Analysis of Percent Change

Since group effect was a significant determiner of wages and since the treatment groups consistently started with higher total wages than the control groups, the D-i-D method was abandoned in order to investigate the relative percent changes in wages between groups rather than using the counterfactual method. As shown in Figure 2.9, the control groups actually outperform the treatment groups in all but one cluster. While this is by no means evidence that new toxic facilities negatively impact wages, it does offer further encouragement to repeat the initial analysis in a manner that standardizes the data in some way rather than relying simply on the total wages data.

2.6 Third Difference-in-Differences Analysis - Wages Standardized by Number of Returns

Building on the results of previous analyses, the final iteration of analysis examines not the total wages of each county, but the standardized value of total wages divided by number of returns for each county. The result of this analysis, therefore, differs from the previous analyses in that it gives a more individualized look at impacts of facilities. By looking at wage per return, this analysis reflects the average level of pay that can be expected for a resident of the county. It also mediates the difference in populations between control and

treatment groups, allowing for a more direct comparison. As with previous analyses, the effect of adding a new facility (treatment) is an insignificant factor in determining wages, as is apparent by examining the p-values given in Table 2.5. Even more dramatic in this analysis, though, is that in all but two clusters, the treatment counties are actually outperformed by the counterfactual projection as seen in Figure 2.10. These results cast even more doubt on industry claims of benefits to wages.

	Cluster 1 Rural North	Cluster 2 Borders	Cluster 3 Appalachia	Cluster 4 Suburbs	Cluster 5 Southern	Cluster 6 Urban
Intercept: γ (initial control wages)	24.55 p<.0001***	24.55 p<.0001***	20.97 p<.0001***	33.80 p<.0001***	23.06 p<.0001***	38.56 p<.0001***
Group Effect: α (initial difference: treatment and control wages)	3.53 p<.0001***	2.21 p=0.019*	2.466 p<.0001***	2.23 p=0.001**	2.61 p<.0001***	1.04 p=0.629
Time Effect: β (change in control wages)	4.51 p<.0001***	4.39 p<.0001***	3.89 p<.0001***	5.32 p<.0001***	4.35 p<.0001***	6.75 p=0.012*
Treatment Effect: δ (Group X Time) (Difference-in-Differences)	-0.063 p=0.896	0.33 p=0.803	0.6661 p=0.509	-0.558 p=0.569	-0.192 p=0.744	-0.903 p=0.767

Table 2.5: Coefficients and P-values for Third D-i-D Analysis: All Counties, Wages standardized by Number of Returns

2.7 Discussion of Results

While toxic industrial facilities fulfill a function in society and can sometimes bring economic benefit, this research demonstrates that blanket claims that new facilities bring jobs and increased wages to the counties that are considering constructing them can no longer be accepted at face value. As demonstrated in the treatment effect δ values in Table 2.5, the addition of a new toxic facility does not have a consistent impact on wages. In clusters two and three, there was a positive effect associated with the new facilities, while in clusters one, four, five, and six showed negative impacts. In none of the clusters was the effect of the new facilities a statistically significant determiner of wages, meaning that not only is the effect inconsistent, it does not have a strong driving influence on wage values compared to other factors. This is important because

if community members are investing their hopes in these facilities to raise wages, they cannot confidently expect to have those hopes fulfilled.

In light of these results, it is evident that planners and citizens should approach facility proposals with scrutiny rather than assuming that facilities will bring increases to local wages. Claims that facilities will raise county wages should be investigated independently on a case-by-case benefit to be sure that they have merit. While facilities may bring benefit, it is not safe to assume that this is guaranteed. Further, communities should approach cost benefit analyses performed by industry representatives with the skepticism they are due. By starting from a place of doubt regarding the benefits, communities could empower their planners to analyze more closely and with more suspicions the likely wage benefits to the county. A skeptical mindset could also serve planners in examining more realistically the negative impacts of facilities, by removing some of the optimism that could serve to obscure criticism of health and environmental effects of the facilities.

In addition to local analysis of new proposals, researchers should continue to investigate broader trends in wage benefits of new facilities. As the US continues to advance into a post-industrial economy in which jobs are increasingly in the knowledge and service sectors rather than the industrial (Johnston et al. 2000; Mackinnon and Cumbers 2007), it will be increasingly important to interrogate hegemonic assumptions that were formed in an industrial economy about the wage benefits of new facilities, toxic and otherwise. This is significant work that needs to be pursued especially in light of the power politicians find in promising to bring jobs to their constituents.

Figure 2.7: First D-i-D: All Counties, Total Wages

Figure 2.8: Second D-i-D: Filtered, Total Wages

Figure 2.9: Percent Change in County Wages from 1999-2005

Figure 2.10: Third D-i-D: Wages Standardized by Number of Returns

Chapter 3

Conclusion

3.1 Results of Analysis

The core of this research has been the testing of the hypothesis that new toxic facilities impact local wages. Through the application of difference-in-difference analysis, wage trends in counties that received new facilities were compared against expected trends in the absence of new facilities based on control counties that did not receive new facilities. Through three increasingly refined models, the effect of adding new toxic facilities consistently failed to be a significant predictor of wages compared to initial values and time and group effects. Further, in the most refined model, presented in Section 2.6, four of the six clusters showed a negative, though still not statistically significant, effect from the addition of new facilities.

All analysis examined impacts at the county level. While actual impacts of facilities could exist at other scales, planning decisions are often made by counties so county level impacts are important to the decision making process concerning the review of placement proposals for new facilities. This is further significant for communities that are marginalized within their own counties and lack full representation in county decisions. Where planners might not be concerned about negative externalities impacting a small portion of their constituency if they expect broad economic benefits for the county, they may be more sympathetic if they approach the decisions with skepticism of broader wage increases.

The consistent insignificance of treatment effect in predicting wages throughout the various Difference-

in-Differences setups along with the existence of some negative δ values demonstrates the fallacy of assuming that new toxic facilities will bring increases in local wages. Therefore, it is important that residents, county planners, and researchers consider the merit of individual proposals for new toxic facilities on a case by case in terms of both risks and benefits. To achieve this, it is important to conduct research on proposed and existing facilities in an embedded manner; individually and with full concern for the uniqueness of local conditions and with support of local knowledge. Furthermore, activists and educators should work to challenge hegemonic jobs and money claims by developers while encouraging citizens to develop informed opinions on the costs and benefits of proposals.

3.2 Reflection on Integration of GISc, PE, and EJ

By setting up a confirmatory statistical study to test the hegemonic notion that toxic facilities bring wage increases, we have been able to reframe the conversation from an argument against an accepted knowledge to a failure of that assumption to prove itself to be true. By placing the onus on the benefit claims, we have been able to negate the validity of the claims while simultaneously challenging their status as beyond question. This method only works by making a legitimate attempt to verify the hegemonic claims with the same rigor that would be applied if the claimants took up the task of defending their claims. Therefore, it is important to be sure that the analysis is beyond reproach and repeatable. By using publicly available data and open source, scripted analysis, it is possible to make available not only the findings, but the process through which they were developed. This allows others to pursue and adapt the analysis to test the robustness of the rejection of the claims by running the analysis for different setups in the same way that I analyzed for multiple setups within my research.

By challenging a hegemonic assumption, we also developed an opportunity to advocate for more embedded research on the issue and on increasing public awareness and community self-determination by offering an opportunity to subvert hegemonic control over the opinions of the populace. This goal can only be achieved, though, if the results are made available in such a way that they are approachable to three distinct audiences and purposes: academics to be used to advocate for more research, resistance workers to use in debates, and to communities at large to contribute to broader education on the matter. Thus far, only the first objective

has been met, so direct evaluation of the effectiveness of these methods at reaching those goals is not yet possible. These objectives have, however, been at the forefront of guiding the research from the beginning and have dictated choices of data, software, methods, and presentation in an effort to balance rigor with approachability and to leave room to tailor final published products to a variety of audiences and formats.

The success of the work in introducing counter analysis as a framework for bringing together critical geography and GISc in this thesis is, likewise, incomplete. While it has been successfully employed as an internal framework for locating my GISc based analysis within a broader social justice agenda, its more general success will depend on the impact that the work has on the GISc and critical geography communities upon publication. The work in that sense is only just begun, because while it has been demonstrated, it has not yet been subjected to critique nor has it had yet had the opportunity to develop past a single application. The long term success of establishing counter analysis within geography, will likely depend, again, on balancing rigor and approachability. This will be doubly difficult, because it must satisfy both academic communities. The rigor of the analysis must satisfy the GISc community while the rigor of application of theory must satisfy the critical geography community. Likewise, the critical theory must be presented in such a way as to be accessible to GIScientists and the methods and results must be accessible to critical geographers. This thesis has tended emphasize the GISc rigor more than the theoretical rigor, but as the work progresses it will be necessary to develop further the theoretical aspect.

3.3 Implications and Directions for Future Work

In both aspects of this analysis, the results are encouraging and indicate a need for further development and application of the methods employed. In the case of the Difference-in-Difference analysis, the use of hierarchical clustering based on county and neighbors' attributes was demonstrated as a method for accounting for spatial nature of data in the selection of control and treatment groups to maximize confidence that the assumption of constant behavior is met. Space is accounted for in this process both by acknowledging the importance of neighbors' attributes in determining the economic climate of a county as well as an acknowledgment that sometimes, especially in the case of urban centers, near counties are not necessarily the most similar, but rather that the most similar counties may be widely distributed. Future work could

include past economic behavior in addition to demographic characteristics in the clustering process. This work would also benefit from being repeated for various treatment periods in different economic conditions to test for changes in overall patterns as economic climate changes. The period of 2000 - 20005 used in this analysis was chosen to avoid the post-2008 recession and to take advantage of the demographic data available from the 2000 census. It would be particularly interesting, though, now that this work has been accomplished, to repeat the analysis for more recent years precisely to see how results vary due to the economic recession.

The approach of using GISc to challenge hegemony in support of social/environmental justice and political ecology/economy research, has been useful in guiding this analysis and in finding purpose to continue the work. I will continue to explore the approach and hope to work with others to develop a more robust framework for its application. In particular, I plan to explore, discuss, and debate the epistemological shift necessary to move from an empirical nomothetic GISc analysis to analyses that challenge accepted notions and advocate for building more complex understandings of situations and processes previously assumed to be globally predictable and simple. More broadly, I plan to continue to advocate for the development of counter analytic methods for methods and frameworks for pursuing critical research using GISc methods and for developing GISc methods by reflecting on them through the lens of critical theory. This work will continue through my own work, through conversations with colleagues, and through education of future geographers.

Bibliography

- Angrist, J.D. and J.S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. Princeton University Press.
- Bertrand, M., E. Duflo, and S. Mullainathan (2004). "How Much Should We Trust Differences-In-Differences Estimates?" In: *The Quarterly Journal of Economics* 119.1, pp. 249–275.
- Bullard, R.D. (2005). *The quest for environmental justice: human rights and the politics of pollution*. Sierra Club Books.
- Bullard, R.D. et al. (2007). *Toxic wastes and race at twenty: 1987-2007*. Tech. rep. United Church of Christ Justice and Witness Ministries.
- Bunge, William (1988). *Nuclear war atlas*. Blackwell.
- Conley, J.F. (2011). "Estimation of exposure to toxic releases using spatial interaction modeling". In: *International Journal of Health Geographics* 10.20.
- Culcasi, Karen (2012). "Mapping the Middle East from Within:(Counter-) Cartographies of an Imperialist Construction". In: *Antipode* 44.4, pp. 1099–1118.
- Cutter, S.L. (1995). "Race, class and environmental justice". In: *Progress in Human Geography* 19.1, pp. 111–122.
- Dolinoy, D.C. and M.L. Miranda (2004). "GIS modeling of air toxics releases from TRI-reporting and non-TRI-reporting facilities: impacts for environmental justice". In: *Environmental health perspectives* 112.17, p. 1717.
- EPA (2012a). *Basics of TRI Reporting*. URL: <http://www.epa.gov/tri/triprogram/bussinesscycle/index.html> (visited on 03/15/2012).

- EPA (2012b). *EPA FRS Facilities State Combined CSV Files Download*. URL: http://www.epa.gov/enviro/html/frs_demo/geospatial_data/geo_data_state_combined.html (visited on 03/15/2012).
- (2012c). *Is My Facility's Six-Digit NAICS Code a TRI-Covered Industry?* URL: <http://www.epa.gov/tri/lawsandregs/naic/ncodes.htm> (visited on 03/15/2012).
- Frazier, Reid R. (2012). *'Cracker' plant will bring jobs, but what about the air?* Public Source. URL: <http://publicsource.org/investigations/cracker-plant-will-bring-jobs-what-about-air> (visited on 09/10/2012).
- Goodchild, Michael F (1992). "Geographical information science". In: *International Journal of Geographical Information Systems* 6.1, pp. 31–45.
- Harris, Trevor and Dan Weiner (1996). *GIS and Society: The Social Implications of how People, Space and Environment are Represented in GIS: Scientific Report for the Initiative 19 Specialist Meeting, March 2-5, 1996, Koinonia Retreat Center, South Haven, Minnesota*. National Center for Geographic Information and Analysis.
- Harris, Trevor and Daniel Weiner (1998). "Empowerment, marginalization, and "community-integrated" GIS". In: *Cartography and Geographic Information Science* 25.2, pp. 67–76.
- Heynen, Nik, Harold A Perkins, and Parama Roy (2006). "The political ecology of uneven urban green space the impact of political economy on race and ethnicity in producing environmental inequality in Milwaukee". In: *Urban Affairs Review* 42.1, pp. 3–25.
- IRS (2009). *2009 Statistics of Income County Income Tables - Users Guide and Record Layout*. URL: <http://www.irs.gov/taxstats/article/0,,id=215866,00.html> (visited on 08/03/2012).
- (2012). *SOI Tax Stats - County Income Data*. URL: <http://www.irs.gov/taxstats/article/0,,id=215866,00.html> (visited on 08/03/2012).
- Jerrett, Michael et al. (2001). "A GIS-environmental justice analysis of particulate air pollution in Hamilton, Canada". In: *Environment and Planning A* 33.6, pp. 955–974.
- Johnston, Ronald John et al. (2000). *The dictionary of human geography, 4th edition*. Blackwell Reference Oxford.

- Kwan, Mei-Po (2002). "Feminist visualization: Re-envisioning GIS as a method in feminist geographic research". In: *Annals of the association of American Geographers* 92.4, pp. 645–661.
- Maantay, J. (2007). "Asthma and air pollution in the Bronx: methodological and data considerations in using GIS for environmental justice and health research". In: *Health & place* 13.1, pp. 32–56.
- Mackinnon, Danny and Andy Cumbers (2007). *An introduction to economic geography: globalization, uneven development and place*. Pearson Education.
- Mennis, J. (2002). "Using geographic information systems to create and analyze statistical surfaces of population and risk for environmental justice analysis". In: *Social Science Quarterly* 83.1, pp. 281–297.
- Njeru, Jeremia (2006). "The urban political ecology of plastic bag waste problem in Nairobi, Kenya". In: *Geoforum* 37.6, pp. 1046–1058.
- Novak, Melanie (1996). *With incinerator plan, Curtis Bay feels dumped-on again*. The Morning Call. URL: http://articles.mcall.com/1996-11-24/news/3117161_1_bethlehem-steel-planning-board-chemicals (visited on 09/10/2012).
- Panda Energy, Inc. (2007). *Panda Energy to Build 1,000 MW Power Plant in Temple, Texas*. PR Newswire. URL: <http://www.prnewswire.com/news-releases/panda-energy-to-build-1000-mw-power-plant-in-temple-texas-58559552.html> (visited on 09/10/2012).
- The Principles of Environmental Justice (EJ)*. (1991). People of Color Environmental Leadership Summit. Washington DC. URL: www.ejnet.org/ej/.
- Peluso, Nancy Lee (1995). "Whose woods are these? counter-mapping forest territories in Kalimantan, Indonesia". In: *Antipode* 27.4, pp. 383–406.
- R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. URL: <http://www.R-project.org/>.
- Rey, Sergio J and Luc Anselin (2007). "PySAL: A Python Library of Spatial Analytical Methods". In: *The Review of Regional Studies* 37.1, pp. 5–27.
- Schuurman, Nadine (2000). "Trouble in the heartland: GIS and its critics in the 1990s". In: *Progress in Human Geography* 24.4, pp. 569–590.

- Shafrin, J. (2006). *Healthcare Economist, Difference in Difference Estimation*. URL: <http://healthcare-economist.com/2006/02/11/difference-in-difference-estimation/> (visited on 09/13/2012).
- Shauk, Zain (2012). *New Dow ethylene plant would employ thousands*. Fuel Fix. URL: <http://fuelfix.com/blog/2012/04/19/dow-to-announce-new-houston-area-plant/> (visited on 09/10/2012).
- Shen, Fern (2012). *With incinerator plan, Curtis Bay feels dumped-on again*. Baltimore Brew. URL: <http://www.baltimorebrew.com/2012/08/31/with-incinerator-plan-curtis-bay-feels-dumped-on-again/> (visited on 09/10/2012).
- Sheppard, Eric (2005). “Knowledge production through critical GIS: genealogy and prospects”. In: *Cartographica: The International Journal for Geographic Information and Geovisualization* 40.4, pp. 5–21.
- Shipp, Brett (2012). *Neighbors vow to fight chemical plant at Super Collider site*. WFAA. URL: <http://www.wfaa.com/news/business/Super-Collider-site-purchased-by-chemical-plant-owner-138444649.html> (visited on 09/10/2012).
- Shrader-Frechette, K.S. (2002). *Environmental justice: creating equality, reclaiming democracy*. Oxford University Press, USA.
- Templet, Paul H and Stephen Farber (1994). “The complementarity between environmental and economic risk: an empirical analysis”. In: *Ecological Economics* 9.2, pp. 153–165.
- UCC-CRJ, Commission for Racial Justice (1987). *Toxic wastes and race in the United States: a national report on the racial and socio-economic characteristics of communities with hazardous waste sites*. Tech. rep. United Church of Christ.
- U.S. Census (2001a). *Census 2000 summary file 1 technical documentation*.
- (2001b). *Census 2000 summary file 3 technical documentation*.
- (2012). *2010 Census (P.L. 94-171) TIGER/Line Shapefiles Technical Documentation*. URL: <http://www.census.gov/geo/www/tiger/tgrshp2010/tgrshp2010.html> (visited on 03/15/2012).
- Wheeler, Timothy B. (2012). *Trash-burning power plant gets another airing*. The Baltimore Sun. URL: http://articles.baltimoresun.com/2012-08-29/features/bs-gr-psc-trash-power-plant-20120829_1_power-plant-curtis-bay-recreation-center-air-pollution (visited on 09/10/2012).

Woods, Brad R. and Jason S. Gordon (2011). "Mountaintop Removal and Job Creation: Exploring the Relationship Using Spatial Regression". In: *Annals of the Association of American Geographers* 101.4, pp. 806–815. DOI: 10.1080/00045608.2011.567947.

Appendix

All code is provided as-is and as was used in the above analysis. All code is being released under the BSD 3-Clause open source license. Code will be further modified in future work and will be made available at <http://geo.wvu.edu/~cdavis>. All Python code was developed and executed within an IPython Notebook Environment, but has been exported to a plain python script for inclusion in this report. R code was developed and executed within RStudio. Code is presented in the order of intended execution.

3.4 Python Code

```
1 #cd-CountyWeightsMatrix-3.0.py
2
3 import shapefile
4 import pysal as ps
5 import numpy as np
6 import pprint
7 import csv
8 from numpy import genfromtxt
9 import pylab
10 from copy import deepcopy
11
12 # Create County Weights Matrix (Queen Contiguity)
```

```

13
14 boundariesFilePath = "F:\Dropbox\School\Thesis_Sp2013\Work\IPython\
    co99-d00-shp\co99-d00-Dissolved.shp" ##"co99-d00-Dissolved.shp" is a 2000
    Census county outline shape file dissolved by GEO_id (after adding correct
    (NAD83) Datum information)
15 boundariesIdVariable = "GeoID"
16
17 [boundariesPath, boundariesFile] = boundariesFilePath.rsplit("\\", 1) ##get
    folder location and filename
18 boundariesPath = boundariesPath + "\\ ## add \\ back onto path
19 boundariesFileRoot = boundariesFile[:-4] ## get filename without extension
20
21 countyWeightsQ = ps.queen_from_shapefile(boundariesFilePath,
    boundariesIdVariable)
22
23 countyWeightsQ.n
24
25 countyWeightsQ.histogram
26
27 countyWeightsQArray = np.array(countyWeightsQ.histogram)
28 pylab.bar(countyWeightsQArray[:,0], countyWeightsQArray[:,1])
29 pylab.xlabel('Number_of_Neighbors')
30 pylab.ylabel('Counties')
31 pylab.title('Neighbor_Distribution')
32 pylab.grid(True)
33 pylab.savefig('Neighbor_Distribution')
34 pylab.show()

```

```

35
36 galName = boundariesFileRoot + ".gal"
37 countyWeightsQ_gal = ps.open(galName, 'w', 'gal') ##export to .gal for backup
38 countyWeightsQ_gal.write(countyWeightsQ)
39
40 dbfName = boundariesFileRoot + ".dbf" ## load dbf from shapefile
41 dbfPath = boundariesPath + dbfName
42 countiesDBF = shapefile.Reader(dbfPath)
43 countiesDBFRecords = countiesDBF.records()
44 countiesDBFFields = countiesDBF.fields
45 #countiesDBFRecords[0]
46
47 countiesDBFDict = {} ##read dbf into dictionary
48 for i in countiesDBFRecords:
49     countiesDBFDict[i[1]] = {countiesDBFFields[j+1][0]:i[j] for j in arange(
        len(countiesDBFFields[1:]))}
50
51 with open('F:\Dropbox\School\Census2000\DEC_00_SF1-DP1-with-ann.csv', mode='r'
    ) as infile: ##from Census
52     reader = csv.reader(infile)
53     headers = reader.next()
54     fields = range(len(headers))
55     census2000_SF1_DP1 = {row[0]:{headers[i]:row[i] for i in fields} for row
        in reader}
56
57 with open('F:\Dropbox\School\Census2000\DEC_00_SF3-DP3-with-ann.csv', mode='r'
    ) as infile: ##from Census

```

```

58     reader = csv.reader(infile)
59     headers = reader.next()
60     fields = range(len(headers))
61     census2000_SF3_DP3 = {row[0]:{headers[i]:row[i] for i in fields} for row
        in reader}
62
63 # For all counties and find values for their neighbors
64
65     counties2000 = {}
66     counties2000.clear()
67     for county in countyWeightsQ.id_order:
68
69 #     if countiesDBFDict[county]['GeoID'][9:11] != "72" and countiesDBFDict[
        county]['GeoID'][9:11] != "15": ## do not include Puerto Rico or Hawai'i
        in analysis
70
71     if countiesDBFDict[county]['GeoID'][9:11] != "72": ## do not include
        Puerto Rico in analysis
72
73     if countyWeightsQ.neighbors[county]: ## only include counties with
        neighbors in analysis
74
75         counties2000[county]={}
76         ##Data about county
77         counties2000[county]['neighbors'] = len(countyWeightsQ.neighbors[
            county]) ### number of neighbors county has #removed from
            analysis in v1.5

```

```

78     ##Area
79     counties2000[county]['totalPop'] = int(census2000_SF1_DP1[county][
        'HC01_VC01']) ## total population of county
80     counties2000[county]['percentWhite'] = float(census2000_SF1_DP1[
        county]['HC02_VC62']) ## percent white, not hispanic or latino
81     counties2000[county]['percentEmployed'] = float(census2000_SF3_DP3
        [county]['HC02_VC03']) ## percent employed (over 16 years old)
82     counties2000[county]['medHhIncome1999'] = int(census2000_SF3_DP3[
        county]['HC01_VC64']) ## 1999 Median Household income
83
84     ##Data about county's neighbors
85     neiPopList = []
86     neiPerWnHList = []
87     neiPerEmpList = []
88     neiMHhIncList = []
89     neiNeighborsList = [] #removed v1.5
90
91     for neighbor in countyWeightsQ.neighbors[county]: ## populate list
        of neighbors' values
92         neiPopList.append(int(census2000_SF1_DP1[neighbor]['HC01_VC01',
            ])) ## total population
93         neiPerWnHList.append(float(census2000_SF1_DP1[neighbor]['
            HC02_VC62'])) ## percent white, not hispanic or latino
94         neiPerEmpList.append(float(census2000_SF3_DP3[neighbor]['
            HC02_VC03'])) ## percent employed (over 16 years old)
95         neiMHhIncList.append(int(census2000_SF3_DP3[neighbor]['
            HC01_VC64'])) ## 1999 Median Household income

```

```

96         neiNeighborsList.append(int(len(countyWeightsQ.neighbors[
           neighbor]))) ## number of neighbors #removed v1.5
97
98     counties2000[county]['neiPopSum'] = np.sum(neiPopList) ## total
           population of neighboring counties
99     counties2000[county]['neiPerWhite'] = np.mean(neiPerWnHList) ##
           percent white, not hispanic or latino, in neighboring counties
100    counties2000[county]['neiPerEmp'] = np.mean(neiPerEmpList) ##
           percent employed (over 16 years old), in neighboring counties
101    counties2000[county]['neiMHhInc'] = np.mean(neiMHhIncList) ## 1999
           Median Household income, in neighboring counties
102    counties2000[county]['neiNeiSum'] = np.sum(neiNeighborsList) ##
           total number of neighboring counties' neighbors (second order
           neighbors) (same conception of neighboring as original) #
           removed v1.5
103
104    len(counties2000)
105
106    with open('F:\Dropbox\School\Thesis_Sp2013\Work\stand\countiesWithNeighborData
           .csv', mode='w') as outfile: ##export dictionary to csv for R
107        writer = csv.writer(outfile, delimiter=',')
108
109        counties = counties2000.keys() ##list all of the counties in the
           dectonary
110        keys = counties2000[counties2000.keys()[0]].keys() ## list all of the keys
           for the first county in the dictionary (should be the same for all
           counties)

```



```

111
112     keys = [ 'geoID' ] + keys
113
114     writer.writerow(keys)
115     for county in counties:
116         countyValues = [county] + counties2000[county].values()
117         writer.writerow(countyValues) ## print geoID and all values for each
            county on a new line of csv
118
119 # Read in TRI Sites and add entries to counties dictionary for number of new
    sites for each year
120
121 triFilePath = "F:\\\\Dropbox\\\\School\\\\Thesis_Sp2013\\\\Work\\\\shapefiles\\\\
    TRIWithDatesAndCounties.shp" ## "TRIWithDatesAndCounties.shp" was created
    in arcGIS by spatially joining the TRI sites to "co99_d00_Dissolved.shp"
122
123 [triPath, triFile] = triFilePath.rsplit("\\", 1) ##get folder location and
    filename
124 triPath = triPath + "\\\" ## add \\ back onto path
125 triFileRoot = triFile[:-4] ## get filename without extension
126
127 triDbfName = triFileRoot + ".dbf" ## load dbf from shapefile
128 triDbfPath = triPath + triDbfName
129 triDBF = shapefile.Reader(triDbfPath)
130 triDBFRecords = triDBF.records()
131 triDBFFields = triDBF.fields
132 #countiesDBFRecords[0]

```

```

133 triDBFDict = {} ##read dbf into dictionary
134 for i in triDBFRecords:
135     triDBFDict[i[1]] = {triDBFFields[j+1][0]:i[j] for j in arange(len(
        triDBFFields[1:]))}
136
137 countiesWTri = deepcopy(counties2000)
138
139 for entry in countiesWTri:
140     for year in arange(1987, 2011, 1):
141         countiesWTri[entry]['tri' + str(year)] = 0 ## add a 0 for each year
142
143 triSitesNotIncluded = [] ## sites not included for any reaseon (throws error
        in next loop)
144 triSitesNIPuertoRico = [] ## sites not included because they're in Puerto Rico
145 triSitesNINoYear = [] ## sites not included because they don't have a start
        year
146 triSitesNINoNeighbors = [] ## sites not included because they're in counties
        without neighbors
147 triSitesNINoCounty = [] ## sites excluded because they do not fall within a
        county boundary polygon (could be due to error or uncertainty in
        shapefiles)
148 triSitesNIOther = []
149
150 for site in triDBFDict:
151     GeoID = str(triDBFDict[site]['GeoID'])
152     triYear = str(triDBFDict[site]['STARTDATE'][0])
153     try:

```

```

154         countiesWTri[GeoID]['tri' + str(triYear)] = countiesWTri[GeoID]['tri'
           + str(triYear)] + 1
155     except:
156         triSitesNotIncluded.append(site) ## all sites not included
157         if countiesDBFDict[county]['GeoID'][9:11] != "72":
           triSitesNIPuertoRico.append(site)
158         elif triYear == '_': triSitesNINoYear.append(site)
159         elif countyWeightsQ.neighbors['GeoID'] == []: triSitesNINoNeighbors.
           append(site)
160         elif GeoID == '_____':
           triSitesNINoCounty.append(site)
161         else: triSitesNIOther.append(site)
162
163 with open('F:\Dropbox\School\Thesis_Sp2013\Work\stand\countiesWithNewTriByYear
           .csv', mode='w') as outfile: ##export dictionary to csv for R
164     writer = csv.writer(outfile, delimiter=',')
165
166     triCounties = countiesWTri.keys() ##list all of the counties in the
           dectinary
167     triKeys = countiesWTri[countiesWTri.keys()[0]].keys() ## list all of the
           keys for the first county in the dictionary (should be the same for
           all counties)
168
169     triKeys = ['geoID'] + triKeys
170
171     writer.writerow(triKeys)
172     for triCounty in triCounties:

```

```

173     countyValuesTri = [triCounty] + countiesWTri[triCounty].values()
174     writer.writerow(countyValuesTri) ## print geoID and all values for
        each county on a new line of csv

```

3.5 R Code 1

```

1  #countyClustersScript_0.3.2.R
2
3  setwd("F:\\Dropbox\\School\\Thesis_Sp2013\\Work\\")
4
5  ##load csv data from python
6  counties <- read.csv("F:\\Dropbox\\School\\Thesis_Sp2013\\Work\\
    countiesWithNeighborData.csv") ## if use geoID as row names, row.names =
        0)
7
8  ##
9  counties.use.unscaled = counties[, -c(1,1)] ##remove geoID column
10 counties.use = scale(counties.use.unscaled) ##scales all variables for use in
        hclust
11
12 counties.dist = dist(counties.use, method = "euclidean") ##compute distance
        matrix
13
14 counties.hclust = hclust(counties.dist, "ward") ##hierarchical clustering of
        scaled county demographic data
15
16 plot(counties.hclust, main='Default_from_hclust', labels=FALSE)
17

```

```

18 ## run for various numbers of clusters
19
20 county.vars = length(counties)
21
22 counts = sapply(2:county.vars, function(ncl) table(cutree(counties.hclust, ncl)))
      ## show number of counties per cluster for 2 through ten clusters
23 names(counts) = 2:county.vars
24 counts
25
26 counties.list <- list()
27 counties.C <- list()
28 for (h in 1:10){
29
30   counties.C[[h]] = cutree(counties.hclust, h)
31
32   counties$geoID[counties.C[[h]] == 1]
33
34   counties.list[[h]] <- sapply(unique(counties.C[[h]]), function(g) counties$
      geoID[counties.C[[h]] == g])
35 }
36
37 #export clusters to csv
38
39 countiesWithClusters <- as.data.frame(counties.C[[1]])
40 countiesWithClusters <- cbind(countiesWithClusters, as.data.frame(counties.C
      [[2]]))

```

```

41 countiesWithClusters <- cbind(countiesWithClusters, as.data.frame(counties.C
    [[3]]))
42 countiesWithClusters <- cbind(countiesWithClusters, as.data.frame(counties.C
    [[4]]))
43 countiesWithClusters <- cbind(countiesWithClusters, as.data.frame(counties.C
    [[5]]))
44 countiesWithClusters <- cbind(countiesWithClusters, as.data.frame(counties.C
    [[6]]))
45 countiesWithClusters <- cbind(countiesWithClusters, as.data.frame(counties.C
    [[7]]))
46 countiesWithClusters <- cbind(countiesWithClusters, as.data.frame(counties.C
    [[8]]))
47 countiesWithClusters <- cbind(countiesWithClusters, as.data.frame(counties.C
    [[9]]))
48 countiesWithClusters <- cbind(countiesWithClusters, as.data.frame(counties.C
    [[10]]))
49
50 countiesWithClusters <- cbind(counties, countiesWithClusters)
51
52 write.csv(countiesWithClusters, file = "countiesWithClusters1_10.csv")
53
54 ## get cluster descriptions and output to text file (for examination of nature
    of clusters)
55 for (nC in 1:10){
56   u <- paste("*****_", nC, "Clusters_*****")
57   capture.output(u, file = "clusterSummaries.txt", append = TRUE)
58   for (cluster in 1:nC){

```

```

59     t <- paste("_____Cluster_Number_", cluster, "_of_", nC, "
               _____")
60     capture.output(t, file = "clusterSummaries.txt", append = TRUE)
61     s <- summary(subset(countiesWithClusters, counties.C[[nC]]==cluster))
62     capture.output(s, file = "clusterSummaries.txt", append = TRUE)
63   }
64 }
65
66 ## make descriptive plot of cluster stats  #nc = 4
67
68 nC = 6
69
70 plot(counties.hclust, main='County_Clusters_by_Demographics', labels=FALSE)
71 rect.hclust(counties.hclust, k = nC)
72
73
74 attributes = labels(countiesWithClusters)[2][[1]][2:county.vars] ## get the
               labels of variables used for clustering
75 limit = data.frame(row.names = c("min", "max"))
76 #summary(subset(countiesWithClusters, counties.C[[nC]]==1)) ## show summary
               for first cluster
77
78 #set ranges for all attributes  ##prep for plots
79 limit[,1] = c(NA, NA)
80 for (q in 2:county.vars){
81   limit[,q] = range(countiesWithClusters[q])
82 }

```

```

83 ##shrink upper limit for population to three times the median:
84
85 limit[2,7]= 3*median(countiesWithClusters$totalPop)
86
87 ##box plots collected by cluster
88
89 for (cluster in 1:nC){
90   fileNameBxByCluster = paste("box_Clust",cluster,"_", nC, ".png", sep = "")
91   png(filename = fileNameBxByCluster, width = 2000, height = 1600, units = "px
      ",pointsize = 24,)
92   par(mfrow=c(2,5))
93   for (q in 2:county.vars){
94     boxplot(subset(countiesWithClusters, counties.C[[nC]]==cluster)[q], main =
      labels(countiesWithClusters)[2][[1]][q], ylim = limit[,q])
95   }
96   title(paste("Cluster_", cluster, "of_", nC), outer = TRUE, line = -1.1 )
97   dev.off()
98 }
99
100 ##box plots collected by attribute
101
102 for (q in 2:county.vars){
103   fileNameClusterbyAtt = paste("box_Attr",q, "_", nC, ".png", sep = "")
104   png(filename = fileNameClusterbyAtt, width = 1000, height = 800, units = "px
      ",pointsize = 24,)
105   par(mfrow=c(1,nC))
106

```



```

107   for (cluster in 1:nC){
108       boxplot(subset(countiesWithClusters, main = cluster, counties.C[[nC]]==
                    cluster)[q],ylim = limit[,q])
109   }
110   title(labels(countiesWithClusters)[2][[1]][q], outer = TRUE, line = -1.1 )
111   dev.off()
112 }

```

3.6 R Code 2

```

1  #differenceInDifferences_6.7_stand.R
2  ## ## ## ## Cdavis script to run difference-in-differences analysis on some
   data ## ## ## ## ##
3  ##### run countyClustersScript_x.x.x.R first in same R session
   #####
4
5  nClusters = 6  ## number of clusters
6
7  ### add data for number of new tri sites by year
8  countiesWithTri<- read.csv( 'F:\\Dropbox\\School\\Thesis_Sp2013\\Work\\
   countiesWithNewTriByYear.csv ' )
9
10 ### read in all irs data for all counties for all years (not just wages, not
   collected by county)
11 countiesWithAllIrs <- read.csv( 'F:\\Dropbox\\School\\Thesis_Sp2013\\Work\\
   IncomeDataAllYears_cleaned.csv ' )
12

```

```

13 ## build table of IRS Data that looks like TRI table (1 row per geoID with all
    years)

```

```

14

```

```

15 irs1989 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 1989)

```

```

16 irs1990 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 1990)

```

```

17 irs1991 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 1991)

```

```

18 irs1992 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 1992)

```

```

19 irs1993 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 1993)

```

```

20 irs1994 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 1994)

```

```

21 irs1995 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 1995)

```

```

22 irs1996 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 1996)

```

```

23 irs1997 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 1997)

```

```

24 irs1998 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 1998)

```

```

25 irs1999 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 1999)

```

```

26 irs2000 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 2000)

```

```

27 irs2001 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 2001)

```

```

28 irs2002 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 2002)

```

```

29 irs2003 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 2003)

```

```

30 irs2004 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 2004)

```

```

31 irs2005 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 2005)

```

```

32 irs2006 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 2006)

```

```

33 irs2007 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 2007)

```

```

34 irs2008 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 2008)

```

```

35 irs2009 <- subset(countiesWithAllIrs , countiesWithAllIrs$Year == 2009)

```

```

36

```

```

37 countiesWithIrs <- data.frame(GeoID = unique(countiesWithAllIrs$GeoID) ,

```

```

38           wages1989 = NA, wages1990 = NA, wages1991 = NA,

```

```

39      wages1992 = NA, wages1993 = NA, wages1994 = NA,
40      wages1995 = NA, wages1996 = NA, wages1997 = NA,
41      wages1998 = NA, wages1999 = NA, wages2000 = NA,
42      wages2001 = NA, wages2002 = NA, wages2003 = NA,
43      wages2004 = NA, wages2005 = NA, wages2006 = NA,
44      wages2007 = NA, wages2008 = NA, wages2009 = NA,
45      standWages1989 = NA, standWages1990 = NA, standWages1991
      = NA,
46      standWages1992 = NA, standWages1993 = NA, standWages1994
      = NA,
47      standWages1995 = NA, standWages1996 = NA, standWages1997
      = NA,
48      standWages1998 = NA, standWages1999 = NA, standWages2000
      = NA,
49      standWages2001 = NA, standWages2002 = NA, standWages2003
      = NA,
50      standWages2004 = NA, standWages2005 = NA, standWages2006
      = NA,
51      standWages2007 = NA, standWages2008 = NA, standWages2009
      = NA
52    )
53
54    for(county in as.single(1:length(countiesWithIrs$GeoID))){ ###If there is a
      value for that year for that county, record it. (This is wildly
      inefficient, but it's the best I've come up with so far.)
55    if(length(as.numeric(subset(irs1989, irs1989$GeoID == countiesWithIrs$GeoID[
      county]))$WagesSalaries))>0){

```

```

56     countiesWithIrs$wages1989[county] <- as.numeric(subset(irs1989, irs1989$
      GeoID == countiesWithIrs$GeoID[county])$WagesSalaries)})
57   if(length(as.numeric(subset(irs1990, irs1990$GeoID == countiesWithIrs$GeoID[
      county])$WagesSalaries))>0){
58     countiesWithIrs$wages1990[county] <- as.numeric(subset(irs1990, irs1990$
      GeoID == countiesWithIrs$GeoID[county])$WagesSalaries)})
59   if(length(as.numeric(subset(irs1991, irs1991$GeoID == countiesWithIrs$GeoID[
      county])$WagesSalaries))>0){
60     countiesWithIrs$wages1991[county] <- as.numeric(subset(irs1991, irs1991$
      GeoID == countiesWithIrs$GeoID[county])$WagesSalaries)})
61   if(length(as.numeric(subset(irs1992, irs1992$GeoID == countiesWithIrs$GeoID[
      county])$WagesSalaries))>0){
62     countiesWithIrs$wages1992[county] <- as.numeric(subset(irs1992, irs1992$
      GeoID == countiesWithIrs$GeoID[county])$WagesSalaries)})
63   if(length(as.numeric(subset(irs1993, irs1993$GeoID == countiesWithIrs$GeoID[
      county])$WagesSalaries))>0){
64     countiesWithIrs$wages1993[county] <- as.numeric(subset(irs1993, irs1993$
      GeoID == countiesWithIrs$GeoID[county])$WagesSalaries)})
65   if(length(as.numeric(subset(irs1994, irs1994$GeoID == countiesWithIrs$GeoID[
      county])$WagesSalaries))>0){
66     countiesWithIrs$wages1994[county] <- as.numeric(subset(irs1994, irs1994$
      GeoID == countiesWithIrs$GeoID[county])$WagesSalaries)})
67   if(length(as.numeric(subset(irs1995, irs1995$GeoID == countiesWithIrs$GeoID[
      county])$WagesSalaries))>0){
68     countiesWithIrs$wages1995[county] <- as.numeric(subset(irs1995, irs1995$
      GeoID == countiesWithIrs$GeoID[county])$WagesSalaries)})

```

```

69   if (length(as.numeric(subset(irs1996 , irs1996$GeoID == countiesWithIrs$GeoID[
      county]))$WagesSalaries))>0){
70     countiesWithIrs$wages1996[county] <- as.numeric(subset(irs1996 , irs1996$
      GeoID == countiesWithIrs$GeoID[county]))$WagesSalaries)}
71   if (length(as.numeric(subset(irs1997 , irs1997$GeoID == countiesWithIrs$GeoID[
      county]))$WagesSalaries))>0){
72     countiesWithIrs$wages1997[county] <- as.numeric(subset(irs1997 , irs1997$
      GeoID == countiesWithIrs$GeoID[county]))$WagesSalaries)}
73   if (length(as.numeric(subset(irs1998 , irs1998$GeoID == countiesWithIrs$GeoID[
      county]))$WagesSalaries))>0){
74     countiesWithIrs$wages1998[county] <- as.numeric(subset(irs1998 , irs1998$
      GeoID == countiesWithIrs$GeoID[county]))$WagesSalaries)}
75   if (length(as.numeric(subset(irs1999 , irs1999$GeoID == countiesWithIrs$GeoID[
      county]))$WagesSalaries))>0){
76     countiesWithIrs$wages1999[county] <- as.numeric(subset(irs1999 , irs1999$
      GeoID == countiesWithIrs$GeoID[county]))$WagesSalaries)}
77   if (length(as.numeric(subset(irs2000 , irs2000$GeoID == countiesWithIrs$GeoID[
      county]))$WagesSalaries))>0){
78     countiesWithIrs$wages2000[county] <- as.numeric(subset(irs2000 , irs2000$
      GeoID == countiesWithIrs$GeoID[county]))$WagesSalaries)}
79   if (length(as.numeric(subset(irs2001 , irs2001$GeoID == countiesWithIrs$GeoID[
      county]))$WagesSalaries))>0){
80     countiesWithIrs$wages2001[county] <- as.numeric(subset(irs2001 , irs2001$
      GeoID == countiesWithIrs$GeoID[county]))$WagesSalaries)}
81   if (length(as.numeric(subset(irs2002 , irs2002$GeoID == countiesWithIrs$GeoID[
      county]))$WagesSalaries))>0){

```

```

82     countiesWithIrs$wages2002[county] <- as.numeric(subset(irs2002, irs2002$
      GeoID == countiesWithIrs$GeoID[county])$WagesSalaries)})
83   if(length(as.numeric(subset(irs2003, irs2003$GeoID == countiesWithIrs$GeoID[
      county])$WagesSalaries))>0){
84     countiesWithIrs$wages2003[county] <- as.numeric(subset(irs2003, irs2003$
      GeoID == countiesWithIrs$GeoID[county])$WagesSalaries)})
85   if(length(as.numeric(subset(irs2004, irs2004$GeoID == countiesWithIrs$GeoID[
      county])$WagesSalaries))>0){
86     countiesWithIrs$wages2004[county] <- as.numeric(subset(irs2004, irs2004$
      GeoID == countiesWithIrs$GeoID[county])$WagesSalaries)})
87   if(length(as.numeric(subset(irs2005, irs2005$GeoID == countiesWithIrs$GeoID[
      county])$WagesSalaries))>0){
88     countiesWithIrs$wages2005[county] <- as.numeric(subset(irs2005, irs2005$
      GeoID == countiesWithIrs$GeoID[county])$WagesSalaries)})
89   if(length(as.numeric(subset(irs2006, irs2006$GeoID == countiesWithIrs$GeoID[
      county])$WagesSalaries))>0){
90     countiesWithIrs$wages2006[county] <- as.numeric(subset(irs2006, irs2006$
      GeoID == countiesWithIrs$GeoID[county])$WagesSalaries)})
91   if(length(as.numeric(subset(irs2007, irs2007$GeoID == countiesWithIrs$GeoID[
      county])$WagesSalaries))>0){
92     countiesWithIrs$wages2007[county] <- as.numeric(subset(irs2007, irs2007$
      GeoID == countiesWithIrs$GeoID[county])$WagesSalaries)})
93   if(length(as.numeric(subset(irs2008, irs2008$GeoID == countiesWithIrs$GeoID[
      county])$WagesSalaries))>0){
94     countiesWithIrs$wages2008[county] <- as.numeric(subset(irs2008, irs2008$
      GeoID == countiesWithIrs$GeoID[county])$WagesSalaries)})

```

```

95   if (length(as.numeric(subset(irs2009 , irs2009$GeoID == countiesWithIrs$GeoID[
      county]))$WagesSalaries))>0){
96     countiesWithIrs$wages2009[county] <- as.numeric(subset(irs2009 , irs2009$
      GeoID == countiesWithIrs$GeoID[county]))$WagesSalaries)}
97
98   ##### Use next code chunk for standardized data (wages/returns)
99
100  if (length(as.numeric(subset(irs1989 , irs1989$GeoID == countiesWithIrs$GeoID[
      county]))$Return))>0){
101    countiesWithIrs$standWages1989[county] <- countiesWithIrs$wages1989[county
      ] / as.numeric(subset(irs1989 , irs1989$GeoID == countiesWithIrs$GeoID[
      county]))$Return)}
102  if (length(as.numeric(subset(irs1990 , irs1990$GeoID == countiesWithIrs$GeoID[
      county]))$Return))>0){
103    countiesWithIrs$standWages1990[county] <- countiesWithIrs$wages1990[county
      ] / as.numeric(subset(irs1990 , irs1990$GeoID == countiesWithIrs$GeoID[
      county]))$Return)}
104  if (length(as.numeric(subset(irs1991 , irs1991$GeoID == countiesWithIrs$GeoID[
      county]))$Return))>0){
105    countiesWithIrs$standWages1991[county] <- countiesWithIrs$wages1991[county
      ] / as.numeric(subset(irs1991 , irs1991$GeoID == countiesWithIrs$GeoID[
      county]))$Return)}
106  if (length(as.numeric(subset(irs1992 , irs1992$GeoID == countiesWithIrs$GeoID[
      county]))$Return))>0){
107    countiesWithIrs$standWages1992[county] <- countiesWithIrs$wages1992[county
      ] / as.numeric(subset(irs1992 , irs1992$GeoID == countiesWithIrs$GeoID[
      county]))$Return)}

```

```

108   if (length(as.numeric(subset(irs1993, irs1993$GeoID == countiesWithIrs$GeoID[
      county]))$Return))>0){
109     countiesWithIrs$standWages1993[county] <- countiesWithIrs$wages1993[county
      ] / as.numeric(subset(irs1993, irs1993$GeoID == countiesWithIrs$GeoID[
      county]))$Return)}
110   if (length(as.numeric(subset(irs1994, irs1994$GeoID == countiesWithIrs$GeoID[
      county]))$Return))>0){
111     countiesWithIrs$standWages1994[county] <- countiesWithIrs$wages1994[county
      ] / as.numeric(subset(irs1994, irs1994$GeoID == countiesWithIrs$GeoID[
      county]))$Return)}
112   if (length(as.numeric(subset(irs1995, irs1995$GeoID == countiesWithIrs$GeoID[
      county]))$Return))>0){
113     countiesWithIrs$standWages1995[county] <- countiesWithIrs$wages1995[county
      ] / as.numeric(subset(irs1995, irs1995$GeoID == countiesWithIrs$GeoID[
      county]))$Return)}
114   if (length(as.numeric(subset(irs1996, irs1996$GeoID == countiesWithIrs$GeoID[
      county]))$Return))>0){
115     countiesWithIrs$standWages1996[county] <- countiesWithIrs$wages1996[county
      ] / as.numeric(subset(irs1996, irs1996$GeoID == countiesWithIrs$GeoID[
      county]))$Return)}
116   if (length(as.numeric(subset(irs1997, irs1997$GeoID == countiesWithIrs$GeoID[
      county]))$Return))>0){
117     countiesWithIrs$standWages1997[county] <- countiesWithIrs$wages1997[county
      ] / as.numeric(subset(irs1997, irs1997$GeoID == countiesWithIrs$GeoID[
      county]))$Return)}
118   if (length(as.numeric(subset(irs1998, irs1998$GeoID == countiesWithIrs$GeoID[
      county]))$Return))>0){

```



```

119     countiesWithIrs$standWages1998[county] <- countiesWithIrs$wages1998[county
      ] / as.numeric(subset(irs1998, irs1998$GeoID == countiesWithIrs$GeoID[
        county]))$Return)}
120   if(length(as.numeric(subset(irs1999, irs1999$GeoID == countiesWithIrs$GeoID[
        county]))$Return))>0){
121     countiesWithIrs$standWages1999[county] <- countiesWithIrs$wages1999[county
      ] / as.numeric(subset(irs1999, irs1999$GeoID == countiesWithIrs$GeoID[
        county]))$Return)}
122   if(length(as.numeric(subset(irs2000, irs2000$GeoID == countiesWithIrs$GeoID[
        county]))$Return))>0){
123     countiesWithIrs$standWages2000[county] <- countiesWithIrs$wages2000[county
      ] / as.numeric(subset(irs2000, irs2000$GeoID == countiesWithIrs$GeoID[
        county]))$Return)}
124   if(length(as.numeric(subset(irs2001, irs2001$GeoID == countiesWithIrs$GeoID[
        county]))$Return))>0){
125     countiesWithIrs$standWages2001[county] <- countiesWithIrs$wages2001[county
      ] / as.numeric(subset(irs2001, irs2001$GeoID == countiesWithIrs$GeoID[
        county]))$Return)}
126   if(length(as.numeric(subset(irs2002, irs2002$GeoID == countiesWithIrs$GeoID[
        county]))$Return))>0){
127     countiesWithIrs$standWages2002[county] <- countiesWithIrs$wages2002[county
      ] / as.numeric(subset(irs2002, irs2002$GeoID == countiesWithIrs$GeoID[
        county]))$Return)}
128   if(length(as.numeric(subset(irs2003, irs2003$GeoID == countiesWithIrs$GeoID[
        county]))$Return))>0){
129     countiesWithIrs$standWages2003[county] <- countiesWithIrs$wages2003[county
      ] / as.numeric(subset(irs2003, irs2003$GeoID == countiesWithIrs$GeoID[

```

```

        county]))$Return)}
130   if (length(as.numeric(subset(irs2004, irs2004$GeoID == countiesWithIrs$GeoID[
        county]))$Return))>0){
131     countiesWithIrs$standWages2004[county] <- countiesWithIrs$wages2004[county
        ] / as.numeric(subset(irs2004, irs2004$GeoID == countiesWithIrs$GeoID[
        county]))$Return)}
132   if (length(as.numeric(subset(irs2005, irs2005$GeoID == countiesWithIrs$GeoID[
        county]))$Return))>0){
133     countiesWithIrs$standWages2005[county] <- countiesWithIrs$wages2005[county
        ] / as.numeric(subset(irs2005, irs2005$GeoID == countiesWithIrs$GeoID[
        county]))$Return)}
134   if (length(as.numeric(subset(irs2006, irs2006$GeoID == countiesWithIrs$GeoID[
        county]))$Return))>0){
135     countiesWithIrs$standWages2006[county] <- countiesWithIrs$wages2006[county
        ] / as.numeric(subset(irs2006, irs2006$GeoID == countiesWithIrs$GeoID[
        county]))$Return)}
136   if (length(as.numeric(subset(irs2007, irs2007$GeoID == countiesWithIrs$GeoID[
        county]))$Return))>0){
137     countiesWithIrs$standWages2007[county] <- countiesWithIrs$wages2007[county
        ] / as.numeric(subset(irs2007, irs2007$GeoID == countiesWithIrs$GeoID[
        county]))$Return)}
138   if (length(as.numeric(subset(irs2008, irs2008$GeoID == countiesWithIrs$GeoID[
        county]))$Return))>0){
139     countiesWithIrs$standWages2008[county] <- countiesWithIrs$wages2008[county
        ] / as.numeric(subset(irs2008, irs2008$GeoID == countiesWithIrs$GeoID[
        county]))$Return)}

```

```

140   if (length(as.numeric(subset(irs2009 , irs2009$GeoID == countiesWithIrs$GeoID[
      county]))$Return))>0){
141     countiesWithIrs$standWages2009[county] <- countiesWithIrs$wages2009[county
      ] / as.numeric(subset(irs2009 , irs2009$GeoID == countiesWithIrs$GeoID[
      county]))$Return)}
142
143 }
144 head(countiesWithIrs)
145
146
147 ## populate cluster 1- nCluster
148 c.tri <- list()
149 c.irs <- list()
150 c.irs.tri <- list()
151 c.Treatment <- list()
152 c.Control <- list()
153
154 for (cluster in 1:nClusters){
155   c.tri[[cluster]] <- countiesWithTri[which(countiesWithTri$geoID %in%
      counties.list[[nClusters]][[cluster]]) ,] ## attach TRI Data
156   c.irs[[cluster]] <- countiesWithIrs[which(countiesWithIrs$GeoID %in%
      counties.list[[nClusters]][[cluster]]) ,] ## attach IRS Data
157
158   c.irs.tri[[cluster]] <- merge(c.tri[[cluster]] , c.irs[[cluster]] , by.x = "
      geoID" , by.y = "GeoID") ## merge all data for irs and tri
159
160 # ##make treatment and control groups

```

```

161 #   c.Treatment[[cluster]] <- subset(c.irs.tri[[cluster]], tri2000 != 0) #
      treatment group: subset with new tri facilities in 2000
162 #   c.Control[[cluster]] <- subset(c.irs.tri[[cluster]], tri2000 == 0) #
      control group: subset with new tri facilities in 2000
163 #   filterTag =""
164
165   ##filter for no new facilities 2001-2005
166   c.Treatment[[cluster]] <- subset(subset(c.irs.tri[[cluster]], tri2000 != 0)
      , tri2001 + tri2002 + tri2003 + tri2004 + tri2005 == 0) # treatment
      group: subset with new tri facilities in 2000
167   c.Control[[cluster]] <- subset(subset(c.irs.tri[[cluster]], tri2000 == 0),
      tri2001 + tri2002 + tri2003 + tri2004 + tri2005 == 0) # control group:
      subset with new tri facilities in 2000
168   filterTag = "Filtered"
169 }
170
171 #####
172 #countiesWithAllIrs[which(countiesWithAllIrs$GeoID=="0500000US01001" &
      countiesWithAllIrs$Year=="2000"),]$Return ## look up number of returns by
      GeoID and Year
173
174 ##### non regression based version:
175
176 ## get mean value of wages for all treatment and control groups
177
178 meanWages.Treatment <- list() ##foo$year[clusters1:10]
179 meanWages.Control <- list()

```

```

180
181
182 for (cluster in 1:nClusters){
183   ## taking 1989 to be year 1 through 2009 as year 21
184   ##treatment groups:
185   meanWages.Treatment$meanWages1989[[cluster]] <- mean(c.Treatment[[cluster]]$
        wages1989, na.rm=TRUE) ##removes NA values (counties without recorded
        wages; could be counties that were merged into other small pop counties
        or counties that did not exist that year)
186   meanWages.Treatment$meanWages1990[[cluster]] <- mean(c.Treatment[[cluster]]$
        wages1990, na.rm=TRUE)
187   meanWages.Treatment$meanWages1991[[cluster]] <- mean(c.Treatment[[cluster]]$
        wages1991, na.rm=TRUE)
188   meanWages.Treatment$meanWages1992[[cluster]] <- mean(c.Treatment[[cluster]]$
        wages1992, na.rm=TRUE)
189   meanWages.Treatment$meanWages1993[[cluster]] <- mean(c.Treatment[[cluster]]$
        wages1993, na.rm=TRUE)
190   meanWages.Treatment$meanWages1994[[cluster]] <- mean(c.Treatment[[cluster]]$
        wages1994, na.rm=TRUE)
191   meanWages.Treatment$meanWages1995[[cluster]] <- mean(c.Treatment[[cluster]]$
        wages1995, na.rm=TRUE)
192   meanWages.Treatment$meanWages1996[[cluster]] <- mean(c.Treatment[[cluster]]$
        wages1996, na.rm=TRUE)
193   meanWages.Treatment$meanWages1997[[cluster]] <- mean(c.Treatment[[cluster]]$
        wages1997, na.rm=TRUE)
194   meanWages.Treatment$meanWages1998[[cluster]] <- mean(c.Treatment[[cluster]]$
        wages1998, na.rm=TRUE)

```

```

195 meanWages.Treatment$meanWages1999[[cluster]] <- mean(c.Treatment[[cluster]]$
    wages1999, na.rm=TRUE)
196 meanWages.Treatment$meanWages2000[[cluster]] <- mean(c.Treatment[[cluster]]$
    wages2000, na.rm=TRUE)
197 meanWages.Treatment$meanWages2001[[cluster]] <- mean(c.Treatment[[cluster]]$
    wages2001, na.rm=TRUE)
198 meanWages.Treatment$meanWages2002[[cluster]] <- mean(c.Treatment[[cluster]]$
    wages2002, na.rm=TRUE)
199 meanWages.Treatment$meanWages2003[[cluster]] <- mean(c.Treatment[[cluster]]$
    wages2003, na.rm=TRUE)
200 meanWages.Treatment$meanWages2004[[cluster]] <- mean(c.Treatment[[cluster]]$
    wages2004, na.rm=TRUE)
201 meanWages.Treatment$meanWages2005[[cluster]] <- mean(c.Treatment[[cluster]]$
    wages2005, na.rm=TRUE)
202 meanWages.Treatment$meanWages2006[[cluster]] <- mean(c.Treatment[[cluster]]$
    wages2006, na.rm=TRUE)
203 meanWages.Treatment$meanWages2007[[cluster]] <- mean(c.Treatment[[cluster]]$
    wages2007, na.rm=TRUE)
204 meanWages.Treatment$meanWages2008[[cluster]] <- mean(c.Treatment[[cluster]]$
    wages2008, na.rm=TRUE)
205 meanWages.Treatment$meanWages2009[[cluster]] <- mean(c.Treatment[[cluster]]$
    wages2009, na.rm=TRUE)
206
207 ##control groups
208 meanWages.Control$meanWages1989[[cluster]] <- mean(c.Control[[cluster]]$
    wages1989, na.rm=TRUE)

```

```

209 meanWages.Control$meanWages1990[[cluster]] <- mean(c.Control[[cluster]]$
      wages1990, na.rm=TRUE)
210 meanWages.Control$meanWages1991[[cluster]] <- mean(c.Control[[cluster]]$
      wages1991, na.rm=TRUE)
211 meanWages.Control$meanWages1992[[cluster]] <- mean(c.Control[[cluster]]$
      wages1992, na.rm=TRUE)
212 meanWages.Control$meanWages1993[[cluster]] <- mean(c.Control[[cluster]]$
      wages1993, na.rm=TRUE)
213 meanWages.Control$meanWages1994[[cluster]] <- mean(c.Control[[cluster]]$
      wages1994, na.rm=TRUE)
214 meanWages.Control$meanWages1995[[cluster]] <- mean(c.Control[[cluster]]$
      wages1995, na.rm=TRUE)
215 meanWages.Control$meanWages1996[[cluster]] <- mean(c.Control[[cluster]]$
      wages1996, na.rm=TRUE)
216 meanWages.Control$meanWages1997[[cluster]] <- mean(c.Control[[cluster]]$
      wages1997, na.rm=TRUE)
217 meanWages.Control$meanWages1998[[cluster]] <- mean(c.Control[[cluster]]$
      wages1998, na.rm=TRUE)
218 meanWages.Control$meanWages1999[[cluster]] <- mean(c.Control[[cluster]]$
      wages1999, na.rm=TRUE)
219 meanWages.Control$meanWages2000[[cluster]] <- mean(c.Control[[cluster]]$
      wages2000, na.rm=TRUE)
220 meanWages.Control$meanWages2001[[cluster]] <- mean(c.Control[[cluster]]$
      wages2001, na.rm=TRUE)
221 meanWages.Control$meanWages2002[[cluster]] <- mean(c.Control[[cluster]]$
      wages2002, na.rm=TRUE)

```

```

222   meanWages.Control$meanWages2003[[cluster]] <- mean(c.Control[[cluster]]$
      wages2003, na.rm=TRUE)
223   meanWages.Control$meanWages2004[[cluster]] <- mean(c.Control[[cluster]]$
      wages2004, na.rm=TRUE)
224   meanWages.Control$meanWages2005[[cluster]] <- mean(c.Control[[cluster]]$
      wages2005, na.rm=TRUE)
225   meanWages.Control$meanWages2006[[cluster]] <- mean(c.Control[[cluster]]$
      wages2006, na.rm=TRUE)
226   meanWages.Control$meanWages2007[[cluster]] <- mean(c.Control[[cluster]]$
      wages2007, na.rm=TRUE)
227   meanWages.Control$meanWages2008[[cluster]] <- mean(c.Control[[cluster]]$
      wages2008, na.rm=TRUE)
228   meanWages.Control$meanWages2009[[cluster]] <- mean(c.Control[[cluster]]$
      wages2009, na.rm=TRUE)
229 }
230
231 ## convert meanWages lists to dataframes for convenience
232 meanWages.Control.DF <- as.data.frame(meanWages.Control)
233 meanWages.Treatment.DF <- as.data.frame(meanWages.Treatment)
234
235 ###differences calculations combined into one loop
236 ## calculate difference of control group
237 # controlDifference = mean(control, t=1) - mean(control, t=0)
238 ## calculate actual difference of treatment group
239 # treatmentDifference = mean(treatment, t=1) - mean(treatment, t=0)
240 ## calculate difference between the differences of treatment and control
      groups (treatment effect)

```



```

241 # treatmentEffect = treatmentDifference - controlDifference
242
243 ##also for the purposes of plotting:
244 ## calculate counterfactual value of treatment group at t=1
245 ## counterFactual = mean(treatment, t=0) + controlDifference
246
247 controlDifference <- numeric()
248 treatmentDifference <- numeric()
249 differenceInDifferences <- numeric()
250 counterFactual <- list()   ### for plotting
251
252 for (cluster in 1:nClusters){
253   controlDifference[cluster] <- meanWages.Control.DF$meanWages2005[cluster] -
      meanWages.Control.DF$meanWages1999[cluster] #for t0 is 1999 and t1 is
      2005
254   treatmentDifference[cluster] <- meanWages.Treatment.DF$meanWages2005[cluster
      ] - meanWages.Treatment.DF$meanWages1999[cluster] #for t0 is 1999 and t1
      is 2005
255   differenceInDifferences[cluster] <- treatmentDifference[cluster] -
      controlDifference[cluster]
256   counterFactual[cluster] <- meanWages.Treatment$meanWages1999[[cluster]] +
      controlDifference[[cluster]]
257 }
258
259
260 all.Differences <- as.data.frame(cbind(controlDifference, treatmentDifference,
      differenceInDifferences))

```

```

261
262
263 #####Plotting
264
265 plot(c(1:nClusters, 1:nClusters, 1:nClusters),c(all.Differences$
      controlDifference, all.Differences$treatmentDifference, all.Differences$
      differenceInDifferences), col = c("blue", "red", "green"))
266 plot(1:nClusters, meanWages.Control$meanWages1999, col="blue", ylim = range(
      range(meanWages.Treatment$meanWages1999), range(meanWages.Control$
      meanWages1999)))
267 par(new=T)
268 plot(1:nClusters, meanWages.Treatment$meanWages1999, col="red", ylim = range(
      range(meanWages.Treatment$meanWages1999), range(meanWages.Control$
      meanWages1999)))
269 par(new=F)
270
271 plot(1:nClusters, meanWages.Control$meanWages2005, col="blue", ylim = range(
      range(meanWages.Treatment$meanWages2005), range(meanWages.Control$
      meanWages2005)))
272 par(new=T)
273 plot(1:nClusters, meanWages.Treatment$meanWages2005, col="red", ylim = range(
      range(meanWages.Treatment$meanWages2005), range(meanWages.Control$
      meanWages2005)))
274 par(new=F)
275
276 #D-i-D Plot
277 for (cluster in 1:nClusters){

```

```

278
279   plotLabel <- paste("D-i-D_for_Cluster", as.character(cluster))#, "of", as.
      character(nClusters))
280
281
282   plot(c(1999, 1999, 2005, 2005), c(meanWages.Control$meanWages1999[cluster],
      meanWages.Treatment$meanWages1999[cluster], meanWages.Control$
      meanWages2005[cluster], meanWages.Treatment$meanWages2005[cluster]),
      main = plotLabel, pch=16, xlab = "Year", ylab = "Wages")
283
284   arrows(1999, meanWages.Control$meanWages1999[cluster], 2005, meanWages.
      Control$meanWages2005[cluster], col="blue")
285   arrows(1999, meanWages.Treatment$meanWages1999[cluster], 2005, meanWages.
      Treatment$meanWages2005[cluster], col="red")
286
287   points(2005, counterFactual[[cluster]])
288   arrows(1999, meanWages.Treatment$meanWages1999[cluster], 2005,
      counterFactual[[cluster]], lty = "dashed")
289
290   legend("right", legend = c("control", "treatment", "counterfactual"), lty=c
      (1,1,2), lwd=c(1,1,1), col=c("blue","red", "black"))
291
292   dev.copy(png, paste("DiD", as.character(cluster), "_", as.character(
      nClusters), filterTag, ".png", sep = ""), width = 600, height = 480,
      units = "px")
293   dev.off()
294

```

```

295 }
296 #
      #####

297
298 ## Cdavis script to run changes-in-changes analysis on some data
299
300 controlChange <- numeric()
301 treatmentChange <- numeric()
302 changesInChanges <- numeric()
303 # counterFactual <- list() ### for plotting
304
305 for (cluster in 1:nClusters){
306   controlChange[cluster] <- (meanWages.Control.DF$meanWages2005[cluster] -
      meanWages.Control.DF$meanWages2001[cluster]) - (meanWages.Control.DF$
      meanWages1995[cluster] - meanWages.Control.DF$meanWages1995[cluster]) #
      delta after minus delta before for 1995-1999 and 2001-2005
307   treatmentChange[cluster] <- (meanWages.Treatment.DF$meanWages2005[cluster] -
      meanWages.Treatment.DF$meanWages2001[cluster]) - (meanWages.Treatment.
      DF$meanWages1995[cluster] - meanWages.Treatment.DF$meanWages1995[cluster
      ]) #delta after minus delta before for 1995-1999 and 2001-2005
308   changesInChanges[cluster] <- treatmentChange[cluster] - controlChange[
      cluster]
309   # counterFactual[cluster] <- meanWages.Treatment$meanWages1999[[cluster]] +
      controlDifference[[cluster]]
310 }
311

```

```

312 all.changes <- as.data.frame(cbind(controlChange, treatmentChange,
    changesInChanges))
313
314 #####-----Regression based
315 #  $Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \epsilon_{ist}$ 
316 # Where:
317 #  $\gamma_s$  group effect (time invariant)
318 #  $\lambda_t$  time effect (group invariant)
319 #  $\delta$  treatment effect
320 #  $D_{st}$  indicator variable for treatment condition
321 #  $\epsilon_{ist}$  is error (epsilon is assumed to be 0)
322
323 #build dataframe for D-i-D regression (treatment year 2000)
324 counties.TriIrs <- merge(countiesWithTri, countiesWithIrs, by.x = "geoID", by.
    y = "GeoID", all = FALSE)
325
326 counties.TriIrsClust <- merge(counties.TriIrs, countiesWithClusters, by.x = "
    geoID", by.y = "geoID", all=FALSE)
327
328 counties.forRegress <- data.frame(geoID=counties.TriIrsClust$geoID, treatment=
    as.numeric(counties.TriIrsClust$tri2000 >= 1), wages1999=counties.
    TriIrsClust$wages1999, wages2005=counties.TriIrsClust$wages2005, cluster =
    counties.TriIrsClust$counties.C[[6]], bin1 = 1, bin0 = 0, filter0105 =
    as.numeric(counties.TriIrsClust$tri2001 + counties.TriIrsClust$tri2002 +
    counties.TriIrsClust$tri2003 + counties.TriIrsClust$tri2004 + counties.
    TriIrsClust$tri2005 != 0))
329

```

```

330 wageData <- data.frame(wages = c(counties.forRegress$wages1999, counties.
    forRegress$wages2005), pre = c(counties.forRegress$bin1, counties.
    forRegress$bin0), post = c(counties.forRegress$bin0, counties.forRegress$
    bin1), treatment = c(counties.forRegress$treatment, counties.forRegress$
    treatment), cluster = c(counties.forRegress$cluster, counties.forRegress$
    cluster), filter = c(counties.forRegress$filter0105, counties.forRegress$
    filter0105))
331
332 ###use next line only if filtering for later facility additions
333 #wageData <- subset(wageData, filter == 0) ## filter out counties with new
    facilities 01-05
334
335 fitTestAll <- lm(wages ~ post + treatment + post*treatment, data = wageData)
336 summary(fitTestAll)
337
338 fitTest1 <- lm(wages ~ post + treatment + post*treatment, data = subset(
    wageData, cluster== 1))
339 summary(fitTest1)
340
341 fitTest2 <- lm(wages ~ post + treatment + post*treatment, data = subset(
    wageData, cluster== 2))
342 summary(fitTest2)
343
344 fitTest3 <- lm(wages ~ post + treatment + post*treatment, data = subset(
    wageData, cluster== 3))
345 summary(fitTest3)
346

```

```

347 fitTest4 <- lm(wages ~ post + treatment + post*treatment, data = subset(
      wageData, cluster== 4))
348 summary(fitTest4)
349
350 fitTest5 <- lm(wages ~ post + treatment + post*treatment, data = subset(
      wageData, cluster== 5))
351 summary(fitTest5)
352
353 fitTest6 <- lm(wages ~ post + treatment + post*treatment, data = subset(
      wageData, cluster== 6))
354 summary(fitTest6)
355 #
356 #
      #####
357
358 ### Wages standardized by number of returns per county
359
360 meanWages.Treatment <- list() ##foo$year[clusters1:10]
361 meanWages.Control <- list()
362
363
364 for (cluster in 1:nClusters){
365   ## taking 1989 to be year 1 through 2009 as year 21
366   ##treatment groups:
367   meanWages.Treatment$meanWages1989[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages1989, na.rm=TRUE) ##removes NA values (counties without

```

*recorded wages; could be counties that were merged into other small pop
counties or counties that did not exist that year)*

```

368 meanWages.Treatment$meanWages1990[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages1990, na.rm=TRUE)
369 meanWages.Treatment$meanWages1991[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages1991, na.rm=TRUE)
370 meanWages.Treatment$meanWages1992[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages1992, na.rm=TRUE)
371 meanWages.Treatment$meanWages1993[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages1993, na.rm=TRUE)
372 meanWages.Treatment$meanWages1994[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages1994, na.rm=TRUE)
373 meanWages.Treatment$meanWages1995[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages1995, na.rm=TRUE)
374 meanWages.Treatment$meanWages1996[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages1996, na.rm=TRUE)
375 meanWages.Treatment$meanWages1997[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages1997, na.rm=TRUE)
376 meanWages.Treatment$meanWages1998[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages1998, na.rm=TRUE)
377 meanWages.Treatment$meanWages1999[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages1999, na.rm=TRUE)
378 meanWages.Treatment$meanWages2000[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages2000, na.rm=TRUE)
379 meanWages.Treatment$meanWages2001[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages2001, na.rm=TRUE)

```



```

380 meanWages.Treatment$meanWages2002[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages2002, na.rm=TRUE)
381 meanWages.Treatment$meanWages2003[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages2003, na.rm=TRUE)
382 meanWages.Treatment$meanWages2004[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages2004, na.rm=TRUE)
383 meanWages.Treatment$meanWages2005[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages2005, na.rm=TRUE)
384 meanWages.Treatment$meanWages2006[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages2006, na.rm=TRUE)
385 meanWages.Treatment$meanWages2007[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages2007, na.rm=TRUE)
386 meanWages.Treatment$meanWages2008[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages2008, na.rm=TRUE)
387 meanWages.Treatment$meanWages2009[[cluster]] <- mean(c.Treatment[[cluster]]$
      standWages2009, na.rm=TRUE)
388
389 ##control groups
390 meanWages.Control$meanWages1989[[cluster]] <- mean(c.Control[[cluster]]$
      standWages1989, na.rm=TRUE)
391 meanWages.Control$meanWages1990[[cluster]] <- mean(c.Control[[cluster]]$
      standWages1990, na.rm=TRUE)
392 meanWages.Control$meanWages1991[[cluster]] <- mean(c.Control[[cluster]]$
      standWages1991, na.rm=TRUE)
393 meanWages.Control$meanWages1992[[cluster]] <- mean(c.Control[[cluster]]$
      standWages1992, na.rm=TRUE)

```

```

394 meanWages.Control$meanWages1993[[cluster]] <- mean(c.Control[[cluster]]$
      standWages1993, na.rm=TRUE)
395 meanWages.Control$meanWages1994[[cluster]] <- mean(c.Control[[cluster]]$
      standWages1994, na.rm=TRUE)
396 meanWages.Control$meanWages1995[[cluster]] <- mean(c.Control[[cluster]]$
      standWages1995, na.rm=TRUE)
397 meanWages.Control$meanWages1996[[cluster]] <- mean(c.Control[[cluster]]$
      standWages1996, na.rm=TRUE)
398 meanWages.Control$meanWages1997[[cluster]] <- mean(c.Control[[cluster]]$
      standWages1997, na.rm=TRUE)
399 meanWages.Control$meanWages1998[[cluster]] <- mean(c.Control[[cluster]]$
      standWages1998, na.rm=TRUE)
400 meanWages.Control$meanWages1999[[cluster]] <- mean(c.Control[[cluster]]$
      standWages1999, na.rm=TRUE)
401 meanWages.Control$meanWages2000[[cluster]] <- mean(c.Control[[cluster]]$
      standWages2000, na.rm=TRUE)
402 meanWages.Control$meanWages2001[[cluster]] <- mean(c.Control[[cluster]]$
      standWages2001, na.rm=TRUE)
403 meanWages.Control$meanWages2002[[cluster]] <- mean(c.Control[[cluster]]$
      standWages2002, na.rm=TRUE)
404 meanWages.Control$meanWages2003[[cluster]] <- mean(c.Control[[cluster]]$
      standWages2003, na.rm=TRUE)
405 meanWages.Control$meanWages2004[[cluster]] <- mean(c.Control[[cluster]]$
      standWages2004, na.rm=TRUE)
406 meanWages.Control$meanWages2005[[cluster]] <- mean(c.Control[[cluster]]$
      standWages2005, na.rm=TRUE)

```

```

407   meanWages.Control$meanWages2006[[cluster]] <- mean(c.Control[[cluster]]$
      standWages2006, na.rm=TRUE)
408   meanWages.Control$meanWages2007[[cluster]] <- mean(c.Control[[cluster]]$
      standWages2007, na.rm=TRUE)
409   meanWages.Control$meanWages2008[[cluster]] <- mean(c.Control[[cluster]]$
      standWages2008, na.rm=TRUE)
410   meanWages.Control$meanWages2009[[cluster]] <- mean(c.Control[[cluster]]$
      standWages2009, na.rm=TRUE)
411 }
412
413 ## convert meanWages lists to dataframes for convenience
414 meanWages.Control.DF <- as.data.frame(meanWages.Control)
415 meanWages.Treatment.DF <- as.data.frame(meanWages.Treatment)
416
417 ###differences calculations combined into one loop
418 ## calculate difference of control group
419 # controlDifference = mean(control, t=1) - mean(control, t=0)
420 ## calculate actual difference of treatment group
421 # treatmentDifference = mean(treatment, t=1) - mean(treatment, t=0)
422 ## calculate difference between the differences of treatment and control
      groups (treatment effect)
423 # treatmentEffect = treatmentDifference - controlDifference
424
425 ##also for the purposes of plotting:
426 ## calculate counterfactual value of treatment group at t=1
427 ## counterFactual = mean(treatment, t=0) + controlDifference
428

```

```

429 controlDifference <- numeric()
430 treatmentDifference <- numeric()
431 differenceInDifferences <- numeric()
432 counterFactual <- list()    ### for plotting
433
434 for (cluster in 1:nClusters){
435   controlDifference[cluster] <- meanWages.Control.DF$meanWages2005[cluster] -
      meanWages.Control.DF$meanWages1999[cluster] #for t0 is 1999 and t1 is
      2005
436   treatmentDifference[cluster] <- meanWages.Treatment.DF$meanWages2005[cluster
      ] - meanWages.Treatment.DF$meanWages1999[cluster] #for t0 is 1999 and t1
      is 2005
437   differenceInDifferences[cluster] <- treatmentDifference[cluster] -
      controlDifference[cluster]
438   counterFactual[cluster] <- meanWages.Treatment$meanWages1999[[cluster]] +
      controlDifference[[cluster]]
439 }
440
441
442 all.Differences <- as.data.frame(cbind(controlDifference, treatmentDifference,
      differenceInDifferences))
443
444
445 ####Plotting
446
447 plot(c(1:nClusters, 1:nClusters, 1:nClusters),c(all.Differences$
      controlDifference, all.Differences$treatmentDifference, all.Differences$

```

```

        differenceInDifferences), col = c("blue", "red", "green"))
448 plot(1:nClusters, meanWages.Control$meanWages1999, col="blue", ylim = range(
        range(meanWages.Treatment$meanWages1999), range(meanWages.Control$
        meanWages1999)))
449 par(new=T)
450 plot(1:nClusters, meanWages.Treatment$meanWages1999, col="red", ylim = range(
        range(meanWages.Treatment$meanWages1999), range(meanWages.Control$
        meanWages1999)))
451 par(new=F)
452
453 plot(1:nClusters, meanWages.Control$meanWages2005, col="blue", ylim = range(
        range(meanWages.Treatment$meanWages2005), range(meanWages.Control$
        meanWages2005)))
454 par(new=T)
455 plot(1:nClusters, meanWages.Treatment$meanWages2005, col="red", ylim = range(
        range(meanWages.Treatment$meanWages2005), range(meanWages.Control$
        meanWages2005)))
456 par(new=F)
457
458 #D-i-D Plot
459 for (cluster in 1:nClusters){
460
461     plotLabel <- paste("D-i-D_for_Cluster", as.character(cluster))#, "of", as.
        character(nClusters), "_byReturns")
462
463

```

```

464   plot(c(1999, 1999, 2005, 2005, 2005), c(meanWages.Control$meanWages1999[
      cluster], meanWages.Treatment$meanWages1999[cluster], meanWages.Control$
      meanWages2005[cluster], meanWages.Treatment$meanWages2005[cluster],
      counterFactual[[cluster]]), main = plotLabel, pch=16, xlab = "Year",
      ylab = "Wages/Reporter")
465
466   arrows(1999, meanWages.Control$meanWages1999[cluster], 2005, meanWages.
      Control$meanWages2005[cluster], col="blue")
467   arrows(1999, meanWages.Treatment$meanWages1999[cluster], 2005, meanWages.
      Treatment$meanWages2005[cluster], col="red")
468
469   arrows(1999, meanWages.Treatment$meanWages1999[cluster], 2005,
      counterFactual[[cluster]], lty = "dashed")
470
471   legend("bottomright", legend = c("control", "treatment", "counterfactual"),
      lty=c(1,1,2), lwd=c(1,1,1), col=c("blue","red", "black"))
472
473   dev.copy(png, paste("DiD", as.character(cluster), "_", as.character(
      nClusters), filterTag, "_byReturns_20130513", ".png", sep = ""), width =
      600, height = 480, units = "px")
474   dev.off()
475
476 }
477 #
      #####

```

478

```

479 ## Cdavis script to run changes-in-changes analysis on some data
480
481 controlChange <- numeric()
482 treatmentChange <- numeric()
483 changesInChanges <- numeric()
484 # counterFactual <- list() ### for plotting
485
486 for (cluster in 1:nClusters){
487   controlChange[cluster] <- (meanWages.Control.DF$meanWages2005[cluster] -
      meanWages.Control.DF$meanWages2001[cluster]) - (meanWages.Control.DF$
      meanWages1995[cluster] - meanWages.Control.DF$meanWages1995[cluster]) #
      delta after minus delta before for 1995-1999 and 2001-2005
488   treatmentChange[cluster] <- (meanWages.Treatment.DF$meanWages2005[cluster] -
      meanWages.Treatment.DF$meanWages2001[cluster]) - (meanWages.Treatment.
      DF$meanWages1995[cluster] - meanWages.Treatment.DF$meanWages1995[cluster
      ]) #delta after minus delta before for 1995-1999 and 2001-2005
489   changesInChanges[cluster] <- treatmentChange[cluster] - controlChange[
      cluster]
490   # counterFactual[cluster] <- meanWages.Treatment$meanWages1999[[cluster]] +
      controlDifference[[cluster]]
491 }
492
493 all.changes <- as.data.frame(cbind(controlChange, treatmentChange,
      changesInChanges))
494
495 #####-----Regression based
496 #  $Y_{ist} = \gamma_s + \lambda_t + \delta_{st} + \epsilon_{ist}$ 

```

```

497 # Where:
498 # \gamma_s group effect (time invariant)
499 # \lambda_t time effect (group invariant)
500 # \delta treatment effect
501 # D_{st} indicator variable for treatment condition
502 # \epsilon_{ist} is error ( epsilon is assumed to be 0)
503
504 #build dataframe for D-i-D regression (treatment year 2000)
505 counties.TriIrs <- merge(countiesWithTri, countiesWithIrs, by.x = "geoID", by.
    y = "GeoID", all = FALSE)
506
507 counties.TriIrsClust <- merge(counties.TriIrs, countiesWithClusters, by.x = "
    geoID", by.y = "geoID", all=FALSE)
508
509 counties.forRegress <- data.frame(geoID=counties.TriIrsClust$geoID, treatment=
    as.numeric(counties.TriIrsClust$stri2000 >= 1), wages1999=counties.
    TriIrsClust$standWages1999, wages2005=counties.TriIrsClust$standWages2005,
    cluster = counties.TriIrsClust$"counties.C[[6]]", bin1 = 1, bin0 = 0,
    filter0105 = as.numeric(counties.TriIrsClust$stri2001 + counties.
    TriIrsClust$stri2002 + counties.TriIrsClust$stri2003 + counties.TriIrsClust$
    tri2004 + counties.TriIrsClust$stri2005 != 0))
510
511 wageData <- data.frame(wages = c(counties.forRegress$wages1999, counties.
    forRegress$wages2005), pre = c(counties.forRegress$bin1, counties.
    forRegress$bin0), post = c(counties.forRegress$bin0, counties.forRegress$
    bin1), treatment = c(counties.forRegress$treatment, counties.forRegress$
    treatment), cluster = c(counties.forRegress$cluster, counties.forRegress$

```



```

        cluster), filter = c(counties.forRegress$filter0105 , counties.forRegress$
        filter0105))
512
513 ###use next line only if filtering for later facility additions
514 #wageData <- subset(wageData, filter == 0) ## filter out counties with new
facilities 01-05
515
516 fitTestAll <- lm(wages ~ post + treatment + post*treatment, data = wageData)
517 summary(fitTestAll)
518
519 fitTest1 <- lm(wages ~ post + treatment + post*treatment, data = subset(
        wageData, cluster== 1))
520 summary(fitTest1)
521
522 fitTest2 <- lm(wages ~ post + treatment + post*treatment, data = subset(
        wageData, cluster== 2))
523 summary(fitTest2)
524
525 fitTest3 <- lm(wages ~ post + treatment + post*treatment, data = subset(
        wageData, cluster== 3))
526 summary(fitTest3)
527
528 fitTest4 <- lm(wages ~ post + treatment + post*treatment, data = subset(
        wageData, cluster== 4))
529 summary(fitTest4)
530

```

```

531 fitTest5 <- lm(wages ~ post + treatment + post*treatment, data = subset(
      wageData, cluster== 5))
532 summary(fitTest5)
533
534 fitTest6 <- lm(wages ~ post + treatment + post*treatment, data = subset(
      wageData, cluster== 6))
535 summary(fitTest6)

```