36-402 Data Exam 2: Invesigating Economies

Anusha Bhat

2024-04-25

1. Introduction

1.1 Identifying the Scientific Questions

Census surveys are incredibly invaluable for learning information on current demographics of certain areas, which can then be used to gauge what issues currently afflict an area and what policies and programs can be developed to address and resolve them. To conduct these surveys, the U.S. Bureau of Economic Analysis breaks up urban areas into smaller "Metropolitan Statistical Areas", which are determined by trends in residence and commuting. A particular variable of interest to the U.S. Bureau of Economic Analysis is "gross metropolitan products"–a metric similar to gross national product.

Currently, there are three prominent theories regarding the arisal of trends in the per-capita gross metropolitan product for various metropolitan statistical areas. The first theory claims that population effects per-capita output and the proportion of the city's economy that is in high-valued industries. The second theory claims that population effects the proportion of the city's economy that is in high-valued industries, which in turn effects the per-capita output. The final theory claims that the industries in a city's economy are obtained by chance, however, these industries effect per-capita output which influences population size. We aim to investigate which of these theories best fits the data.

We will conduct our investigation using causal inference through comparing and contrasting directed acyclic graphs (DAGs) and exploratory data analysis. Specifically, we will determine the DAG for each theory, use each DAG to estimate the causal effect of population on the per-capita gross metropolitan product as well as the effect of the information and communications technology on the per-capita gross metropolitan product, and identify and test conditional independences that differ between ech DAG. Once we determine the best-fitting theory, we will then investigate whether the local economy of an urban area can be improved more by increasing population or increasing the share of information and communications technology. We will determine this by comparing the effects of doubling the population of Pittsburgh to the effects of increasing the share of information and communications technology in it's respective local economy by 10% on the per-capita gross metropolitan product. Investigating the factors that influence per-capita gross metropolitan product will be useful for city planning and addressing the socio-economic issues that often plague urban areas.

1.2 Elementary Data Analysis

The data for this report was provided by Dr. Shalizi, and contains several recorded variables from surveys conducted in various metropolitan statistical areas in 2006. These variables included represent the name of the area, the per-capita gross metropolitan product (pcgmp)-measured in dollars per person per year-the population of the area, and the proportion of the city's economy that is dependent on the finance, professional and technical services (prof tech), information and communications technologies (ICT), and management services industries. There are 366 observations in our dataset, 357 of which have at least one variable with an "NA" value for one of the industries. This is due to some variables not being recorded for some cities, however, we are able to proceed with our analysis since a majority of the data is available for each city.

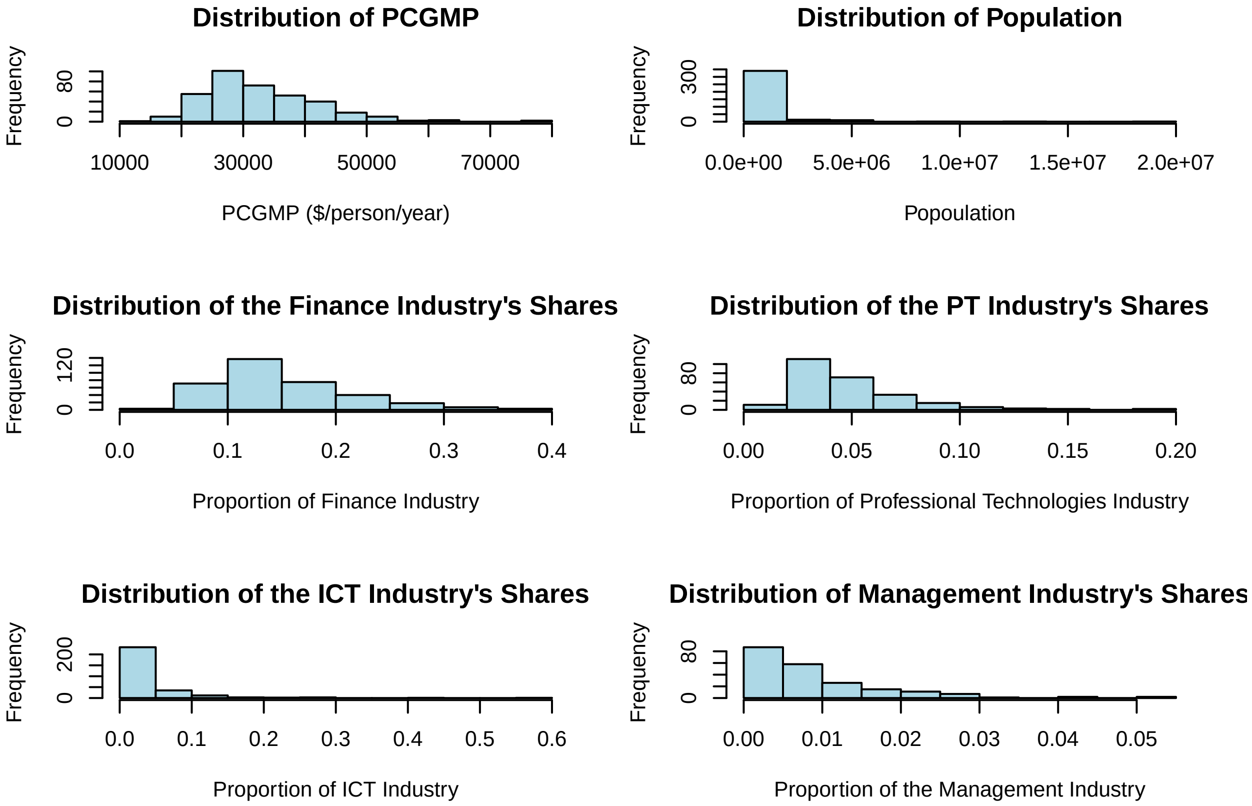

Figure 1: Distribution of the numerical variables.

We can observe that pcgmp ranges from $10,000-$65,000 per year with an outlier above $75,000 per year. The population ranges from 0-10,000,000 individuals and the ICT industry shares range from 0-0.15%. The distributions for the finance's industry's shares and the professional technologies' industry shares both have a somewhat bell-curve shape skewed to the right by a long right tail.

Below is a table displaying the averages and their standard errors for each of the variables in our data set (omitting for NA observations and excluding the names of the cities since that is not numerical) as well as a second table including the specific observations for Pittsburgh. We can observe that for Pittsburgh and the means of all of the cities, the finance industry appears to have the largest share of the economies.

Table 1: Means and Standard Errors For Each Variable

	Mean	Standard Error
Population	6.808977e+05	6.606324e+09
PCGMP	3.292276e+04	2.322587e+05
Finance	1.491661e-01	1.070000e-05
Prof Tech	4.735540e-02	2.100000e-06
ICT	3.963010e-02	8.700000e-06
Management	8.870100e-03	2.000000e-07

Table 2: Observations For Pittsburgh

MSA	PCGMP	Population	Finance	Prof Tech	ICT	Management
Pittsburgh, PA	38350	2361000	0.2018	0.0777	0.03434	0.02946

In figure 2, we observe the relationship between population and each of the numerical variables. All of the relationships appear to have a positive association, however, they do not appear to be linear. This indicates that for our models, we must use non-parametric regression.

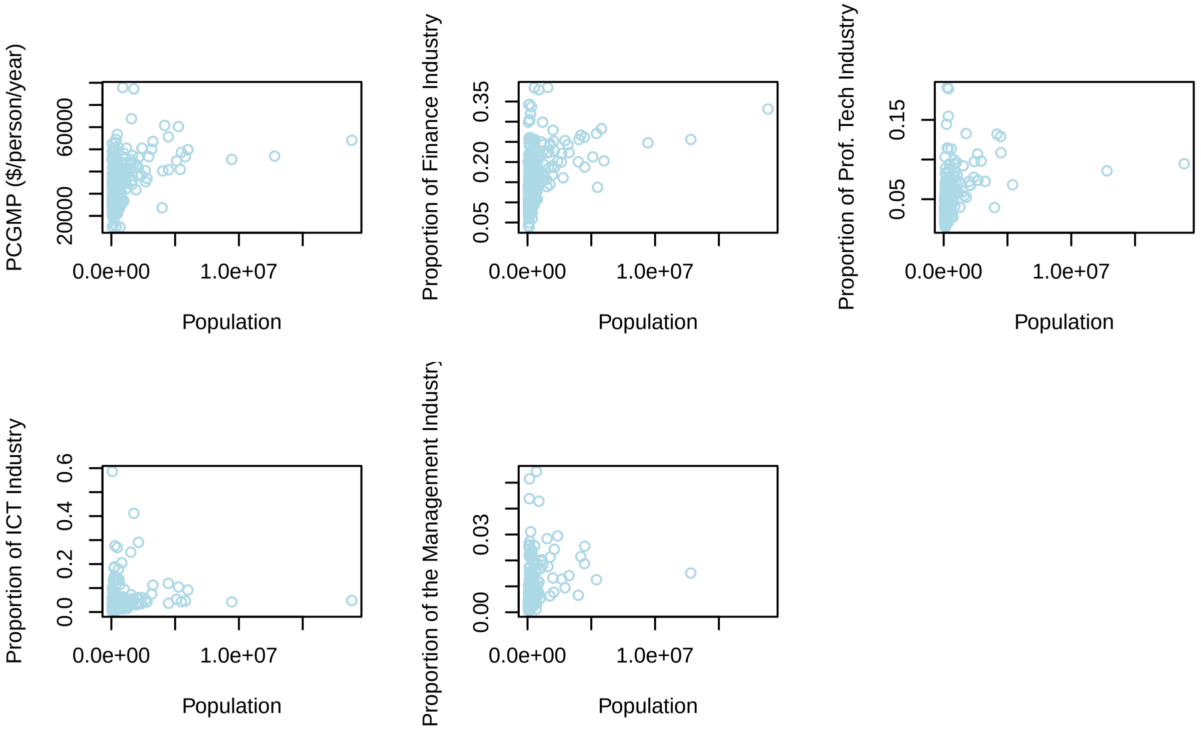

Figure 2: Scatterplots of population v.s. each of the numerical variables.

2. Analyses

2.1 Creating Directed Acyclic Graphs

2.1.1 DAG for Theory 1

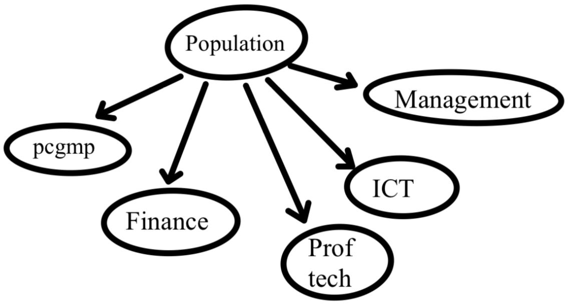

Figure 3: DAG for theory 1.

The first theory describes that "increasing population causes higher per-capita output" and it also "causes more of the city's economy o be in high-value industries". This implies that there is a cause and effect relationship between population and pcgmp, as well as, between population and the four industries in our dataset–i.e., changing population changes the distribution of the other variables. This reveals that in a DAG, population must be the parent of the other numerical variables in our dataset, leading us to the DAG shown in figure 3.

2.1.2 DAG for Theory 2

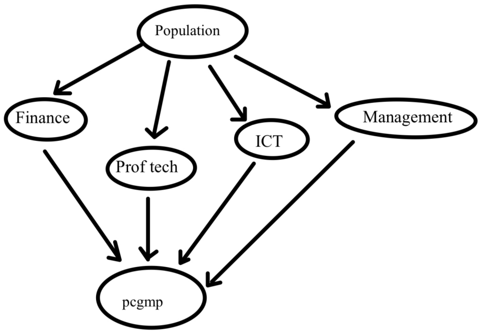

Figure 4: DAG for theory 2.

The second theory describes that higher-valued industries are located in cities with larger populations due to a higher amount of customers that they can serve. This implies that "population causes industry shares, and industry shares cause per-capita output", i.e. changing population alters the distribution of industry shares and changing industry shares alters the distribution of pcgmp. The highlight a parent-child relationship where population is the parent and each of the industries are the children. Another parent-child relationship that is highlight by this theory is that each of the industries effect pcgmp so pcgmp is the child of all of the industries. This leads us to the DAG in figure 4.

2.1.3 DAG for Theory 3

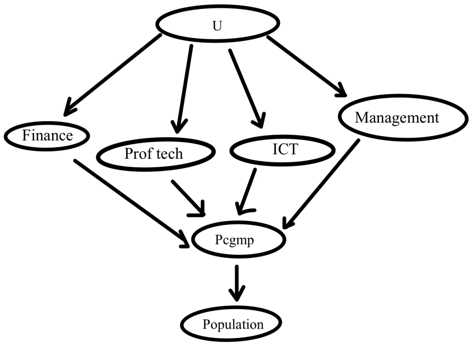

Figure 5: DAG for theory 3.

The third theory describes that industry shares of the economy are determined by chance. For example, for some cities, ICT will have a majority share of the economy, whereas, in other cities, the finance industry may have a majority share of the economy, and these proportions are determined largely by luck/chance. For our DAG, this indicates that each of the industries should have a parent “U” that indicates unobserved variables which influence the luck/chances mentioned previously. The theory further explains that some industries pay more, and since income is necessary for pcgmp, this indicates that the industry shares effect pcgmp–i.e. the industry shares are all the parents of pcgmp. Finally, the theory states that “people move to places where the output is high”, indicating that pcgmp effects population, highlighting the fact that in our DAG, pcgmp should be the parent of our population. Putting these pieces together, we obtain the DAG in figure 5.

2.2 How to Estimate Causal Effects

2.2.1 Population on Per-Capita GMP

For theories 1 and 2, we observe that pcgmp is a descendant of population in their respective DAGs. In theory 1 pcgmp is a child of population, whereas, in theory 2 it is a grandchild. However, in both DAGs, there are no backdoor paths from population to pcgmp, therefore, we do not need to control for any other variables when estimating the causal effect of population on pcgmp. We will specifically observe the causal effect by doubling Pittsburgh’s population. First, we will perform a non-parametric kernel regression with the model “pcgmp ~ population” using the Pittsburgh observation from the original dataset. We will repeat this model using the Pittsburgh observation but with the population doubled. Additionally, the kernel regression models should use tol and ftol values of 0.01 due to the size of the dataset. Once we have obtained and run our two models, we can predict a pcgmp value for each model, then subtract the prediction from the original observation from the prediction of the new observation (with Pittsburgh's population doubled). This difference is an estimate of the causal effect since we have essentially done $Do(X = 2 * Pittsburgh' spopulation)$. To gauge uncertainties of the estimate, we will conduct leave-one-out cross-validation (LOOCV) for each model as well as case-resampling bootstrapping with 95% confidence for 100 repetitions.

For theory 3, population is the child of pcgmp, indicating that there is no effect of population on pcgmp. The only cause-and-effect relationship between these two variables is pcgmp on population. Therefore, we can not estimate the effect of population on pcgmp with this theory.

2.2.2 Industry Shares on Per-Capita GMP

For theory 1, we do not need to control for anything since the only path from ICT to pcgmp is through population, however, controlling for population will close this single path, so, we won't be able to get an effect of the causal estimate if we control for population. For theories 2 and 3, we observe that pcgmp is a child of the ICT industry in their respective DAGs. Additionally, in theory 2, there is a back-door path from the ICT industry to pcgmp through population and the other industries, so for this theory, we will need to control for either population or all of the other industries to block the back-door paths. In theory 3, there is a back-door path from the ICT industry to pcgmp through the unobserved variable, U and the other industries. However, since U is unobserved, we can not control for it, so, we can only control for the other industries.

We will specifically observe the causal effect by increasing the ICT industry's share in Pittsburgh's economy by 10%.

For theory 1, first, we will perform a non-parametric kernel regression with the model "pcgmp ~ ICT" using the Pittsburgh observation from the original dataset. We will repeat this model using a an observation with a 10% increase in the ICT value of the original observation that Pittsburgh had in 2006. Additionally, the kernel regression models should use tol and ftol values of 0.01 due to the size of the dataset. Once we have obtained and run our two models, we can predict a pcgmp value for each model, then subtract the prediction on the original observation from the prediction of the new observation (with Pittsburgh's ICT share increased). This difference is an estimate of the causal effect since we have essentially done Do(X=1.1 * Pittsburgh's ICT share). For theory 2, we will perform the same procedure as mentioned for theory 1, except the model should be "pcgmp ~ ICT + population" since we have to control for either population or all of the industrial variables. Again, we will repeat this same procedure for theory 3, except we will use the model "pcgmp ~ ICT + finance + prof tech + management" using the same logic. To gauge uncertainties of the estimates, we will conduct LOOCV for each model as well as case-resampling bootstrapping with 95% confidence for 100 repetitions.

2.3 Estimating Causal Effects

2.3.1 Population on Per-Capita GMP

Below is a table of the estimate of the causal effect of population on pcgmp for theories 1 and 2, along with bootstrapped standard errors and a confidence interval for the estimates.

Table 3: Causal Effect of Population on PCGMP With Bootstrapped Std. Error and CI

	Estimate	Std. Error	CI Low	CI High
Theories 1 & 2	6164.726	39324.76	-4934.253	50679.36

We can observe that doubling the population of Pittsburgh has an estimated causal effect of $6,164.726 million per year, i.e. increasing population leads to an increase in pcgmp.

2.3.2 ICT Industry Shares on Per-Capita GMP

Below is a table of the estimate of the causal effect of ICT shares on pcgmp for theories 2 and 3, and bootstrapped standard errors and confidence intervals for each estimate.

Table 4: Causal Effect of ICT on PCGMP With Bootstrapped Std. Error and CI

	Estimate	Std. Error	CI Low	CI High
Theory 1	789.31600	5252.343	-245.4927	7182.44200
Theory 2	225.14160	3273.275	-4041.4900	587.61970
Theory 3	25.13914	1050.379	-1409.8870	75.57363

We can observe that for each theory, all of the causal effect estimates are positive, indicating that increasing ICT leads to an increase in pcgmp.

2.4 Conditional Independences in Each Theory

In theory 1, professional technologies is conditionally independent from ICT given population and pcgmp. Conditioning on population blocks the only open path from professional technologies to ICT in the DAG, allowing for the two variables to become independent. Controlling for pcgmp in addition does not add any information since pcgmp does not effect the relationship between professional technologies and ICT, however, it allows us to identify a conditional independence relationship that holds in theory 1 but not in theory 2 or 3. It does not hold in theory 2 since professional technologies and ICT are independent only given population using the back-door criterion. Additionally, since pcgmp is a child of both professional technologies and ICT in the DAG for theory 2, we do not need to condition on it for independence. Moreover, in theory 3 the same logic applies as to why we would not need pcgmp for conditional independence. In this theory, professional technologies and ICT are independent given U using the back-door criteria (not population).

In theory 2, pcgmp is conditionally independent from population given all of the industrial variables, since conditioning on those variables blocks all paths between pcgmp and population to make them independent from one another. This relationship does not hold in theory 1 since pcgmp is a child of population and there are no back-door paths between the two variables, therefore, they can not be conditionally independent from one another. In theory 3, this does not hold since population is the child of pcgmp and there are no back-door paths between the two variables to block, therefore, again, they can not be conditionally independent.

In theory 3, professional technologies is independent from population given pcgmp. All of the back-door and front-door paths from professional technologies to population go throug pcgmp, so, conditioning on pcgmp will make the two variables independent. This relationship does not hold in theory 1 since professional technologies is a child of population and there are no back-door paths to block between the two variables so they can not be conditionally independent. Furthermore, this relationship does not hold for theory 2 for the same reasoning as why it does not hold for theory 1.

2.5 Testing Conditional Independences in Each Theory

We can test the conditional independencies for each theory by fitting a Generalized Additive Model (gam) representing each relationship and observing whether the partial responses are significantly non-zero for the variables that we controlled for. For example, if we claimed that X is independent from Y given Z, then we will predict Y from both X and Z to make conclusions on the validity of this relationship based on the partial response functions. If the partial response functions are not significantly non-zero, this indicates that increasing the variables we predicted Y from does not effect the value of the response variable, i.e. the conditional relationship holds for the data. Measures of uncertainty will be given in the form of boot-strapped confidence intervals using 200 repetitions and 95% confidence level.

2.5.1 Testing Theory 1

For theory 1, we will test whether or not professional technologies is conditionally independent from ICT given population and pcgmp. We will use the model "gam(ICT ~ s(prof tech) + s(pop))", with NA's excluded–where ICT is Y, professional technologies is X, and population is Z.

We observe that for both of the partial response functions in figure 6, the confidence intervals overlap 0, so, the functions are not significantly non-zero. Therefore, the conditional relationship for theory 1 holds true.

2.5.2 Testing Theory 2

For theory 2, we will test whether or not pcgmp is conditionally independent from population given all of the industry variables. We will use the model "gam(population ~ s(pcgmp) + s(finance) + s(prof tech) + s(ict)

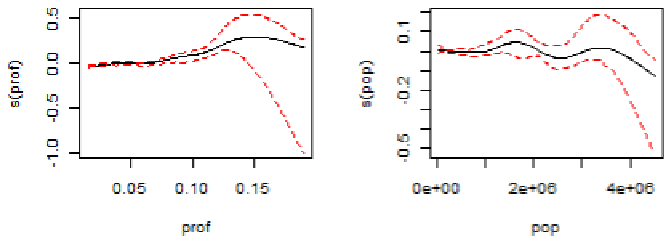

Figure 6: Partial response functions of professional technologies industry shares and population for gam testing conditional independence for theory 1.

+ s(management))" with NA's excluded–where population is Y, pcgmp is X, and the industrial variables are Z in this case.

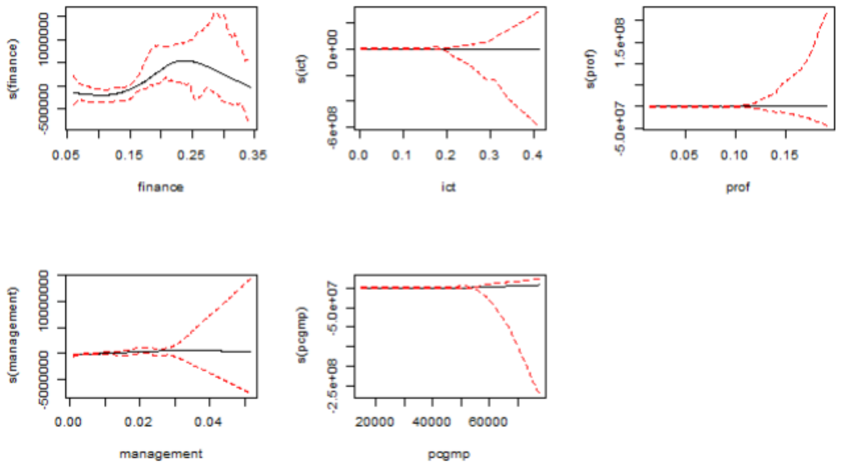

Figure 7: Partial response functions of all industry shares and pcgmp for gam testing conditional independence for theory 2.

We observe that for all of the partial response functions in figure 7, the confidence intervals overlap 0, so, the functions are not significantly non-zero. Therefore, the conditional relationship for theory 2 holds true.

2.5.2 Testing Theory 3

For theory 2, we will test whether or not professional technologies is conditionally independent from population given pcgmp. We will use the model "gam(population ~ s(prof) + s(pcgmp))" with NA's excluded–where population is Y, professional technologies is X, and pcgmp is Z in this case.

We observe that for both of the partial response functions in figure 8, the confidence intervals overlap 0, so, the functions are not significantly non-zero. Therefore, the conditional relationship for theory 3 holds true.

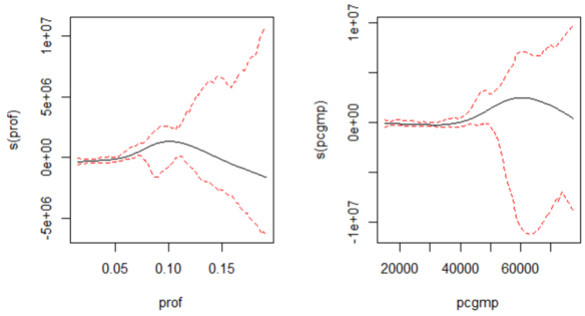

Figure 8: Partial response functions of professional technologies and pcgmp for gam testing conditional independence for theory 3.

Conclusion

We find that doubling Pittsburgh's population was associated with a positive causal effect estimate, indicating that population leads to an increase in pcgmp. For theory 3 we could not assess this due to the structure of the DAG, however, this estimates holds valid for theory 1 and 2. We find that increasing the ICT industry's share by 10% for Pittsburgh is associated with a positive causal effect estimate for all of the theories, controlling for the respective variables necessitated by their respective DAGS. This indicates that increasing the ICT shares leads to an increase in pcgmp. These two results suggest that pcgmp is a descendant of both population and ICT. Theory 2 is the only theory where this holds true. We could additionally use the conditional independence analysis to determine which theory best fits the data, however, more analysis must be done to do this since we found that all of the conditional independence relationships held true for each theory as all of the partial response functions included 0 within their confidence intervals. Thus, based on the analysis of the causal effects of population and ICT on pcgmp, we conclude that theory 2 best fits this data. To restate the the theory, this theory claims that higher-valued industries tend to settle in cities with higher populations since their is a larger customer base, so population influences industry shares and industry shares then influence per-capita output. Based on our analysis using theory 2, we find that doubling Pittsburgh's population has a causal effect estimate of increasing pcgmp by $6,164.726 million/year (95% CI[-4934.253, 50679.36]). Additionally, we find that increasing Pittsburgh's ICT share by 10% has a causal effect estimate of increasing pcgmp by $225.1416 million/year (95% CI[-4041.4900, 587.61970]).

Future work can be done to further investigate factors that influence economies, such as repeating the procedures in this report using other variables, collecting more variables for analysis, and testing other models in addition to kernel regression and GAM.