Capstone Project Report: Regulatory Studies Center

Causal Analytic Toolkit

Daniel X. Kim and Cindy Huang

5/8/2017

Daniel X. Kim
Cindy Huang
Capstone Assignment: Final Paper
May 08, 2017

Contents

Daniel X. Kim
Cindy Huang
Capstone Assignment: Final Paper
May 08, 2017

1. Executive Summary

The George Washington Regulatory Studies Center (RSC) is an academic center established in 2009 housed in the Trachtenberg School of Public Policy and Public Administration. The RSC recently developed computer software, Causal Analytics Tool (CAT), in hopes to help policy experts and students approach data analysis in an objective and unbiased manner.

In this report, the authors hope to achieve the following:

- Show what problem the CAT is trying to solve
  - In order to find the purpose of the software, the team conducted a literature review of the topic causality. The team gathered information about different theories on the concept of causality. In addition to the theoretical background, the team also searched for various examples of the current government research results that might suffer from researcher bias.
- Introduce the software's functionality
  - The team examined the software's functions and capabilities with a purposefully biased econometrics model for international trade and deforestation. The team used a model that is purposefully biased in order to examine the CAT's capability to correct researcher bias.
- Analyze the market competitors for the CAT
  - The team conducted market analysis to find the CAT's competitors and drew lessons from those competing programs' development processes.

The team has conducted the research using mixed methods including the qualitative methods (e.g. in-depth interviews) and quantitative methods to examine the software's functionality.

Following our analyses on the above-mentioned points, we found that the CAT can maximize its market potential by branding itself as a statistical bias checking program that could suggest corrections to the research outputs that might be based on researcher bias.

2. Client Information: GW Regulatory Studies Center

The RSC works to improve regulatory policy through research, education, and outreach (Regulatory Studies Center, 2017). The RSC strives to become the hub for the regulatory research and a training ground for both experts and students who want to better understand the effects of regulation and the various ways to analyze the benefits and costs of such regulations.

While the Causal Analytics Program is the first project for the RSC to invest in the software development, the RSC hopes to apply this toolkit to its core organizational goals. In 2017, one of the Center's research focus is to serve as a source for "objective information on regulatory matters," and within this goal, the CAT program is expected to contribute as a tool that can generate unbiased and objective information (Regulatory Studies Center, 2017).

The CAT's developers plan to continuously develop upon the already available features of the CAT with the new updates, and new packages of R (the statistical computing and graphics program) are released. More recently, the CAT has been augmented with a Predictive Analytics Toolkit (PAT), developed with support from the American Chemistry Council, that used prediction of in vivo carcinogenicity of chemicals from high throughput screening assay data as the motivating example.

3. Background Information

3.1 Researcher Bias in the current literature

Causal analysis is "a method of searching for the cause or causes of certain effects." Causation makes sense of data, guides policies, and learn from success and failures (Pearl, 2006). According to Cox, causation is the fundamental question in public policy field because the decision makers must link alternative choices to expected results (Cox, 2017).

Many of the current public policymaking decisions, however, depend heavily on models based not on objective data but on the experts' assumptions. In public health sector, many research findings are based on the unverified assumptions such as the health effects caused by air pollution (Cox, 2012). For example, Yim and Barrett, two professors at MIT, argued that "UK combustion emissions cause 13,000 premature deaths in the UK per year, while an additional 6000 deaths in the UK are caused by non-UK European Union (EU) combustion emissions" (Yim and Barrett, 2012). Their argument was directly challenged by the National Health Service report in 2012 with the following comment; "although particulate matter has been associated with premature mortality in other studies, a definitive cause-and-effect link has not yet been demonstrated" (NHS, 2012). Although some might argue that the modeling assumptions are an inevitable part of the research process, using assumption-based inferences may limit the external validity of findings, and the results might offer biased insight for preventive action (Glass 67-68).

According to Cox, "The availability of powerful statistical modeling packages has made it easy to search for combinations of modeling assumptions that imply desired (e.g., publishable) results – the problem known as p-hacking, which undermines the credibility of many published scientific findings and reported significance levels based on statistical modeling" (Cox, 2017).

3.2 The RSC's approach to the Problem

To help researchers keep objectivity in their data analysis, Cox suggested using probabilistic approach and data science that are often used in the field of artificial intelligence, which could minimize the researcher interruption on data analysis. With support from the GW Regulatory Studies Center, he developed the Causal Analytics Toolkit (CAT) to provide a data-driven approach to determine strong correlational relationships and avoid the application of biased models. Using an Excel add-in feature with R packages allows users to compute causal analysis without having to code the regression models (Cox, 2016). The CAT, in other words, is a bias-checking program that could notify the researchers of any unintentional researcher bias, similar to a grammar checking program for students.

The CAT integrates Bayesian network learning algorithms[1] with various visualization effects that could demonstrate the correlations between the variables in the dataset in an easy-to-use package. The main goal of the program is to identify strong correlations among data that *could* be causal. If two variables are not shown to be potentially causally related using the CAT, then the data itself do not support a confident causal inference or prediction. Such finding could indeed be used to dispute premature causal conclusions based on expert judgment (Cox, 2017).

The CAT applies non-parametric methods that would eliminate unverified modeling assumptions, which means the analysis is only driven by data without the researchers' any preconceived ideas or models that could bias the analysis results (Cox, 2017). Cox cited four challenges in causal analytics and how the CAT responds to these challenges:

- Ambiguous concentration-response associations: the CAT uses information and partial dependence to draw conclusions, not associations.
- Unverified modeling assumptions: the CAT uses non-parametric methods
- No gold standard for causation: the gold standard is identified by how much does the concentration improve prediction of responses?
- The uncertain credibility of causal conclusions: the CAT strives to provide strong credibility and objectivity of information-based conclusions (Cox, 2016).

Table 1. Examples of Researcher biases

Pro (Claim)	Con (Caveat)
"Epidemiological evidence is used to quantitatively relate PM2.5 exposures to the risk of early death. We find that UK combustion emissions cause 13,000 premature deaths in the UK per year, while an additional 6000 deaths in the UK are caused by non-UK European Union (EU) combustion emissions" (Yim and Barrett, 2012).	"[A]lthough particulate matter has been associated with premature mortality in other studies, a definitive cause-and-effect link has not yet been demonstrated" (NHS, 2012)
"[A]bout 80,000 premature mortalities [per year] would be avoided by lowering PM2.5 levels to 5 g/m3 nationwide" in the U.S. 2005 levels of PM2.5 caused about 130,000 premature mortalities per year among people over age 29, with a simulation-based 95% confidence interval of 51,000 to 200,000 (Fann et al., 2012).	"Analysis assumes a causal relationship between PM exposure and premature mortality based on strong epidemiological evidence… However, epidemiological evidence alone cannot establish this causal link" (EPA, 2011, Table 5-11).
"[D]ata on the impact of improved air quality on	"In their primary analyses, which were

[1] Cox wrote about the uncertainty in risk analysis models and identified ten methods to understanding them using two main strategies: "finding robust decisions what work acceptably well for many models (those in the uncertainty set); and an adaptive risk management using "well-designed and analyzed trial and error" (Cox, 1611). Of the ten methods, the Bayesian Model Averaging as one of the best developed models for statistical inference in cases when the statistical model is uncertain.

children's health are provided, including… the reduction in the rates of childhood asthma events during the 1996 Summer Olympics in Atlanta, Georgia, due to a reduction in local motor vehicle traffic" (Buka et al., 2006). "During the Olympic Games, the number of asthma acute care events decreased 41.6% (4.23 vs. 2.47 daily events) in the Georgia Medicaid claims file," coincident with significant reductions in ozone and other pollutants (Friedman et al., 2001).	adjusted for seasonal trends in air pollutant concentrations and health outcomes during the years before and after the Olympic Games, the investigators did not find significant reductions in the number of emergency department visits for respiratory or cardiovascular health outcomes in adults or children." In fact, "relative risk estimates for the longer time series were actually suggestive of increased ED [emergency department] visits during the Olympic Games" (Health Effects Institute, 2010)
"Our findings suggest that control of particulate air pollution in Dublin led to an immediate reduction in cardiovascular and respiratory deaths." (Clancy et al., 2002). "The results could not be more clear, reducing particulate air pollution reduces the number of respiratory and cardiovascular related deaths immediately" (Harvard School of Public Health, 2002).	"Serious epidemics and pronounced trends feign excess mortality previously attributed to heavy black smoke exposure" (Wittmaack, 2007)." "Thus, a causal link between the decline in mortality and the ban on coal sales cannot be established" (Pelucchi et al., 2009).

Daniel X. Kim
Cindy Huang
Capstone Assignment: Final Paper
May 08, 2017

3. THE CAT in Use

3.1 Program Testing with Deforestation v. International Trade

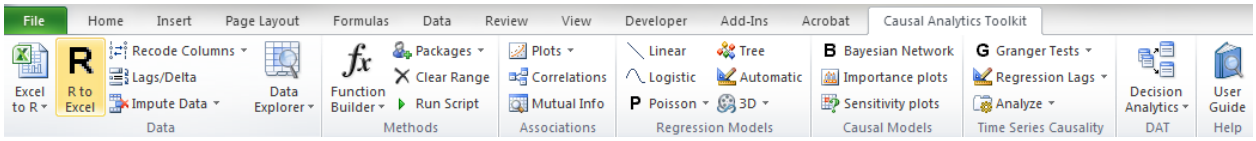
Examining the usefulness of the CAT, the team decided to examine the correlation between the international trade and deforestation. Prior studies have widely recognized that deforestation is affected by agricultural material trade in the international market (Robalino and Herrera, 2010). However, would increase in trade of agricultural raw material decrease a nation's Forest Area? There is a limited body of literature focusing on the direct relation between international trade and deforestation. Thus, we conducted an empirical analysis using the CAT to examine the causal relationship between deforestation and international trade.

3.2 Methodology

To measure international trade, several direct indicators were examined, including Trade (% of GDP), Agricultural raw materials imports (% of merchandise imports), Agricultural raw materials exports (% of merchandise exports). Several indirect indicators such as Net barter terms of trade index, net forest depletion (% of GNI) and 16 other variables were also examined to see if the CAT would indicate there are variables we might have overlooked. Since the CAT is embedded with some functions that can automatically choose appropriate regression models based on the data structure, no pre-specified models are required prior to analysis. Nevertheless, it is still necessary to identify the independent variables (i.e. predictors) that may help explain the variation in the dependent variable.

Given these data, the CAT's "Automatic" function is first used to automatically identify a regression model. As a result, it identifies a simple ordinary least square (OLS) model. However, an OLS model is typically used for linear correlation, which could generate biased results for this analysis because deforestation might be caused differently by different countries. Therefore, samples were divided into two different categories; the first is the multiple regression models among the Annex-I countries, and the other is the same model for non-Annex-I countries.

In addition to regression analysis, several causal analytics tools in the CAT are used to explore the relation between Forest Area and Trade. The "Correlations" function is used to demonstrate correlations between variables. "Granger Tests" is used to test whether values in a time series is predictable from prior values of another time series; in this case, whether values in Forest Area can be predicted by prior values of Trade. "Sensitivity Plots" is to quantify and visualize marginal effects of trade on Forest Area. Last, "Importance Plots" is used to illustrate the relative importance of each predictor in predicting the variation in the response.



3.3 Data

Data for this analysis are mostly obtained from public databases provided by the World Bank. The time period covered is from 1990 to 2015. The number of countries studied is six countries (of the 177 countries). The dependent variable, the Forest Area of each nation, was cross-checked with the database offered by Food and Agriculture Organization (FAO) of the United Nations. Table 2 summarizes all the variables.

Because the CAT is still in its developing phase, we confronted many difficulties to fit the dataset into a format that was *readable* by the program. Since the CAT did not cooperate with the missing data (i.e. blank cells in the database), we only succeeded in analyzing 21 variables (of the 150 variables we gathered) for each of the six countries (of the 177 countries) – three Annex-I countries and three non-Annex-I countries. We present the six countries' results in this report: United States, Japan, Australia (Annex-I) and Uganda, Nicaragua, and Myanmar (non-Annex-I).

Table 2. Variables

	Database	The CAT
Countries	177	6
Variables	150	21

3.4 Result

Table 3 shows the correlations among the four most important indicators for the deforestation[2]. Pearson correlations refer to the correlation between the variable without controlling for the effects of the other predictors, while partial correlations indicate the correlation between the variables after adjusting for all other predictor variables via linear regression. The results indicate that correlation between Forest Area and trade is relatively strong[3] for all six countries. There were interesting differences between the developed and developing countries regarding what variable are correlated with the Forest Area.

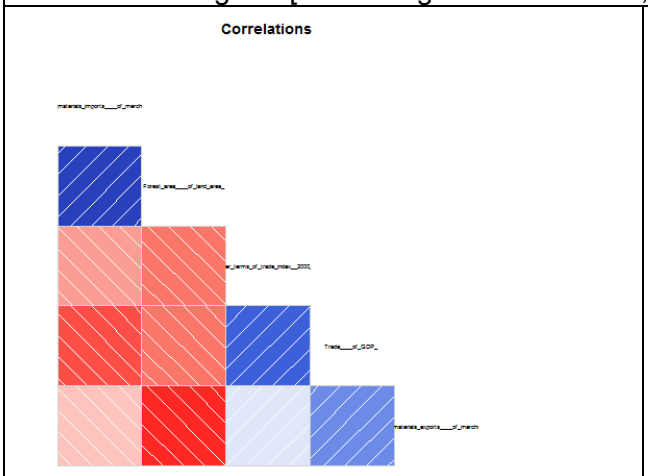
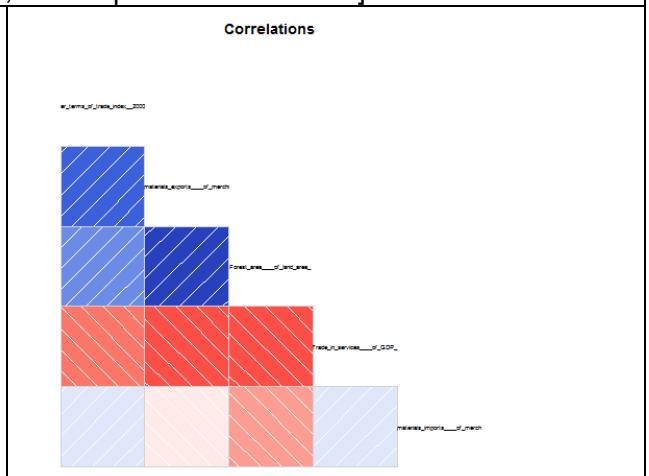
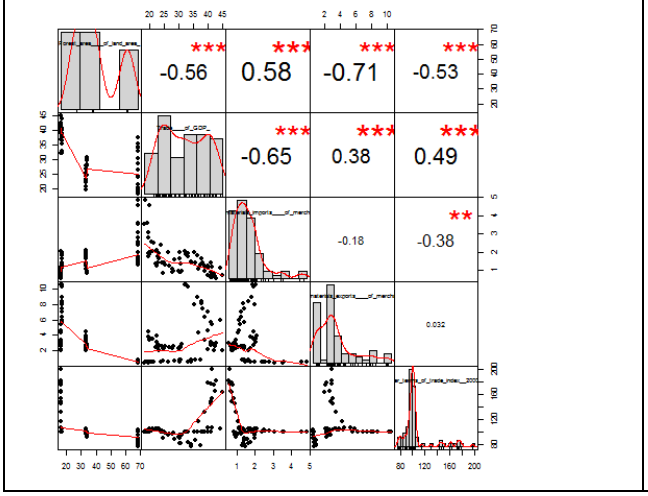
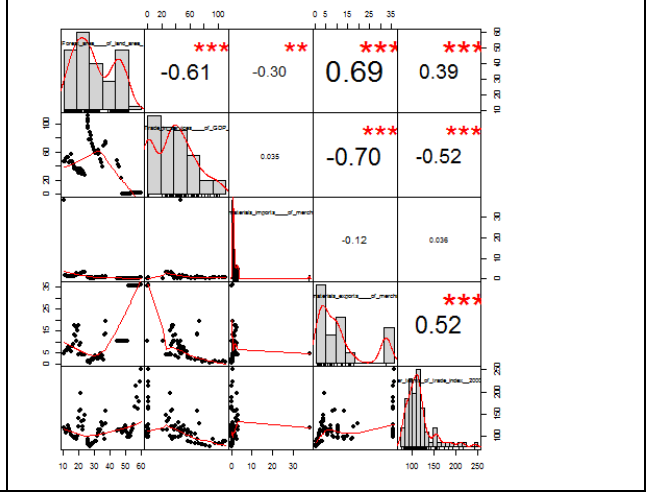
Table 3. Correlations between deforestation and trade

	Developed Countries (US, Japan, and Australia)		Developing Countries (Uganda, Nicaragua, and Myanmar)	
	Pearson Correlations	Partial Spearman Correlations	Pearson Correlations	Partial Spearman Correlations
	Forest Area	Forest Area	Forest Area	Forest Area
Trade(%GDP)	- 0.56	-0.50	- 0.61	0.12

[2] (the authors have results for all 21 variables, but for the limitation of space in this report, the authors decided to present only five)
[3] Weak if between .3 and -.3

Agricultural Import	0.58	0.17	-0.30	-0.86
Agricultural Exports	- 0.71	- 0.79	0.69	0.15
Net Barter Terms of Trade Index	- 0.53	- 0.89	0.39	0.53

Table 4. Visualizing Correlation

Developed Countries	Developing Countries
Correlation Diagram [Red = negative correlation, Blue = positive correlation]	
	
Scatterplot matrix, with histograms, kernel density overlays, absolute correlations, and significance asterisks (0.05, 0.01, 0.001)	
	

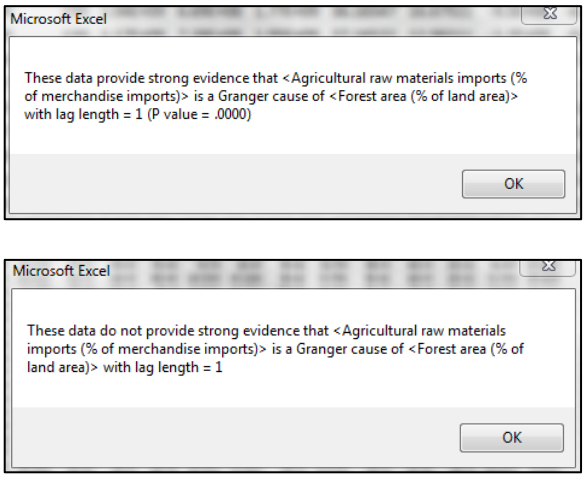
For developed countries, according to the Pearson correlations, only Agricultural Raw Material Imports (% of merchandise imports) indicates a positive correlation with the Forest Area, while

others all show a negative correlation. When controlling for the effects of all other predictors (partial Spearman correlations), all correlations stay approximately the same. The results indicate that an increase in Trade/Agricultural Exports/Net Barter terms of trade is likely associated with less Forest Area, while an increase in Agricultural Exports is *weakly* associated with more Forest Area, holding all other predictor variables constant.

For developing countries, while trade still seems to be negatively correlated with the Forest Area, the signs for import and export of agricultural raw materials are reverse compared to the developed countries' correlations. The export of Agricultural Raw Material is positively associated, while the import is negatively associated with Forest Area. The Net Barter Terms of Trade index is also positive for the developing countries. Such result indicates that an increase in the agricultural import is strongly associated with less Forest Area for developing countries.

According to Granger tests, Agricultural Raw Material Import variable is a Granger cause of Forest Area for developing countries, while it is NOT for developed countries. In other words, the value of Forest Area at a time point is predictable from the value of trade at a prior time point for developing countries. This does fulfill an essential condition for causality – a cause must precede its effect (only for developing countries).

Picture 1. Output Message for Granger test from the CAT



Regression analysis reveals more details on the relationship between trade and deforestation. For the developed countries, the variables with statistically significant p-values are Agricultural Raw Materials Export and Import. For the developing countries, on the other hand, shows many more variables that influence the Forest Area in the nation; Trade (% of GDP), Agricultural Raw Materials Exports, Net Barter Terms Of Trade Index, Transport Services of Service Imports, Travel Services of Service Imports, ICT Service Exports, Communications, Computer Exports, Insurance And Financial Services Exports, and Travel Services Exports. All of the statistically significant variables are negatively correlated except Agricultural Raw Materials Exports, Net

Barter Terms of Trade, Travel Services, and ICT Service. For some of these variables (e.g. ICT Service Exports), further research is needed to explain its relationship with the Forest Area.

Such difference in the statistically significant variables could be referring to the Kuznets Curve theory. For developing countries, the development relates to environmental degradation while for developed countries, the economic growth relates to environmental conservation. For developing countries, the multiple linear regression shows that many factors not only in the agricultural sector but also ICT, transport, and travel sectors are significantly correlated with less Forest Area.

More specifically, negative correlation between Forest Area and Trade in developing countries could be described as (1) a 1 percentage point increase in Trade (% of GDP) is associated with an estimated 0.140 percentage point decrease in Forest Area (% land area); (2) a 1 percentage point increase in Agricultural Raw Materials Exports (% of merchandise exports) is associated with an estimated 0.212 increase in Forest Area (% land area); (3) a 1 point increase in Net Barter terms of trade index is associated with an estimated 0.0493 percentage point increase in Forest Area (% land area); (4) a percentage point increase in Transport Services is associated with an estimated 0.0831 percentage point decrease in Forest Area (% land area); (5) a percentage point increase in Travel Services Imports is associated with 0.213 percentage point increase in Forest Area (% land area); (6) a percentage point increase in ICT Service Export is associated with an estimated 0.244 percentage point increase in Forest Area (% land area); (7) a percentage point increase in Communications Computer Export is associated with 0.222 percentage point decrease in Forest Area (% land area); (8) a percentage point increase in Travel Services Exports is associated with 0.443 percentage point decrease in Forest Area (% land area).

To further visualize how Forest Area changes with Trade, sensitivity plots are used to show the partial dependence of Forest Area on Trade (% of GDP) – the marginal effect of Trade on Forest Area after accounting for the average effects of all other predictors in the model. The results are parallel with what we have found in the correlation analysis above. For developed countries, exports are negatively correlated, and imports are insignificant. For developing countries, the relationship is reversed, where exports are positive and imports are negative.

According to the above analysis, there seem to be differing influences of Agricultural Goods Import and Export in developed and developing countries. To illustrate the relative importance among all the predictors in predicting the response, importance plots are used. Table 5 highlights Insurance and Financial Services, Trade, Agricultural Imports And Exports as the most important predictors of Forest Area for developing countries. For developed countries, the importance table showed no significant results. Overall, trade is an important indicator for the nation's Forest Area, but for developing countries, the effect is larger and varies across more variables than developed countries.

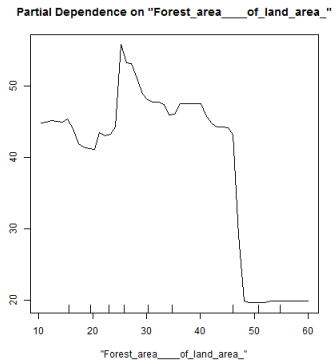
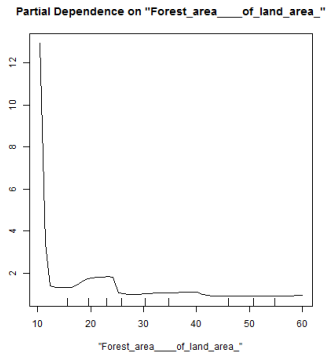
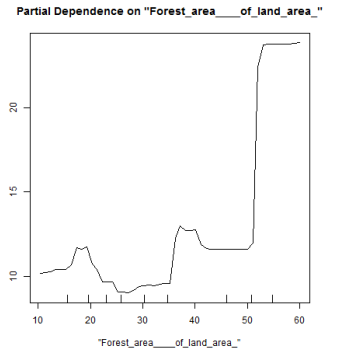
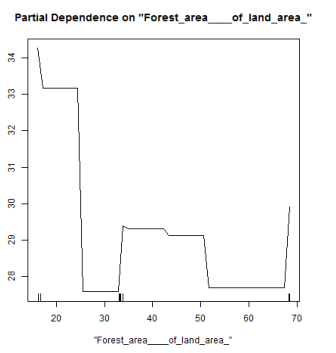
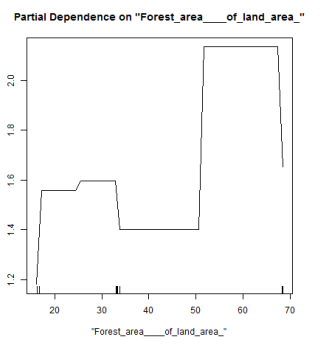
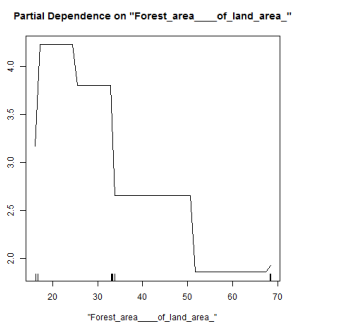
In other words, the CAT was able to redirect the capstone team's focus from only Agricultural Good Imports and Exports to other variables such as ICT Service Exports. After running the

multivariable regression, the team found that many more variables can influence the Forest Area, especially for the developing countries. Seeing that there are many more variables that influence the Forest Area for developing countries, we suggest looking more deeply into the variables that might indicate the effect of Kuznets curve for further research.

Table 5: Regression Results for Forest Area (% Land area)		
Response	Developed Countries	Developing Countries
	Linear	Linear
Direct Predictors		
Trade (% of GDP)	-1.23e-02 -0.10	-1.40e-01 -3.69***
Agricultural imports (% of merchandise imports)	1.36 (1.25)	-7.15e-02 -0.66
Agricultural imports (% of merchandise imports)	-2.21 (-5.05)***	2.12e-01 2.54*
Net Barter terms of trade index_2000	-0.00882 (-0.26)	4.93e-02 2.49*
Other Predictors		
Goods imports (BoP)	3.47e-10 0.01	1.99e-08 0.09
Service imports (BoP)	4.00e-10 0.01	-2.02e-09 -0.19
Imports of goods services and primary income (BoP)	-4.22e-11 -2.30)*	1.90e-09 0.73
Transport Services (% of service imports)	-2.23e-01 -1.13	-8.31e-02 -2.04*
Travel services (% of service imports)	-2.43e-01 -1.44	2.13e-01 2.53*
Net Trade in goods and services (BoP)	3.05e-10 0.01	1.14e-08 1.04
Net Trade in goods (BoP)	1.35e-11 0.31	9.88e-09 0.05
ICT services exports (BoP)	-2.18e-10 -1.16	8.44e-09 0.73
ICT Service Exports (% of service exports)	-4.10e-02 -0.12	2.44e-01 2.56*
Communithe CATions computer (% of service exports)	2.95e+00 0.21	-2.22e-01 -3.35**
Exports of goods and services (BoP)	-3.39e-10 -0.01	-2.20e-08 -0.10
Insurance and financial services (% of service exports)	2.83e+00 0.20	-2.80e+00 -3.69***
Goods exports (BoP)	NA	NA
Service exports (BoP)	NA	1.08e-08

		0.05
Exports of goods in services and primary income (BoP)	3.68e-11 1.78	1.96e-09 0.10
Transport Services (% of service exports)	3.11e+00 0.22	-3.13e-01 -1.93
Travel services (% of service exports)	2.39e+00 0.17	-4.43e-01 -5.23***
Adj. R^2	2.01	0.014
*Note: t-statistics in the parentheses. *** P<0.001; ** P<0.01; * P<0.05.*		

Table 6. Sensitivity Reports from the CAT

	Trade	Agricultural Import	Agricultural Export
Developing			
Developed			

Daniel X. Kim
Cindy Huang
Capstone Assignment: Final Paper
May 08, 2017

Table 7. Importance Plots: Developing Countries

Importance table

From most to least important, the relative importances of these potential causes are as follows:

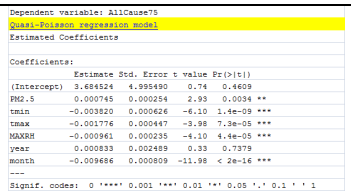
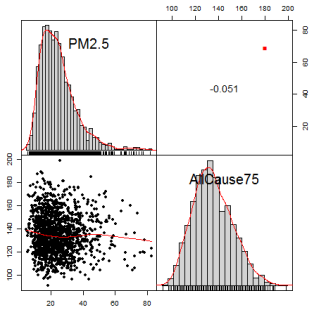
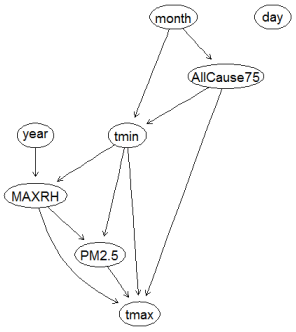
Variable	Importance(%IncMSE)
Insurance_and_financial_services____of_service_exports__BoP_	107.509293
Trade____of_GDP_	94.597787
Agricultural_raw_materials_imports____of_merchandise_imports_	79.391100
Agricultural_raw_materials_exports____of_merchandise_exports__	27.218501
Travel_services____of_service_exports__BoP_	15.346059
ICT_service_exports____of_service_exports__BoP_	14.294970
Net_trade_in_goods_and_services__BoP__current_US__	13.135911
Travel_services____of_service_imports__BoP_	11.351067
Communications__computer__etc.____of_service_exports__BoP_	9.687127
ICT_service_exports__BoP__current_US__	9.189102
Service_imports__BoP__current_US__	7.547342
Net_barter_terms_of_trade_index__2000___100_	5.370496
Net_trade_in_goods__BoP__current_US__	3.867676
Goods_exports__BoP__current_US__	3.403445
Exports_of_goods_and_services__BoP__current_US__	3.352868
Exports_of_goods__services_and_primary_income__BoP__current_US__	3.015171
Imports_of_goods__services_and_primary_income__BoP__current_US__	2.694213
Service_exports__BoP__current_US__	2.347141
Transport_services____of_service_exports__BoP_	2.044496
Goods_imports__BoP__current_US__	2.039379
Transport_services____of_service_imports__BoP_	1.079865

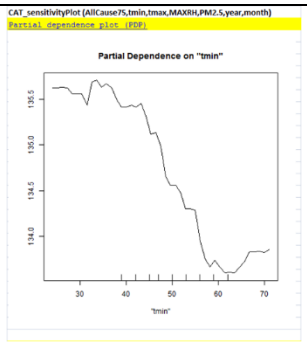
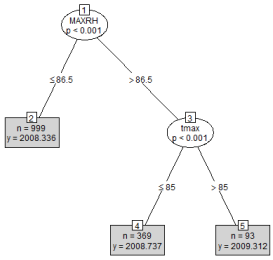
Daniel X. Kim
Cindy Huang
Capstone Assignment: Final Paper
May 08, 2017

4. The CAT Functions

4.1 Program Description and Functions

The CAT software offers various causal techniques without the need to input coding for an R script to produce the results.

Automatic		The **Automatic** icon will identify the appropriate families of regression models, apply the model to the data, and display the results (Cox, 5). The user can scroll down the spreadsheet to view the various models provided in the analysis.
Plots		The **Plots** function provides frequency distributions, scatter plots, correlation, smooth regression curves. The user can see how negatively or positively variable A is associated with variable B.
B Bayesian Network		The **Bayesian Network** function produces a graphical output of a Bayesian Network analysis by using a directed acyclic graph (DAG) structure. Each node contains a conditional probability take and variables that indicate a relationship will be depicted through arrows. The CAT provides further analysis on the discovery algorithm by being able to confirm or refute the Bayesian Network structure with additional non-parametric tests.

Sensitivity plots		The **Sensitivity** plots estimate how one variable varies with another while holding all others fixed at their empirical frequency distribution of values ("Partial Dependence Plot"). The estimate is created by averaging predictions from ensemble ("forest") of trees. The function allows nonlinearities and interactions.
Tree		The **Tree** function is one of the primary features of the CAT. It is a Classification and Regression Tree (CART) that uses the data to identify relationships in a set of data. This algorithm reduces prediction error for the dependent variable. This feature can be useful to confirm or identify a relationship that the user may not believe to exist in a set of data. The CAT also includes basic statistic functions.

4.2 User Experience

Familiarity with managing datasets in Microsoft Excel will improve the user's ability to use the tool. The CAT, along with the R software and packages, is installed as an extension to Excel. Because the CAT application also includes the R software and packages, it is not required to download the R software prior to use. However, the research team discovered that installing the latest version of R facilitated the installation of the CAT program.

Prior to using the dataset in the CAT program, the user must prepare the dataset so that the data can be read by the CAT functions. Once the dataset is ready, the user can manage the dataset in Excel as normal but with the added ability to apply CAT functions.

The CAT provides easy to use functions to create graphs, make test correlations, and run regressions. Prior knowledge of a model is not required before using CAT because the user can choose an "Automatic" function in CAT to suggest a regression model, which chooses an appropriate model based on the structure of the dataset. It is important for the user to understand that although the automatic feature in CAT provides model options for the user, the results does not identify logarithms, polynomials, or interactions in the results. This is due to the program providing limited types of regression models, which includes linear, logistic, and Poisson (quasi-Poisson) models. In order to identify those functional forms, the user needs to take an extra step and use CAT functions such as CAT_tree or CAT_sensitivity functions.

The CAT provides the user with functions to explore correlations among variables. The Bayesian Network is one feature provides a visualized network of correlated variables. Through this visualization, the user can conduct an analysis to identify direct and indirect causes of the variables and test the validity by filtering out misleading connections using linear regressions or another CAT function, important plots. This function visualizes the relative importance of all the predictor variables in predicting the response.

The CAT fully integrates R capabilities in Excel so that the user does not require knowledge of R scripts to conduct analyses, but allows users who know R scripts to input them into Excel cells. The functions available in the program is limited to the R packages that is provided upon installation so if the user requires analyses beyond the installed packages, the user must know the R code to execute the relevant R package (Xie, interview). One example of the limitation of the current version of CAT is the function to test the assumptions of generalized least squares (e.g. normality and homoscedasticity).

In July 2016, the CAT introduced an option for users to use the CAT with the Python programming software with a plug-in. However, the research team was not able to explore the features in the Python plug-in due to the errors that it caused upon installation of the CAT program.

The CAT has installed the most commonly used statistical functions into the one-click interface. However, we feel that some of these functions are more commonly used in general statistics rather than econometrics. For example, machine learning algorithms such as random forest have not been widely used in econometrics. Therefore, to what extent the CAT can be applied to policy analysis needs to be further explored.

Although one of the goals of the CAT program is to provide an easy-to-use interface to conduct causal analysis for users who do not possess familiarity in statistics, the user does require a background in statistics in order to interpret the results from using the CAT functions. In the feasibility of using the program, the CAT achieves its target through introductory descriptions of some functions of the CAT in the User Guide and slides from previous presentations on the CAT. However, a non-statistician may inadvertently make mistakes reading results from the CAT and lead to misinterpretation of the results. So while deep statistical knowledge may not be required to use the CAT, knowledge of statistics is still required to understand the results produced by the program.

Table 8. Functions of the CAT: SWOT Analysis		
	Positive	Negative
Internal Validity	Strength	Weakness
	▪ Accessible by wider consumer who might not have extensive statistical background ▪ Solely data-driven without user bias in models ▪ Various visualization effects allow user to understand the dataset from multiple viewpoints	▪ Difficulties in installing: downloading, installing, and communicating with the developers ▪ Difficulties in getting the CAT to function: takes a long process of trial-and-error to make the CAT show results ▪ Not compatible with various computer configurations: Apple v. Windows
External Validity	Opportunity	Threats
	▪ Could be used to locate the correlations that need further investigation ▪ Could be used to check the data analysis to find any overlooked variables ▪ Potential to be developed into a more effective tool with features found in other similar software (e.g. GIS) ▪ Network?	▪ Misinterpretation of results by users that might not have extensive knowledge in statistics ▪ Might cause other types of biases for data analysis ▪ Still need to understand statistics in order to analyze results and use sophisticated functions of the CAT ▪ Competition with other programs

User Tips

- When installing the CAT to your computer, follow exactly with the User Guide, and do NOT uncheck any boxes in the installation windows
- Before executing any function in the CAT, remember to export all your data columns to R. When exporting, the datasheet has to be named “Data,” or it will not work
- The current version of the CAT does not support missing data, so all observations (rows) with missing values must be eliminated. The CAT has a Clean Rows function to do that
- Always select the dependent variable first when you run regressions or other analysis
- The result tab will be overwritten every time you re-click on the same function, so make sure to rename the tab if you want to save the result

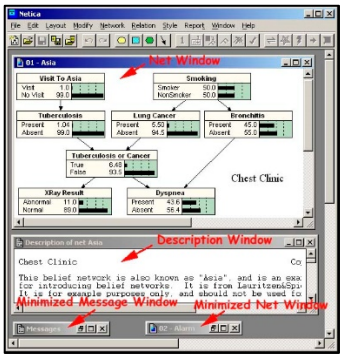
Daniel X. Kim
Cindy Huang
Capstone Assignment: Final Paper
May 08, 2017

5. Market Research

5.1 Competitors

There are other programs available in the market that takes approaches to causal analytics similar to the CAT. There are three programs that we have found that could be the CAT's competition; Netica, Bayesia Lab, and TETRAD.

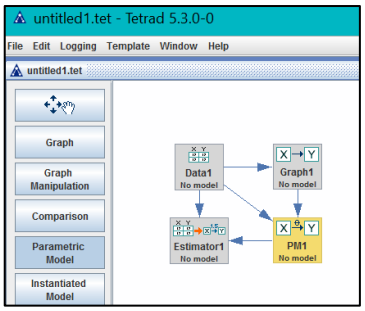
5.1.1 Netica



Netica is a Bayesian Network software that automates probability propagation. Once the variables are entered, the program finds the most likely explanation for the observations. The program was developed by Norsys Software Corp., a company that provides services that specialize in Bayesian network software to companies and government organizations (foreign and domestic) in almost every sector. The program uses belief networks to perform various kinds of inference through algorithms and is compatible with MS Windows and MAC Operating Systems. Netica provides a free version with full features of the program but limited in model size. The full version costs vary depending on the type of use, whether it's educational, personal, commercial, or site license. The program's most recent update includes a new add-on product called GeoNetica that integrates Geographic Information System (GIS) capability to Netica in response to the demand of clients who used Netica in conjunction with a GIS platform. Upcoming features for GeoNetica to improve usability includes learning Bayes net models from GIS data, operating on vector data and raster data, and including an interface with commercially available GIS systems. Another feature of Netica is its variations in the Application Programmer Interfaces (API) group that includes Java, C, C#, Visual Basic/COM, C++, Matlab, and CLisp.

5.1.2 TETRAD

TETRAD is a program that aims to provide sophisticated causal and statistical models in a user-friendly interface for operators with little to no statistical or programming knowledge. The program is limited to models of categorical data and to linear models with a normal probability distribution. The output is three models that describe causal data:

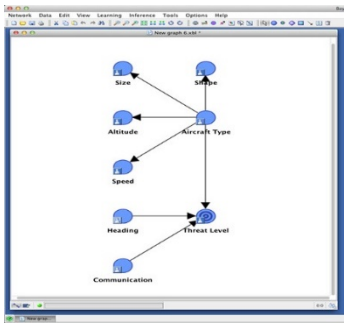


- A directed graph specifying hypothetical causal relations among the variables;
- A specification of the family of probability distributions and kinds of parameters associated with the graphical model; and
- A specification of the numerical values of those parameters.

The program is not intended to replace statistical programming systems, but is a freeware that performs many functions in commercial products such as Netica, Hugin, LISREL, EQS and performs discovery functions that the products do not perform.

5.1.3 Bayesia Lab

BayesiaLab is a desktop application with a graphical user interface that fosters a "'laboratory' environment for machine learning, knowledge modeling, diagnosis, analysis, simulation, and optimization" (BayesiaLab). Like Netica's API application, users can use Bayesia Engine APIs to construct and edit networks. In order to download a free trial of BayesiaLab and the user guide, the user must provide an email and organization status where an email will be sent to download the program. The free trial is available only for 90 days. Immediately upon signing up, the user receives regular email notifications regarding upcoming seminars to better understand about uncertainty, Bayesian Networks, and about BayesiaLab.



There is no requirement to know any scripts prior to using the CAT, unlike Netica and BayesiaLab. In the R product and Python program, the code is already incorporated because of the integration of the CAT with the two freeware programs. However, if the other software programs perform causal analysis better than the CAT and with no errors while downloading or in use, then the competition will become a better option for consumers to use. Although expanding the customer base with the ability to use the program with other operating systems encourage use, it may not be necessary to develop the compatibility for other operating systems since most policymakers in the U.S. government use a Windows operating system on their computers.