Directed Acyclic Graphs: An Application to Modeling Causal Relationships with Worldwide Poverty Data

Gott würfelt nicht. -- Albert Einstein

David A. Bessler
Texas A&M University

Presented to
James S. McDonnell Foundation
21st Century Science Initiative
“Creating Knowledge from Information”
Tarrytown, New York
June 3, 2003
Food and Agricultural Organization (FAO) of the United Nations

- FAO has the charge to understand the role of food production in poverty alleviation.
- The development literature has identified several variables as being related to poverty.
- The causal status of many of these variables is unsettled.
- It is unethical to perform random assignment experiments to provide evidence on the causal status of these variables.
A Partial List of Literature on Causes and Effects of Poverty

- Agricultural Income (Mellor 2000)
- Freedom (Sachs and Warner 1997)
- Income (Sen 1981)
- Income Inequality (Sen 1981)
- Child Mortality (Berhrman and Deolalikar 1988)
Literature Continued

- Birth Rate (Malthus 1798; Sen 1981)
- Rural Population (Rosenweig 1988)
- Foreign Aid (World Bank 2000)
- Life Expectancy (Wheeler 1980)
- Illiteracy (Birdsall 1988)
- International Trade (Ricardo 1817; Bhagwati 1996)
Measures of Poverty


- Economic Measures: e.g., % of Population Living on One or Two Dollars or Less per Day
- Biological Measures: e.g., deficits in calorie intake
Data Sources

World Bank Development Indicators
80 Countries: % of Population Living off of One and Two Dollars per Day or Less.

Heritage Foundation
Index of Economic and Political Freedom on 80 countries.

FAO (United Nations)
% of Population that is Under-Nourished.
### Table 1  
**Countries Studied**

<table>
<thead>
<tr>
<th>Algeria</th>
<th>Burkino</th>
<th>Ecuador</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armenia</td>
<td>C. Afr. Rep.</td>
<td>Egypt</td>
</tr>
<tr>
<td>Azerbaijan</td>
<td>Chile</td>
<td>El Salvador</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>China</td>
<td>Estonia</td>
</tr>
<tr>
<td>Belarus</td>
<td>Columbia</td>
<td>Ethiopia</td>
</tr>
<tr>
<td>Bolivia</td>
<td>Costa Rica</td>
<td>Gambia</td>
</tr>
<tr>
<td>Botswana</td>
<td>Cote Divor</td>
<td>Georgia</td>
</tr>
<tr>
<td>Brazil</td>
<td>Czech</td>
<td>Ghana</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>Domin Rep</td>
<td>Guatemala</td>
</tr>
</tbody>
</table>

Countries listed in this table were selected from 2001 World Bank Development Indicators for which $1/day and $2/day population figures were available.
### Table 1
**Countries Studied Continued**

<table>
<thead>
<tr>
<th>Country</th>
<th>Country</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honduras</td>
<td>Lao Pdr</td>
<td>Mongolia</td>
</tr>
<tr>
<td>Hungry</td>
<td>Latvia</td>
<td>Morocco</td>
</tr>
<tr>
<td>India</td>
<td>Lesotho</td>
<td>Mozambique</td>
</tr>
<tr>
<td>Indonesia</td>
<td>Lithuania</td>
<td>Namibia</td>
</tr>
<tr>
<td>Jamaica</td>
<td>Madagascar</td>
<td>Nepal</td>
</tr>
<tr>
<td>Jordan</td>
<td>Mali</td>
<td>Niger</td>
</tr>
<tr>
<td>Kazakhstan</td>
<td>Mauritania</td>
<td>Nigeria</td>
</tr>
<tr>
<td>Kenya</td>
<td>Mexico</td>
<td>Pakistan</td>
</tr>
<tr>
<td>Korea S.</td>
<td>Moldova</td>
<td>Panama</td>
</tr>
</tbody>
</table>

Countries listed in this table were selected from 2001 World Bank Development Indicators for which $1/day and $2/day population figures were available.
Table 1
Countries Studied  Continued

<table>
<thead>
<tr>
<th>Paraguay</th>
<th>Slovak Rep</th>
<th>Turkmenstan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peru</td>
<td>Slovenia</td>
<td>Ukraine</td>
</tr>
<tr>
<td>Poland</td>
<td>So Africa</td>
<td>Uruguay</td>
</tr>
<tr>
<td>Portugal</td>
<td>Sri Lanka</td>
<td>Uzbekistan</td>
</tr>
<tr>
<td>Romania</td>
<td>Tanzania</td>
<td>Venzuela</td>
</tr>
<tr>
<td>Russia</td>
<td>Thailand</td>
<td>Yemen</td>
</tr>
<tr>
<td>Rwanda</td>
<td>Trinidad</td>
<td>Zambia</td>
</tr>
<tr>
<td>Senegal</td>
<td>Tunisia</td>
<td>Zimbabwe</td>
</tr>
<tr>
<td>Sierra Leon</td>
<td>Turkey</td>
<td></td>
</tr>
</tbody>
</table>

Countries listed in this table were selected from 2001 World Bank Development Indicators for which $1/day and $2/day population figures were available.
Inference on Causal Flows

- Oftentimes we are uncertain about which variables are causal in a modeling effort.
- Theory may tell us what our fundamental causal variables are in a controlled system.
- It is common that our data may not be collected in a controlled environment.
Use of Subject Matter Theory

Theory may be a good source of information about direction of causal flow among variables. However, theory usually invokes the *ceteris paribus* condition to achieve results.

Data are often observational (non-experimental) and thus the *ceteris paribus* condition may not hold. We may not ever know if it holds because of unknown variables operating on our system.
Experimental Methods

- If we do not know the "true" system, but have an idea that one or more variables operate on that system, then experimental methods can yield appropriate results.

- Experimental methods work because they use randomization, random assignment of subjects to alternative treatments, to account for any additional variation associated with the unknown variables on the system.
Observational Data

In the case where no experimental control is used in the generation of our data, such data are said to be observational (non-experimental).
Causal Models Are Well-Represented By Directed Graphs

One reason for studying causal models, represented here as $X \rightarrow Y$, is to predict the consequences of changing the effect variable ($Y$) by changing the cause variable ($X$). The possibility of manipulating $Y$ by way of manipulating $X$ is at the heart of causation.

“Causation seems connected to intervention and manipulation: one can use causes to ‘wiggle’ their effects.”

-- Hausman (1998, page 7)
Directed Acyclic Graphs

- Pictures summarizing the causal flow among variables -- there are no cycles.

- Inference on causation is informed by asymmetries among causal chains, causal forks, and causal inverted forks.
A Causal Fork

For three variables X, Y, and Z, we illustrate X causes Y and Z as:

\[ Y \leftarrow X \rightarrow Z \]

Here the unconditional association between Y and Z is non-zero, but the conditional association between Y and Z, given knowledge of the common cause X, is zero.

*Knowledge of a common cause screens off association between its joint effects.*
An Example of a Causal Fork

- X is the event, the student doesn’t learn the material in Econ 629.
- Y is the event, the student receives a grade of “D” in Econ 629.
- Z is the event, the student fails the PhD prelim in Economic Theory.

Grades are helpful in forecasting whether a student passes his/her prelims: \( P(Z \mid Y) > P(Z) \)

If we add the information on whether he/she understands the material, the contribution of grade disappears (we do not know candidate’s name when we mark his prelim): \( P(Z \mid Y, X) = P(Z \mid X) \)
An Inverted Fork

- Illustrate X and Z cause Y as:
  
  \[ \begin{align*}
  X & \rightarrow Y \\
  Y & \leftarrow Z
  \end{align*} \]

- Here the unconditional association between X and Z is zero, but the conditional association between X and Z, given the common effect Y is non-zero:

\[ \text{Knowledge of a common effect does not screen off the association between its joint causes.} \]
The Causal Inverted Fork: An Example

- Let Y be the event that my daughter’s cell-phone won’t work
- Let X be the event that she did not pay her phone bill
- Let Z be the event that her battery is dead

Paying the phone bill and the battery being dead are independent: \( P(X|Z) = P(X) \).

Given I know her battery is dead (she remembers that she did not charge it for a week) gives some information about bill status: \( P(X|Y,Z) < P(X|Y) \). (although I don’t know her bill status for sure).

\[ X \rightarrow Y \leftarrow Z \]
The Literature on Such Causal Structures Has Been Advanced in the Last Decade Under the Label of Artificial Intelligence

- Pearl, *Biometrika*, 1995
- Glymour and Cooper, editors, *Computation, Causation and Discovery*, MIT Press, 1999
Causal Inference Engine
- PC Algorithm

1. Form a complete undirected graph connecting every variable with all other variables.

2. Remove edges through tests of zero correlation and partial correlation.

3. Direct edges which remain after all possible tests of conditional correlation.

4. Use screening-off characteristics to accomplish edge direction.
Assumptions
(for PC algorithm on observational data to give same causal model as a random assignment experiment)

1. Causal Sufficiency
2. Causal Markov Condition
3. Faithfulness
4. Normality
Causal Sufficiency

No two included variables are caused by a common omitted variable.

No hidden variables that cause two included variables.
Causal Markov Condition

The data on our variables are generated by a Markov property, which says we need only condition on parents:

\[ P(W, X, Y, Z) = P(W) \cdot P(X|W) \cdot P(Y) \cdot P(Z|X,Y) \]
There are no cancellations of parameters.

\[ A = b_1 B + b_3 C \]
\[ C = b_2 B \]

It is not the case that: \[-b_2 b_3 = b_1\]

Deep parameters \(b_1\), \(b_2\) and \(b_3\) do not form combinations that cancel each other.
Table 2
Examples of Edges Removed

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gini -- Ag Inc</td>
<td>rho(Gini, Ag Inc)</td>
<td>-0.1266</td>
<td>0.2612</td>
</tr>
<tr>
<td>Gini -- Life Exp</td>
<td>rho(Gini, Life Exp)</td>
<td>-0.0920</td>
<td>0.4157</td>
</tr>
<tr>
<td>Gini -- % Rural</td>
<td>rho(Gini, % Rural)</td>
<td>-0.0298</td>
<td>0.7921</td>
</tr>
<tr>
<td>Gini -- Child Mort</td>
<td>rho(Gini, Child Mort)</td>
<td>0.1103</td>
<td>0.3283</td>
</tr>
<tr>
<td>Gini -- GDP/Person</td>
<td>rho(Gini, GDP/Person)</td>
<td>-0.0416</td>
<td>0.7131</td>
</tr>
<tr>
<td>Gini -- Illiteracy</td>
<td>rho(Gini, Illiteracy)</td>
<td>0.0709</td>
<td>0.5315</td>
</tr>
<tr>
<td>Gini -- Foreign Aid</td>
<td>rho(Gini, Foreign Aid)</td>
<td>0.0829</td>
<td>0.4637</td>
</tr>
<tr>
<td>Gini -- Under-Nourish</td>
<td>rho(Gini, Under-Nourish.)</td>
<td>-.1736</td>
<td>0.1222</td>
</tr>
<tr>
<td>Life Exp -- Birth Rate</td>
<td>rho(Life Exp, Birthrate</td>
<td>Child Mort)</td>
<td>0.0093</td>
</tr>
<tr>
<td>Life Exp -- Illiteracy</td>
<td>rho(Life Exp, Illiteracy</td>
<td>Child Mort)</td>
<td>0.0312</td>
</tr>
<tr>
<td>&lt;$2/Day -- Life Exp</td>
<td>rho(&lt;$2/Day, Life Exp</td>
<td>Child Mort)</td>
<td>-0.1199</td>
</tr>
</tbody>
</table>
### Table 2
A few more Removed Edges

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ag Inc -- Child Mort</td>
<td>rho(Ag Inc, Child Mort</td>
<td>Birthrate)</td>
<td>-0.0234</td>
</tr>
<tr>
<td>Ag Inc -- % Undr-Nourished</td>
<td>rho(Ag Inc, % Undr-Nourished</td>
<td>Birthrate)</td>
<td>-0.1202</td>
</tr>
<tr>
<td>Ag Inc -- % Rural</td>
<td>rho(Ag Inc, % Rural</td>
<td>&lt;$2/Day)</td>
<td>-0.0319</td>
</tr>
<tr>
<td>GDP/Person -- Foreign Aid</td>
<td>rho(GDP/Person, Foreign Aid</td>
<td>Child Mort)</td>
<td>-0.1096</td>
</tr>
<tr>
<td>Unfree -- Ag Inc</td>
<td>rho(Unfree, Ag Inc</td>
<td>Illiteracy)</td>
<td>-0.1368</td>
</tr>
<tr>
<td>Ag Inc – Foreign Aid</td>
<td>rho(Ag Inc, Foreign Aid</td>
<td>Child Mort)</td>
<td>-0.0529</td>
</tr>
<tr>
<td>Unfree -- Foreign Aid</td>
<td>rho(Unfree, Foreign Aid</td>
<td>Child Mort)</td>
<td>-0.0546</td>
</tr>
<tr>
<td>Gini -- %Undr-Nourished</td>
<td>rho(Gini, % Undr-Nourished</td>
<td>Child Mort)</td>
<td>0.1357</td>
</tr>
<tr>
<td>&lt;$2/Day -- Gini</td>
<td>rho(&lt;$2/Day, Gini</td>
<td>% Undr-Nourised)</td>
<td>0.1075</td>
</tr>
<tr>
<td>Unfree -- Birthrate</td>
<td>rho(Unfree, Birthrate</td>
<td>Child Mort)</td>
<td>-0.0754</td>
</tr>
<tr>
<td>% Rural -- Foreign Aid</td>
<td>rho(% Rural, Foreign Aid</td>
<td>&lt;$2/Day)</td>
<td>0.0313</td>
</tr>
</tbody>
</table>
Continue to Remove Edges Considering All Possible Conditional Correlations

- Significance level for removal is crucial (here we use 10% significance).

- Advanced methods for edge removal based on Statistical Loss functions and on Bayesian Posterior odds is currently being explored.
“Rising Tide Lifts All Boats?”
Regressions Based on $1/day Graph

\[
\% \text{ $1/Day} = 27.45 - .004 \text{ GDP/Per.} ; \quad R^2 = .60 \\
(2.65) \quad (.001)
\]

(estimated std. errors in parentheses)

Here regressing % $1/day on GDP/Person gives us the expected negative and significant estimate.

Recall from the graph, however, that no line connects GDP and $1/day. We removed the edge by conditioning on Child Mortality.

\[
\% \text{ $1/Day} = 2.75 - .0004 \text{ GDP/Per.} + .237 \text{ Chld Mrt} ; \quad R^2 = .84 \\
(2.82) \quad (.001) \quad (.022)
\]

This last regression shows GDP/Per is not significant in the $1/day regression.
“Rising Tide Lifts All Boats?”
Regressions Based on $2/day Graph

\[
\% \text{ $2/Day} = 57.96 - .007 \text{ GDP/Person} ; \quad R^2 = .81
\]
\[(3.39) \quad (.001)\]

Here regressing \% $2/day on GDP/Person gives us the expected negative and significant estimate!

Notice from the $2/day graph that we have a connection between GDP and $2/day. So conditioning on Child Mortality does not eliminate GDP as an actor in explaining \%$2/day.

\[
\% \text{ $2/Day} = 28.42 - .0033 \text{ GDP/Person} + .287 \text{ Child Mort} ; \quad R^2 = .91
\]
\[(4.22) \quad (.001) \quad (.034)\]
Regression Analysis: Backdoor and Front Door Paths

The previous results on the “rising tide” debate are generalized as necessary conditions for estimating the magnitude of the effect of a causal variable with regression analysis.

To estimate the effect of X on Y using regression analysis, one must block any “backdoor path” from X to Y via the ancestors of X. We “block” backdoor paths by conditioning on one or more ancestors of X.

To estimate the effect of X on Y using regression analysis one must not condition on descendants of X. One must “not block” the front door path.
Front Door Path: Consider the Effect of Agricultural Income on % < $2/day

From above we have the following causal chain:

Ag Income/Person $\rightarrow$ GDP/Person $\rightarrow$ %2/Day

Since GDP/Person is caused by AG Income/Person, we cannot have GDP/Person in the regression equation to measure the effect of Agricultural Income/Person on %2/Day – do not block the front door!

**Biased Regression (biased in terms of the coefficient on Ag. Inc.)**

\[
%2/Day = 57.99 - .0007 \text{ Ag Inc.} - .0068 \text{ GDP} ; \quad R^2 = .37
\]

\[
(3.60) \quad (.0014) \quad (.0018)
\]

**Unbiased Regression:**

\[
%2/Day = -51.73 - .0038 \text{ Ag Inc.} ; \quad R^2 = .23
\]

\[
(4.34) \quad (.0018)
\]
Backdoor Paths: Consider the Effect of Child Mortality on Poverty (%<$2/Day)

We have the following sub-graph:

Child Mortality ← Illiteracy Rate

↓

Birth Rate → %$2/Day

The front door path would suggest that we regress $2/Day on Child Mortality. But there exists a backdoor path, through Illiteracy Rate to $2/Day. **We must “block” the backdoor path** by conditioning on Illiteracy Rate.

Note: the edge between Illiteracy Rate and Child Mortality is directed using advanced loss function scoring methods (Illiteracy → Child Mort.).
Comparison of $2/Day on Child Mortality: Two Regressions

Biased Regression (fails to block the backdoor)

$2/Day = 17.85 + .339 \text{ Child Mort.} ; \quad R^2 = .65
\quad (2.92) \quad (.032)

Unbiased Regression (blocks the backdoor)

$2/Day = 16.91 + .265 \text{ Child Mort.} + .25 \text{ Illiteracy Rate} ; \quad R^2 = .66
\quad (2.71) \quad (.06) \quad (.16)

Caution: Do not interpret the estimated coefficient on illiteracy as unbiased. We violate the front door path rule for this coefficient!
Conclusions

Given our set of variables: Illiteracy, Freedom, and Agricultural Income are exogenous movers of (root causes of) poverty.

Given the assumptions in the directed graphs literature, we can consider manipulations of poverty by manipulations in one or more of these causes.

Whether or not any of these can be “easily” manipulated is, of course, another question.

Use of regression techniques to measure the quantitative relationship between causes and effects requires that we block backdoor paths and not block front door paths.
Caution

Our methods assume:

- Causal Sufficiency
- Markov Property
- Faithfulness
- Normality

Failure of any of these may change results.
More Caution: Duhem’s Thesis

Foreign Aid may be better measured (for our purposes) as Foreign Aid for Poverty Alleviation (the variable we use is Total Foreign Aid).

International trade might well be measured without natural resource exports (Dutch Disease).

Dynamic representation of poverty should be pursued. This will require a richer data set.
Acknowledgements

Motivation for the study

Aysen Tanyeri-Abur, FAO

Motivation for our study of Directed Graphs

Clark Glymour, CMU

Judea Pearl, UCLA
June 3 2003

Dear David,

Thanks for your slides about causality and DAGS. I certainly think that poverty is a very important topic and well worth studying. DAGS can sometimes be helpful in suggesting causality, but are not without problems. It is useful to require a time passage between cause and effect but this is not always clear in DAGS. I can only partly agree with your central point on page 11; a theory is helpful only if it is correct (a bad theory can be misleading) but how can you tell if the theory is correct?

I think that poverty data will always be non-experimental. Your statement on slide 15 is quite controversial although some writers equate causality and controlability (Hoover does), most do not (Pearl seems to be unclear in discussions). I think that control is a much deeper question and needs a separate definition.

I hope these comments are helpful.

Yours sincerely,

Clive W.J. Granger
Professor of Economics