Empirical investigation of the endogenous regulation of production by assets for CGE Models

Stuart J. Nettleton

Advanced Analytics Institute, University of Technology, Sydney, PO Box 123, Broadway, NSW, 2007, Australia. Fax: +612 9514 2611; Tel: +612 9514 2000; E-mail: stuart.nettleton@uts.edu.au

31 October 2012

1. Abstract

Asset turnover has been used for approximately a century in corporate capital allocation. Capital expansion coefficients have been used in the Leontief Dynamic Model but the use of asset turnover ratios to regulating production growth in Computable General Equilibrium models has been limited. This research investigates the hypothesis that there is a causal relationship between productive assets and production of commodities. The hypothesis is tested in global economic data using static and chain probabilistic graphical model selection. It was found that the hypothesis is supported for a significant number of commodities. The confirmation of the hypothesis establishes that production to assets ratios for commodities are endogenous regulators of production growth.

2. Background

The concept of asset turnover using annual sales or sales per day as a proxy for time evolved in the early twentieth century, notwithstanding that the relationship between production, capital and time had long been a topic of interest for economists. Asset turnover was a key component in Return on Net Assets (RONA) analysis developed by Pierre Samuel DuPont (1870-1954) to manage capital allocation across diversified business units at E. I. du Pont de Nemours and Company and to measure management efficiency (Davis 1950; Chandler 1977, p.446; Nettleton 2012).

RONA Analysis was implemented at General Motors where it helped underpin Alfred P. Sloan's revolutionary "Organization Study," which devolved management autonomy to business units while retaining strong central financial control (Sloan 1990, pp.140–8). RONA Analysis became widely adopted throughout US industry, where it was also known as DuPont Analysis.

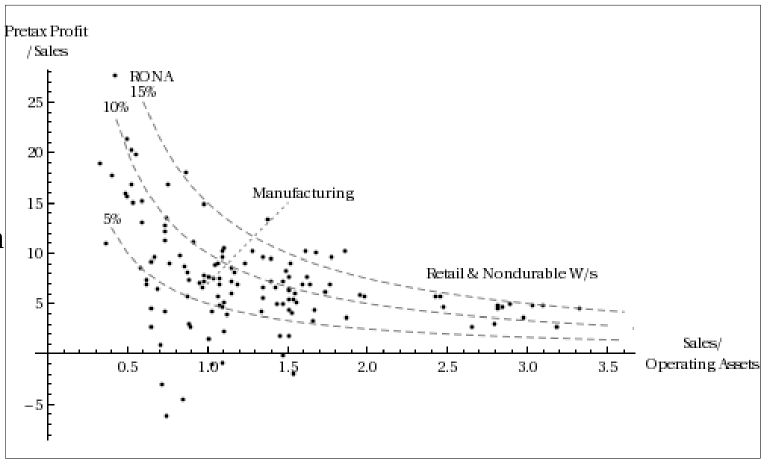

Figure 1: Components of the RONA (DuPont) Method of Ratio Analysis by Industry Sector (Pretax Profit is before interest and non-operating expenses; Operating Assets is Total Assets less non-interest bearing liabilities). Source: US Department of Commerce Quarterly Financial Report for Manufacturing, Mining, Trade and Selected Service Industries QFR Database Tables for the year to Second Quarter in 2002, 2007 and 2012 (Retail lags one quarter).

The relationship between profit margin and asset turnover, the two components of RONA, is shown in Figure 1 (US

Census Bureau 2012).

Industry asset turnover ratios are provided in Figure 2. It may be noted that retail and non-durable wholesale have high sales to assets ratios ranging from approximately 2.5 to almost 5.0 times. The asset turnover ratios for all manufacturing is approximately 1.0 times, implying one dollar of assets per dollar of sales.

Industry	2002	2007	2012
Wholesale Trade, Nondurable Goods	3.91	5.07	4.92
Food and Beverage Stores	3.33	3.69	4.03
All Wholesale Trade	3.19	3.78	3.68
All Retail Trade	2.85	2.89	2.82
All Other Retail Trade	3.1	3.04	2.82
Wholesale Trade, Durable Goods	2.66	2.98	2.8
Clothing and General Merchandise Stores	2.48	2.46	2.43
Motor Vehicles and Parts	1.48	1.54	1.67
Petroleum and Coal Products	1.13	1.51	1.66
Printing and Related Support Activities	1.99	1.77	1.65
Furniture and Related Products	1.96	1.63	1.56
Iron, Steel, and Ferroalloys	1.03	1.68	1.54
Textile Mills and Textile Product Mills	1.44	1.51	1.53
Fabricated Metal Products	1.43	1.78	1.52
Transportation Equipment	1.46	1.51	1.51
Apparel and Leather Products	1.74	1.62	1.49
Plastics and Rubber Products	1.35	1.55	1.48
Foundries	1.51	1.86	1.38
Food	1.6	1.51	1.35
Management and Technical Consulting Services			1.34
Aerospace Products and Parts	1.4	1.4	1.29
Primary Metals	1.01	1.35	1.19
All Other Professional and Technical Services (except Legal Services)			1.17
Paper	0.88	1.11	1.08
Wood Products	1.87	1.48	1.08
Machinery	1.03	1.24	1.07
All Nondurable Manufacturing	1.08	1.15	1.04
All Manufacturing	1.1	1.15	1.01
All Durable Manufacturing	1.11	1.15	0.98
Basic Chemicals, Resins, and Synthetics	0.65	1.07	0.98
All Professional and Technical Services (except Legal Services)			0.97
Nonferrous Metals	0.89	1.05	0.89
Scientific Research and Development Services			0.85
Computer Systems Design and Related Services			0.77
All Other Chemicals	0.86	0.74	0.75
Computer and Peripheral Equipment	1.1	0.9	0.74
Miscellaneous Manufacturing	1.1	0.92	0.67
Computer and Electronic Products	0.75	0.74	0.65
Nonmetallic Mineral Products	0.96	1.1	0.65
Electrical Equipment, Appliances, and Components	1.11	0.88	0.62
All Other Electronic Products	0.72	0.69	0.62
Chemicals	0.83	0.74	0.6
Communications Equipment	0.54	0.71	0.59
Beverage and Tobacco Products	0.87	0.76	0.56
Telecommunications			0.54
All Information			0.53
Publishing Industries, except Internet			0.53
Broadcasting, except Internet			0.5
Motion Picture and Sound Recording			0.5
All Other Information			0.49
Pharmaceuticals and Medicines	0.98	0.6	0.41
All Mining	0.37	0.43	0.33

Figure 2: Industry Asset Turnover Ratios(Sales to Operating Assets as defined in Figure 1 for the year to Second Quarter, sorted by 2012) Source: US Department of Commerce Quarterly Financial Report for Manufacturing, Mining, Trade and Selected Service Industries Historical QFR Database Tables [2000 - Present]

Ex post asset turnover ratios are influenced by many factors such as changes to sales quantities and prices, varying stock levels and capacity utilisation. Asset turnover risk is measured by the volatility of the asset turnover ratio. This is an important consideration in residual income valuation, which is a widely used technique that evolved from DuPont Analysis (Penman 2007, pp.367–374,653). It has been shown in growth planning and forecasting that asset turnover ratios are stable for long periods and while capacity increases in steps, asset turnover reverts to the mean within a period of one to two years (Penman 2007, p.512).

Asset turnover ratios embody production technologies and infrastructure as well as thee other organisational strategies embodied in operations, marketing and distribution. Therefore asset turnover is usually analysed in management accounts by its various asset components, such as days sales in inventories, days sales in debtors and factory investment per dollar of annual sales. From this it may be noted that asset turnover is interchangeably referred to as either a sales to assets ratio, as in DuPont Analysis, or the inverse assets to sales ratio. For example, can be expressed as either low days sales in inventories or high sales per dollar of inventory investment.

There have been many finance and investment studies in predicting company performance using DuPont Analysis (Penman 2007, p.522; Soliman 2008). However an appreciation of the use of financial techniques such as asset turnover in governing production growth has been less apparent in Computable General Equilibrium (CGE) modelling for regional and multi-regional industry performance. The Leontief Dynamic Model and CGE benchmarking techniques have explicitly incorporated asset turnover ratios.

The recursive Leontief Dynamic Model developed by the pioneer of Input-Output analysis, Nobel Laureate Wassily Leontief (Leontief 1936; 1941, p.48; 1953, pp.53–90), and David Hawkins (Hawkins 1948, p.312; Hawkins & Simon 1949, pp.245–8) has the form $(I-A+B)x^i - Bx^{i+1} = f'$, where f' is final demand, x^i is production in year i and A is the static technical or direct requirements coefficient matrix of the industrial production network and B is the stock or capital expansion coefficients matrix. A proportion of production output of $B(x^{i+1}-x^i)$ is notionally sequestered for investment. The model has since been extended (Miller & Blair 2009, pp.642–6) by incorporating a matrix of replacement capital coefficients (D) $(I-A+B-D)x^i - Bx^{i+1} = f'$.

The use of the Leontief Dynamic Model in computable general equilibrium optimisation is affected by the inherent assumption that all sectors produce at full capacity, which leads to two major issues (Miller & Blair 2009, p.649). Firstly, the Leontief Dynamic Model is rendered unsuitable in the majority of intertemporal CGE applications where capacity utilisation varies with industry activity. Secondly, capital additions may become inadequate for future production and even become negative if production declines (effectively transforming productive assets back to commodities that are available for final consumption).

There are other disadvantages of recursive approaches. In practical situations the underlying annual transfer function is usually not a set of simple recursive functions of input variables. For example, CGE and financial projections share the same features of balance sheet and income and cash flow statements that represent models of stocks and flows comprising state, information and endogenous variables. These include many discontinuous elements such as tests, switches and external constraints that frustrate the use of recursive models. It has been shown that spreadsheet-like topological processing of acyclic networks is a better approach when the equations are functions of intermediate variables rather than only the initial variables (Nettleton 2010b; Nettleton 2011).

A third, albeit more theoretical issue is a mathematical inconsistency in the Leontief Dynamic Model. It can be demonstrated using the Make V and Use U framework that production in the current period is a function of assets in the previous period as well as in the current period (Nettleton 2010c, pp.13–4). This violates the conditional independence principle that production in a given period is dependent only upon the assets in the current period.

These issues with the Leontief Dynamic model can be stabilised using a framework accounting stocks and flows model that employs external asset turnover constraints, or governors, to mediate intertemporal production growth (Nettleton 2010a, pp.331–43). Benchmarking CGE models make use of multiregional commodity Make V and Use U matrices prepared through standard United Nations' SNA93 national accounting techniques (ten Raa 2005). Productive gross margin is $(U-V^T).s$ where s is the vector of industry activity levels, which is similar to von Neumann's productive process intensities (Von Neumann 1938, p.3). Asset turnover constraints take the place of static material resource inequalities and may best be characterised as top-down asset intensity governors of productive intensities. For each commodity in each country in each period, the production constraint is given by (Nettleton 2010a, pp.335–339):

$$
\begin{aligned}
&V^T \cdot s_t \leq ninvt_{t-1} \cdot \frac{Production}{Assets} \\
&where: \\
&V^T \text{ is the Make matrix} \\
&s_t \text{ is the industry activity matrix} \\
&ninvt_{t-1} \text{ is the net investment at the end of the previous period}
\end{aligned}
$$

Although supported by the extensive use of DuPont Analysis in industry and the application of production governors in benchmarking CGE, the research question of whether assets indeed govern production has not been investigated directly in data. This may be expressed as the hypothesis that there exists a causal influence from productive assets to production in regard to the production of commodities.

The Global Trade Analysis Project (GTAP) (McDonald & Thierfelder 2004, pp.3–5; Purdue University Department of Agricultural Resources 2012) collates, harmonises and provides consistent world economic data that is suitable for CGE calculations. The GTAP Social Accounting Matrix (SAM) is a comprehensive dataset that facilitates testing of the hypothesis defined above. GTAP's SAM has three unique features:

1. Commodity rows are uniformly valued at sellers' prices, which GTAP terms “market prices”.
2. There are three factors of production, Labour, Land and Capital (which includes natural resources). Factor inputs are classified both as the non-commodity inputs required by the production process and the gross value added by production. Labour is readily identifiable as the wage and salary compensation of employees. Land and Capital together comprise the gross operating surplus and mixed income accruing from production after labour cost but before to interest, other rents and any other tangible, non-produced assets required to carry on production. Gross operating surplus plus mixed income less the consumption of fixed capital provides the net operating surplus.
3. Commodities are consumed by six agents: industrial production activities, private households, government, investment, global transport services and other regions. The role of these agents is purely consumption, which is achieved through the artifice of a Regional Household. The Regional Household receives all factor payments and taxes from production activity and distributes these receipts to the agents. Capital account investment and taxes is funded by capital account savings, received from the Regional Household, together with depreciation and any net trade balance in capital investment.

Only a single aggregated figure for each country's invested capital is available in the GTAP SAM. This means a surrogate for productive assets by commodity class is required for empirical investigations. A potential proxy for capital is provided in the factor return on Capital & Land by commodity. While the factor return on Capital & Land is the gross operating surplus and mixed income, its Dupont Analysis analogy is net assets multiplied by return on net assets.

Due to an absence of viable alternatives, factor return on capital has often been used as a proxy for capital by commodity (Reimer 2006, p.406). In circumstances where source data is unsatisfactory due to issues such as negative factor returns, GTAP harmonises the factor return on Capital & Land using appropriate estimates of return on capital (Huff et al. 2000, p.10).

However, it is quite possible that in an empirical investigation from data the factor return for Capital

& Land may be insufficiently robust as a surrogate. There are a number of major issues. Firstly, factor returns are both inputs to production and outputs from production. The aggregate value of factor inputs for each class of commodities is equal to the difference between the value of material outputs and the total value of domestic and imported material inputs and non-factor service inputs. Therefore, it may be expected that there would be a strong causal relationship between production and factor returns, which might be termed the profit causality, in contrast to the asset turnover causality where there is a causal link from assets to production. It quite possible that the profit direction could heavily mask any causal link from assets to production.

A second major issue is that the factor return on Capital & Land is a joint distribution of productive assets and the rates of return achieved in sustainable commercial enterprises in the commodity class across different countries. The factor return for Capital & Land in different countries may not have the same distribution due to differing production functions, infrastructure and returns to scale. It might also be considered that the factor return for Capital & Land may be affected by business conditions in different country and over time in a single country.

Two further substantive issues are working with world economic data, which is inherently a complex, noisy and uncertain domain, and computational challenges such as data size and combining the algorithms of computing science with the probabilistic foundations of statistics to infer a single unified model that simultaneously solves for multiple variables.

These challenges to the suitability of the factor return for Capital & Land as a surrogate and of its use in a difficult problem mean that there is an inherently large bias against establishing the hypothesis.

3. Methodology

3.1. Data Source

GTAP SAM data is investigated at the lowest level, that is, without aggregation by commodity or by country. The GTAP 8 flexagg8 databases for 2004 and 2007 were exported to 19.3Gb sql-script files using GTAP utility programs. Binary operations in Mathematica (Wolfram 2011) to condition descriptions and remove zero entries delivered reduced file sizes of about 652Mb that could be used to create standalone Hsqldb 9 (Hypersonic SQL Group 2012) databases. The commodity components of each of the variable class by country were obtained with crosstab queries executed across the Social Accountability Matrix ("ASAM") (McDonald & Thierfelder 2004).

The first column in each dataset contains a discrete country identifier, which means the dataset is "mixed", and the subsequent columns containing commodity class variables industrial production outputs (V), industrial uses inputs (U), the factor inputs of Capital, Land and Labour, and the net non-industrial consumption of output calculated as:

$$\begin{matrix} \textit{Consumption} \\ (\textit{Net}) \end{matrix} = \begin{matrix} \textit{Household} \\ \textit{Consumption} \end{matrix} + \begin{matrix} \textit{Government} \\ \textit{Consumption} \end{matrix} + \begin{matrix} \textit{Investment} \\ \textit{CGDS} \end{matrix} + \begin{matrix} \textit{Net} \\ \textit{Exports} \end{matrix}$$

In the industrial production matrix of inputs and outputs, the production of a commodity draws inputs of other commodities. These are termed Uses ("U"). For example, the production of automobiles consumes steel, electricity and many other commodities. For the production of each commodity, the Uses of all commodities are aggregated as a single “U” variable (i.e. vertically aggregate in the SAM's industrial production matrix).

Other ways to mode Uses are either to aggregate the Outputs of production (i.e. horizontally aggregate in the SAM) or to retain the Uses structure as a full Input Output matrix. The former greatly weakens any causalities between Uses and Factors because of the orthogonality between Inputs and Outputs. The latter approach of preserving an unaggregated Uses matrix is not viable. It results in the number of Use variables rising from 57, which is the number of commodities considered, to 3,249. This has a number of consequences. Firstly, the number of potential edges in each models rise from approximately 61,000 to 11.8 million. Secondly, the computational complexity of algorithms is polynomial in the number of tests, usually $O(n^2)$, where n is the number of variables, and execution time scales with the size of the data set. A study based on the unaggregated Uses matrix was commenced but failed during processing. Thirdly, even had there been success in processing, the exceptionally high number of Uses variables would be so highly disproportionate to the number of other class variables (Production (V), Capital, Land, Labour and Consumption) that overfitting could have been a major issue and the analysis of causalities potentially misleading.

3.2. Static Mixed Directed Gaussian Graphical Model

a) Methodological Background

A Directed Gaussian Graphical Model (DGGM) is appropriate for continuous data, which can be modelled with a probability distribution that is the product of factors that conditional models according to the d-separation property and Markov condition (Højsgaard et al. 2012, p.13):

$$f(x)=\prod_{v\in V} f(x_v | x_{pa(v)})$$

Each factor is associated with the distribution of a local node x_v, which depends only upon the joint distribution of the local node's parent nodes $x_{pa(v)}$.

Assumptions underlying the application of graphical models have been concisely summarised (Scutari & Strimmer 2010, pp.8–10). Two important assumptions are that there are no hidden (latent) variables and that the relationships between variables are solely conditional independencies. The latter assumption implies that the global and local distributions of discrete or categorical variables follow a multinomial distribution, the global distributions of continuous variables follow a multivariate Gaussian distribution $N_d(\mu, \Sigma)$ and the local distributions of continuous variables follow a univariate or multivariate Gaussian distribution.

Methods to infer the structure of a joint graphical model generally fall into one of low order conditional independence tests of edge likelihood or log-likelihood, heuristic search through score optimisation such as the hill-climbing algorithm and Bayesian Markov Chain Monte Carlo sampling (Højsgaard et al. 2012, p.42). In this investigation two directed acyclic graph (DAG) selection methods are applied, consistent with the approach of previous researchers (Højsgaard et al. 2012, p.62).

The first selection method for the Static model is the R pcalg PC() function, which is an example of low order conditional independence tests of edge log-likelihood (Spirtes & Glymour 1991; Spirtes et al. 1993). A skeleton of undirected edges is detected based on a threshhold p-value for local edges of 0.05. Various models that are Markov equivalent have the same undirected graph skeleton and same immoralities and so cannot be distinguished in model selection (Frydenberg 1990; Pearl & Verma 1991). The PC() algorithm orients edges to determine a complete partially directed acyclic graph (cpDAG) equivalence class rather than a specific Directed Gaussian Graphical Model. In addition to causal edges in the cpDAG, there are undirected edges and bidirection edges that have one orientation in a DAG of the equivalence class and the reverse orientation in another DAG.

The likelihood of an edge between two variables in the PC() algorithm is function of the empirical mutual information between the variables and thereby to the extent to which the variables are correlated. The partial correlations between random variables are calculated from the concentration matrix K, which is the inverse of the covariance matrix $\mathrm{K}=\Sigma^{-1}$. A weighted covariance matrix Σ is calculated with the R stats package cov.wgt(), using the S-Plus "ML" method. As the covariance matrix Σ is nearly singular, the concentration matrix K is calculated from Σ using the R corpcor package pseudoinverse() function. The partial correlation matrix is then derived using the gRbase conc2pcor() function, which calculates the partial correlation between variables *u* and *v* as follows:

$$p_{uv}|V/\{u,v\}=\frac{-k_{uv}}{\sqrt{k_{uu}k_{vv}}}$$

The PC algorithm is vulnerable to overfitting and measures are implemented to explicitly penalising complexity by regulation and restricting the hypothesis space (Koller 2012). Overfitting arises because the sampled *u* and *v* may have mutual information greater than zero some of the time, notwithstanding that *u* and *v* may be independent in the empirical distribution. As more edges lead increase the likelihood score, additional edges may be added up to the point where the likelihood score is maximised because the network is fully connected.

The second model selection method in the Static case is heuristic search through score optimisation using the R bnlearn hc() and mmhc() functions (Scutari 2010; 2012). These functions optimise a Bayesian Information Criterion (BIC) goodness-of-fit score across all possible network structures generated from the current DAG using perturbations that add, remove and reverse edges. The main advantages of using a BIC score across the network are that underfitting data is unlikely because of the asymptotic consistency of BIC scoring, and overfitting is minimised by forcing a trade-off between fit and complexity that penalises spurious edges.

A hill-climbing algorithm has the potential to become trapped in a local equivalence class where the BIC score doesn't change with the perturbations. There are often multiple local equivalence classes neighbouring the I-minimum. Therefore methods are implemented to overcome local equivalence class plateaus and explore the global space. These methods include techniques such as random restarts that tend to exacerbate the already high computational demands of the hill-climbing perturbations.

The mmhc() function was developed as a hybrid algorithm to adapt the hc() function by restricting the hypothesis space (Tsamardinos et al. 2003). This first part of the procedure is a "max-min parents and children" forward selection of the skeleton based on maximisation of the minimum association measure observed with any subset of the nodes selected in the previous iterations (Tsamardinos et al. 2006). Markov blankets of variables are detected by restricting the search space using conditional independence tests using a default significance level of 0.05. The second part of the procedure is a hill-climbing algorithm that finds the optimal network structure in the restricted space using BIC network scores.

b) Additional Data Preparation

Crosstab results for each country in 2004 and 2007 are normalised by dividing all data for each country by the aggregate of all commodity production in that country in the respective year and multiplying by 10^4. The number of observations was maximised by vertically appending the 2007 rows to the 2004 rows, while retaining the country identifier. This resulted in a file size of approximately 1.1Mb excluding Consumption variables and 1.4Mb including Consumption variables.

3.3. Chain Mixed Directed Gaussian Graphical Model

a) Methodological Background

A chain graph is similar to a DAG in having directed and undirected edges but a chain graph has no bidirectional edges or no semi-directed cycles. A chain graph is derived from a forest where nodes have at least one parent, edges are undirected and the result may include many disconnected components. Minimising the overall BIC score leads to one or more trees and is the preferred approach. In contrast minimising edge log-likelihood constrains edges weights to be non-negative and results in a single tree.

Structure learning or model selection in chain analysis requires that the search space be restricted to the edges of a conditional model between blocks of variables. It is an appropriate technique when it is clear that variables can be classed *a priori* into meaningful blocks. There are two steps needed to in this research. First, a model similar to the static case is developed solely for 2004 variables specifying the country classifier as a prior. Following this, a model is developed for the whole of the data (both 2004 and 2007 variables), specifying the 2004 model as a prior.

Model selection does not assume any order within bocks but respects the mutual order of the blocks. Thus the only causal edges derived in this analysis are the directed edges between the 2004 and 2007 blocks. Causal edges between these blocks result from an accounting effect that has become very diluted by factors that mitigate against finding strong support for the hypothesis:

- As in the Static case, dilution of causal asset turnover edges by the presence of large number of edges with reverse causality (i.e. the profit causality from “V&U” to “Capital and Land”);
- In regard to the Leontief Dynamic Model highlighted above, production in a period is theoretically a function only of average productive assets in that period and not the average production assets of previous periods. Use of a chain modelling framework stretches this theoretical foundation in three ways. Firstly, returns to Capital and Labour in each period are proxies that are one step removed from the average productive assets represented by these indicators. Secondly, although average assets in each period are conditionally independent in regard to causality of production there is, of accounting necessity, a numerical linkage. The closing value of assets in the previous period, which is the same as opening assets in the current period, becomes a common component (albeit only one component in each case) in the notional calculation of average productive assets in each period. Thirdly, in this research data for the immediately preceding periods (2005 and 2006) is unavailable, which imposes a significant intervening time period between the 2004 and 2007 class variables;

- Intertemporal linkages over the period 2004 to 2007 are subject to many exogenous influences in the intervening periods, particularly considering that these exogenous factors affect different countries across the globe in different ways;
- The chain model selection is challenged by the presence of only one observation per country compared to two observations per country in the Static case.

Although the accounting chain effect is very diluted, detection of residual causal edges from "Capital & Land 2004" to "V&U 2007" would indicate a strong relationship between these "Capital & Land" and the industrial production network "V&U."

The LWF Markov properties (S. L. Lauritzen & Wermuth 1989; Frydenberg 1990) specify that chain graphs factorise similarly to a DAG, where conditional independence is represented by d-separation. In addition, chain graphs have the property that each factor (or conditional density) further factorises according to an undirected graph where conditional independence is represented by c-separation. Mixed interaction models may comprise log-linear models for discrete variables, such as country classifiers, and Gaussian models for continuous variables.

A structure that can be represented by a mixed-interaction chain model has many advantages including elegant mathematics, efficient optimisation for high dimensional problems, such as exploiting decomposability and sparse parameterisation. Being tree structures, mixed-interaction chain models have a natural resistance to overfitting, which means that models can be generalised from a small number of samples (Koller 2012).

Components of a chain graph *G* are the connected components of the graph after directed edges have been removed (Barber 2011, p.68). The components represent distributions over the variables of the component, conditioned on the parental components. The conditional distribution is itself a product over the cliques of the undirected component and moralised parental components. Therefore the joint distribution $p(x)$ of chain graph *G* is:

$$p(x)=\prod_i p(X_i|pa_G(X_i))$$
$$p(x)\propto\prod_i \prod_{c\in C_i} \Phi(X_{C_i})$$

where C_i is the union of the cliques in component *i*, together with the moralised parental components of *i*, and Φ are the associated functions defined on each clique.

The dependence graph of the Gaussian graphical model is decomposable and model selection can exploit the closed form expressions for factor graphs. Furthermore, chain graphs can be more expressive than DAGs for marginal distributions such as undirected 4-cycles (Barber 2011, p.69).

This research applies two chain graph scoring methods in the R gRapHD minForest() function (de Abreu et al. 2011; 2010). The first method is BIC scoring, which is the default method for the extended Chow-Liu algorithm (Chow & Liu 1968; Edwards et al. 2010; Kirshner et al. 2004). This algorithm is an application of the maximum-weight spanning tree that by default optimises the BIC score to find the undirected maximum likelihood tree structure closest to the true one in the probability space, under the special constraint that each parent has just one parent. Highly efficient algorithms are guaranteed to find the maximum spanning tree (Kruskal 1956; Prim 1957). However these algorithms are $O(n^2)$ in time, which is generally unavoidable when considering pairs of edges, and a minimal forest of undirected edges cannot be guaranteed.

Bayes Dirichlet scoring is the second scoring method used by this research in the R gRapHD minForest() function. Bayes Dirichlet scoring is a maximum a posteriori (MAP) estimate using the hyper-Dirichlet distribution for conjugate priors of the decomposable graphical model and the logarithm of Bayesian independence factors for edge weights (Buntine 1991; Cooper & Herskovits 1992; Dawid & S. L. Lauritzen 1993; Heckerman et al. 1995; Højsgaard et al. 2012, pp.171–4). As the Dirichlet distribution is conjugate to the multinomial distribution, the posterior can be updated in closed form using sufficient statistics. This Bayesian scoring approach has sufficient statistics both from the data and from additional alpha-hyperparameters. Given a small amount of data, the sufficient statistics from the hyperparameter determine the prior beliefs and the strengths of these beliefs, which helps to smooth out random fluctuations in the data that can affect maximum likelihood estimates (Koller 2012). at At the asymptotic limit, real data sufficient statistics dominate and the same result is observed for both BIC and Bayes Dirichlet scoring.

Following the determination of a BIC minimum forest, a decomposable graph is determined using the R gRapHD package stepw() function. The forward selection function stepw() selects edges to be added to a triangulated graph that maximises the overall score, which by default is BIC. Identifying decomposable graphs from undirected graphs is an NP hard problem in the same way as identifying Bayesian Networks (Chickering 1996; Højsgaard et al. 2012, p.166).

b) Additional Data Preparation

Crosstab query results for each country in 2004 and 2007 are normalised in the same way as for the Static dataset, except that both 2004 and 2007 results are divided by the aggregate of all commodities produced in that country in the 2004 year. The chain dataset is completed by horizontally augmenting the 2004 rows for each country with the 2007 rows for the respective country. Variable names relating to 2004 are identified with a “y4” suffix and those for 2007 with a “y7” suffix. Chain file sizes are similar to Static file sizes.

4. Results

4.1. Static Mixed Directed Gaussian Graphical Model

On a research cluster computer the PC() function is quite fast, taking approximately 1 minute whether consumption variables are excluded or included. Computing time for the BIC-scored mmhc() function is significantly greater, extending to 54 minutes and 1.6 hours respectively. The greedy search hc() function is considerably more expensive. Excluding consumption variables the function requires 141 hours to complete and fails to finish within 2 weeks when consumption variables are included.

The number of causal edges between classes of variables is shown in Figure 3 for three algorithms (PC, hc and mmhc). The number of undirected and bidirectional edges for the PC algorithm is also shown. For example, from Capital & Land to the industrial production network (“V&U”) there are 51 direct causalities and 37 undirected or bidirected edges. Indirect causalities are not shown. There are a further 36 indirect causal edges from Capital & Land to the industrial production network (“V&U”) that are intermediated by another variable and involve an undirected or bidirectional edge.

Each of the three algorithms select models with significant causalities between variable classes. Edge counts approximately correspond to the likelihood of the variable class causalities. However, this needs to be interpreted with some caution for three reasons. Firstly, in general selected models with fewer edges are preferred.

Causal Edges 2004 & 2007 (Industrial Production & Uses (V&U), Capital, Land & Labour)

Rank	Class From	Class To	Causal Edges (pcalg PC cPDAG)	Undirected Edges (pcalg PC cPDAG)	Causal Edges (bnlearn hc DAG)	Causal Edges (bnlearn mmhc DAG)	Potential Edges
1	Capital & Land	V & U	51	37	36	18	8550
2	V & U	V & U	47	42	60	56	12 882
3	Labour	V & U	38	17	27	16	6498
4	V & U	Capital & Land	37	37	94	64	8550
5	V & U	Labour	36	17	73	48	6498
6	Capital & Land	Capital & Land	35	18	27	10	5550
7	Labour	Labour	34	4	12	7	3192
8	Labour	Capital & Land	23	7	32	6	4275
9	Capital & Land	Labour	21	7	22	5	4275

Figure 3: Causal Edges for Static Model by three algorithms (pcalg PC() with $p<0.05$, and blearn hc() & mmhc() each with edge strength > 0.85 & direction ≥ 0.5)

Secondly, it may be noted from the last column of Figure 3 that there are significant differences in the number of potential edges for each causality. Finally, the three algorithms have different scoring approaches (the PC() algorithm is constraint based, the hc() algorithm is BIC score based and mmhc() is a combination of both).

In Figure 3 it may also be noted that graphical models found by the hill-climbing algorithms hc() and mmhc() have large numbers of causal relationships from the industrial production network ("V&U") to other class variables. For example, in the profit direction where the industrial production network ("V&U") drives Capital & Land. Notwithstanding this issue, thee hc() and mmhc() algorithms confirm the PC() result that "Capital and Land" drives the industrial production network "V&U." In each case this relationship ranks fourth after industrial production network ("V&U") sources.

This is remains the situation in Figure 4, where consumption variables are included. As noted above, the hc() algorithm did not complete. Introducing consumption variables reduces the number of causal edges from Capital & Land to the industrial production network ("V&U") from 51 to 46, placing the relationship on equal footing with the in the reverse or profit driver from "V&U" to Capital & Land.

Causal Edges 2004 & 2007 (Industrial Production & Uses (V&U), Capital, Land, Labour & Consumption)

Rank	Class From	Class To	Causal Edges (pcalg PC cPDAG)	Undirected Edges (pcalg PC cPDAG)	Causal Edges (bnlearn hc DAG)	Causal Edges (bnlearn mmhc DAG)	Potential Edges
1	V & U	V & U	54	24	na	55	12 882
2	V & U	Capital & Land	46	17	na	46	8550
3	Capital & Land	V & U	46	17	na	16	8550
4	Capital & Land	Capital & Land	39	8	na	7	5550
5	Labour	V & U	37	12	na	16	6498
6	Labour	Labour	34	6	na	6	3192
7	V & U	Labour	31	12	na	37	6498
8	Labour	Capital & Land	23	7	na	5	4275
9	Consumption	V & U	23	9	na	1	6498
10	V & U	Consumption	22	9	na	16	6498
11	Consumption	Capital & Land	19	5	na	3	4275
12	Labour	Consumption	15	5	na	5	3249
13	Consumption	Labour	15	5	na	2	3249
14	Capital & Land	Labour	13	7	na	6	4275
15	Capital & Land	Consumption	12	5	na	11	4275
16	Consumption	Consumption	11	12	na	5	3192

Figure 4: Causal Edges for Static Model including Consumption (bnlean hc() did not complete)

4.2. Chain Mixed Directed Gaussian Graphical Model

The second stage of chain analysis is generally limited by computational resources. While on a research cluster computer the combined minForest() and stepw() processing time is just 4 minutes for the first stage in 2004 of 247 variables, this rises steeply to 18.3 hours when the secondary stage of 493 variables for 2004 and 2007 is included. The model including consumption variables, which has a total of 697 variables (+41%), has an overall processing time of 66.7 hours (+264%). Bayes Dirichlet scoring increases the processing time to 22.4 hours (+22%) excluding consumption variables and 87.9 hours (+32%) including consumption variables.

Causal edges for the chain graph are shown in Figure 5. These edges are identified with the Chow-Liu algorithm using two different scoring approaches, default BIC scoring and Bayesian Dirichlet scoring. BIC scoring ranks the causal edge from Capital & Land 2004 to V&U 2007 almost equal second, on par with the other factor, Labour 2004, as a driver.

There are two issues with the results of Bayesian Dirichlet scoring. Firstly, it exhibits a strong tendency to hub at the first variable. With the country classifier included all edges hub at the country classifier and there is a total absence of relationships between all other variables. In Figure 5 the problematic country classifier is excluded for Bayesian Dirichlet scoring on the basis that the Static case and the Chain case with BIC scoring each show the the country classifier is not significant. With the exception of edges that hub at the first variable, which happens to be "V&U 2004," it may be noted that the number of causal edges is otherwise similar to BIC scoring.

Causal Edges 2004 to 2007 (Industrial Production & Uses (V&U), Capital, Land & Labour)

Rank	Class From	Class To	Edges (gRapHD BIC)	Edges (gRapHD Dirichelt)	Potential Edges
1	V & U 2004	V & U 2007	523	611	12 996
2	Labour 2004	V & U 2007	338	323	6498
3	Capital & Land 2004	V & U 2007	335	326	8550
4	Capital & Land 2004	Capital & Land 2007	300	328	5625
5	V & U 2004	Capital & Land 2007	273	360	8550
6	Labour 2004	Labour 2007	273	298	3249
7	V & U 2004	Labour 2007	226	305	6498
8	Labour 2004	Capital & Land 2007	215	228	4275
9	Capital & Land 2004	Labour 2007	197	206	4275

Figure 5: Causal Edges for Chain Model by gRapHD ChowLiu Algorithm using BIC scoring and Dirichlet scoring excluding country classifier

The second issue with the results of Bayesian Dirichlet scoring is that the results are less concise that for default BIC scoring. Bayesian Dirichelt scoring identifies 11.3% more intertemporal causal edges than BIC scoring. In contrast to other experience, Bayesian Dirichelt scoring does not provide a sparser graph than the default BIC method in this research (Højsgaard et al. 2012, pp.171–4).

Causal Edges 2004 & 2007 (Industrial Production & Uses (V&U), Capital, Land, Labour & Consumption)

Rank	Class From	Class To	Edges (gRapHD BIC)	Edges (gRapHD Dirichelt)	Potential Edges
1	V & U 2004	V & U 2007	540	639	12 996
2	Capital & Land 2004	V & U 2007	360	333	8550
3	Labour 2004	V & U 2007	343	373	6498
4	Consumption 2004	V & U 2007	331	328	6498
5	V & U 2004	Capital & Land 2007	301	366	8550
6	Capital & Land 2004	Capital & Land 2007	301	295	5625
7	V & U 2004	Labour 2007	283	368	6498
8	Labour 2004	Labour 2007	280	291	3249
9	Labour 2004	Capital & Land 2007	222	226	4275
10	Consumption 2004	Capital & Land 2007	219	202	4275
11	Capital & Land 2004	Labour 2007	212	219	4275
12	V & U 2004	Consumption 2007	209	269	6498
13	Consumption 2004	Labour 2007	202	209	3249
14	Consumption 2004	Consumption 2007	177	200	3249
15	Capital & Land 2004	Consumption 2007	168	174	4275
16	Labour 2004	Consumption 2007	160	164	3249

Figure 6:Causal Edges for Chain Model including Consumption

It may be also be noted that the number of causal edges between class variables in both BIC and Bayesian Dirichlet scoring is significantly greater than in the Static case. This implies that the Static model is the more robust although, as outlined above, Chain model selection can be regarded as working on residuals and therefore a far more onerous test of the hypothesis.

The number of intertemporal causal edges for the Chain Model including Consumption variables is shown in Figure 6. The default BIC scoring method elevates the rank of the number of causal edges from Capital and Land 2004 to the industrial production network "V&U" 2007 well above the Labour driver.

Single Commodity Causal Edges (Percentage of Potential Commodity Edges)

Consumption Variables Excluded	Static PC	Static bnlearn hc	Static bnlearn mmhc	Chain Chow-Liu BIC Scoring	Chain Chow-Liu Bayes Dirichlet
Capital	43%	50%	29%	35%	31%
Land	66%	33%	0%	38%	22%

Consumption Variables Included	Static PC	Static bnlearn hc	Static bnlearn mmhc	Chain Chow-Liu BIC Scoring	Chain Chow-Liu Bayes Dirichlet
Capital	45%	na	24%	31%	24%
Land	38%	na	5%	38%	38%

Figure 7: Percentage of Single Commodity Causal Edges identified by Model Selection Algorithms from Capital & Land to the Industrial Production Network (V or U)

5. Discussion of Results

Causal edges resulting from model selection in both the Static and Chain models showed that "Capital and Land" is a strong driver of the industrial production network "V&U." However a test of the hypothesis needs to limit the analysis of causal edges to single commodities by eliminating inter-commodity effects where assets in one commodity may be causal to production of a different commodity.

The percentage of potential single commodity causal edges from Capital and Land to the industrial production network (V or U) is shown in Figure 7. These percentages are based on a maximum of 57 edges from Capital to V or U, and a maximum of 18 edges from Land to V or U.

There are two issues to be noted. Firstly, double counting has been eliminated in these results. For example, if a commodity has a causal edge to both V and U then the causality to the network is counted once only.

Secondly, the results for the Static case include both direct and indirect causal edges. However, the latter are few in number. With consumption variables excluded there is only one additional propagated edge in each of the PC(), mmhc() and hc() algorithms. Including consumption variables results in three propagated edges with the PC() algorithm and one propagated edge with the mmhc() algorithm.

Causal edges are mapped in Figure 8 to indicate the commonality in model selection. It may be noted that the algorithms generally detect similar edges, and even across both the Static and Chain models.

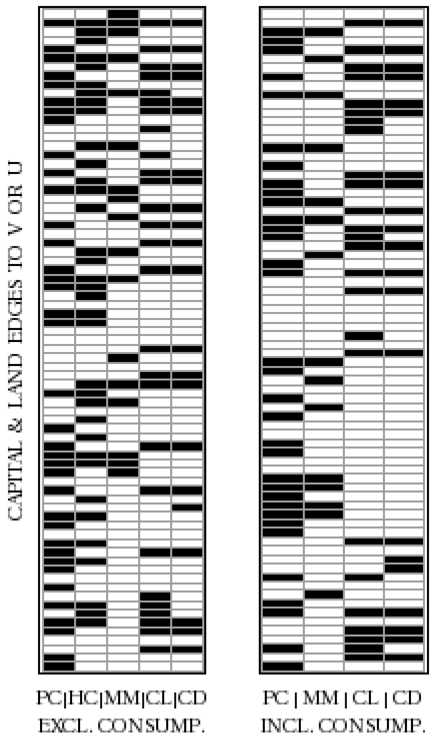

Figure 8: Map of Single Commodity Causal Edges by Algorithm Algorithm key: PC(), HC (bnlearn hc()), MM (bnlearn mmhc()), CL (Chow Liu BIC scored) and CD (Chow Liu Bayesian Dirichlet scored)

It may be noted in Figure 7 that for the Static network, excluding Consumption variables, Capital drives the industrial production network (V or U) in almost half of cases and approximately one-third of cases in the more onerous intertemporal Chain network. Land is also a significant driver in both models, albeit the algorithms demonstrate greater volatility with regard to this causality. In each case the percentage of detected single commodity causalities is marginally reduced by the inclusion of consumption variables.

Notwithstanding potential masking of the hypothesis of a causal edge from Capital and Land to the industrial production network (V or U) by the reverse causality of profit, which removes such edges from consideration, it is possible to conclude that model selection across a number of algorithms confirms a large percentage of significant single commodity causal edges from the productive assets of a commodity to production of the commodity.

6. Conclusion

The use of asset turnover, which is the ratio of assets to sales, in RONA (DuPont) Analysis has been outlined. Asset turnover has also been applied in the Leontief Dynamic Model, in the form of capital expansion coefficients, and as an external constraint to regulate intertemporal production in benchmarking Computable General Equilibrium. The question addressed in this research is whether the use of asset turnover in regulating growth in CGE models is an inherent feature to be found empirically from data.

The hypothesis that there is a causal link between the productive assets for a commodity and the production of that commodity has been tested in global economic data derived from a fully disaggregated Global Trade Analysis Project (GTAP8) Social Accountability Matrix. The GTAP8 database contains consistent data for the 2004 and 2007 years.

Factor returns to Capital and Land were been selected as a proxy for industry assets. Model selection techniques for mixed Gaussian probabilistic graphical models were used to detect the presence of causal edges from Capital and Land to the industrial production network. The presence of the causal edges from assets to production is masked by the presence of edges with reverse causality representing profit.

Static and Chain model selection cases were investigated. The PC, hill-climing and max-min hill climbing algorithms were applied to a Static case of a notional single year by combining normalised 2004 and 2007 SAMs. The Chow-Liu algorithm with BIC and Bayesian Dirichlet scoring was applied to a dynamic case of Chain causality from 2004 to 2007, where a 2004 graphical model was the Bayesian prior for a 2007 graphical model.

Model selection from data in both the Static and Chain cases confirms that for a significant number of commodities, Capital and Land is a causal driver of the corresponding industrial production network for the commodity. It is concluded that the hypothesis of a causal relationship between productive assets and production is supported. This answers the research question confirming that production to assets ratios for commodities are endogenous regulators of production growth.

7. References

de Abreu, G.C.G., Edwards, D. & Labouriau, R., 2011. *gRapHD: Efficient selection of undirected graphical models for high-dimensional datasets*, Available at: http://http://cran.r-project.org/web/packages/gRapHD/index.html.

de Abreu, G.C.G., Edwards, D. & Labouriau, R., 2010. High-Dimensional Graphical Model Search with the gRapHD R Package. *Journal of Statistical Software*, 37(1), pp.1–18.

Barber, D., 2011. *Bayesian reasoning and machine learning*, Cambridge University Press.

Buntine, W., 1991. Theory refinement on Bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*. pp. 52–60. Available at: http://dl.acm.org/citation.cfm?id=2100669 [Accessed October 18, 2012].

Chandler, A.D., 1977. *The visible hand: The managerial revolution in American business*, Belknap Pr. Available at: http://books.google.com.au/books?hl=en&lr=&id=hUkqx76sF6oC&oi=fnd&pg=PA1&dq=%22The+Visible+Hand%22+Chandler&ots=CJvmzltSiY&sig=_SyEVEgLc1bxOlSFCnto_FKtubw [Accessed October 23, 2012].

Chickering, D.M., 1996. Learning Bayesian networks is NP-complete. *Lecture Notes in Statistics, New York, Springer Verlag*, pp.121–130.

Chow, C. & Liu, C., 1968. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3), pp.462–467.

Cooper, G.F. & Herskovits, E., 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4), pp.309–347.

Davis, T.C., 1950. How the Du Pont organization appraises its performance. *American Management Association*, 94(7).

Dawid, A.P. & Lauritzen, S.L., 1993. Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3), pp.1272–1317.

Edwards, D., de Abreu, G.C.G. & Labouriau, R., 2010. Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. *BMC bioinformatics*, 11(1), p.18.

Frydenberg, M., 1990. The chain graph Markov property. *Scandinavian Journal of Statistics*, pp.333–353.

Hawkins, D., 1948. Some conditions of macroeconomic stability. *Econometrica: Journal of the Econometric Society*, 16, pp.309–322.

Hawkins, D. & Simon, H.A., 1949. Note: Some conditions of macroeconomic stability. *Econometrica: Journal of the Econometric Society*, 17(3 & 4), pp.309–322.

Heckerman, D., Geiger, D. & Chickering, D.M., 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3), pp.197–243.

Højsgaard, S., Edwards, D. & Lauritzen, S., 2012. *Graphical Models with R*, Springer Verlag. Available at: http://books.google.com.au/books?hl=en&lr=&id=em5QdpWmljAC&oi=fnd&pg=PR3&dq=%22Graphical+Models+with+R%22&ots=-YWA25vcVy&sig=naG_GBOQ0eWIvRz4vZRD-JF_es8 [Accessed July 16, 2012].

Huff, K., McDougall, R. & Walmsley, T.L., 2000. Contributing input-output tables to the GTAP data base. Available at: http://docs.lib.purdue.edu/gtaptp/1/ [Accessed November 15, 2012].

Hypersonic SQL Group, 2012. HSQLDB. Available at: http://hsqldb.org/ [Accessed October 12, 2012].

Kirshner, S., Smyth, P. & Robertson, A., 2004. Conditional Chow-Liu tree structures for modeling discrete-valued vector time series. In *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*. UAI. Arlington, pp. 317–324. Available at: http://arxiv.org/abs/1207.4142 [Accessed October 13, 2012].

Koller, D., 2012. Probabilistic Graphcial Models (Stanford University). Available at: https://class.coursera.org/pgm/ [Accessed October 12, 2012].

Kruskal, J.B., 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1), pp.48–50.

Lauritzen, S.L. & Wermuth, N., 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, pp.31–57.

Leontief, W.W., 1953. Dynamic Analysis. In W. W. Leontief, ed. *Studies in the Structure of the American Economy*. Oxford University Press New York.

Leontief, W.W., 1936. Quantitative input and output relations in the economic systems of the United States. *The Review of Economic Statistics*, 18(3), pp.105–125.

Leontief, W.W., 1941. *The structure of American economy, 1919-1929: An empirical application of equilibrium analysis*, Harvard University Press.

McDonald, S. & Thierfelder, K., 2004. *Deriving a Global Social Accounting Matrix from GTAP Versions 5 and 6 Data*, Global Trade Analysis Project (GTAP).

Miller, R.E. & Blair, P.D., 2009. Input-output analysis. *Cambridge Books*.

Nettleton, S.J., 2010a. *Benchmarking climate change strategies under constrained resource usage*. UTS theses submitted as part of the Australasian Digital Theses Program (ADT). Sydney, Australia: University of Technology, Sydney. Available at: http://utsescholarship.lib.uts.edu.au/iresearch/scholarly-works/handle/2100/1012 [Accessed March 30, 2010].

Nettleton, S.J., 2012. *Personal communication with senior executives of E. I. du Pont de Nemours And Company*,

Nettleton, S.J., 2010b. Processing Large Scale Systems of Equations in Financial Models. In Wolfram Technology Conference. Champaign, IL.: Wolfram, Inc. Available at: http://www.wolfram.com/events/techconf2010/speakers_p2.html [Accessed October 30, 2012].

Nettleton, S.J., 2011. Spreadsheet-Like Solution of Large-Scale Systems of Equations: Comprehensive Corporate Model Example. *Wolfram Demonstrations Project*. Available at:

http://demonstrations.wolfram.com/SpreadsheetLikeSolutionOfLargeScaleSystemsOfEquationsCompreh/ [Accessed January 23, 2012].

Nettleton, S.J., 2010c. The Service Science of Climate Change Policy Analysis: applying the Spatial Climate Economic Policy Tool for Regional Equilibria. In International Input Output Association (June 2010). Sydney University, Australia. Available at: http://viXra.org/abs/1210.0144.

Von Neumann, J., 1938. A model of general economic equilibrium. *Readings in Welfare Economics*, 13(1945), pp.1–9.

Pearl, J. & Verma, T.S., 1991. A theory of inferred causation. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.51.7221 [Accessed October 12, 2012].

Penman, S.H., 2007. *Financial statement analysis and security valuation*, McGraw-Hill/Irwin New York, NY. Available at: http://www.ifm.wi.tum.de/files/courses/ss04/financial_statement_analysis/financial_statement_analysis_ch13.pdf [Accessed October 22, 2012].

Prim, R.C., 1957. Shortest connection networks and some generalizations. *Bell system technical journal*, 36(6), pp.1389–1401.

Purdue University Department of Agricultural Resources, 2012. GTAP 8 Database. Available at: https://www.gtap.agecon.purdue.edu/databases/v7/default.asp.

ten Raa, T., 2005. *The Economics of Input Output Analysis*, New York: Cambridge UniversityPress. Available at: www.cambridge.org/9780521841795.

Reimer, J.J., 2006. Global production sharing and trade in the services of factors. *Journal of International Economics*, 68(2), pp.384–408.

Scutari, M., 2012. *CRAN - Package bnlearn*, Available at: http://cran.r-project.org/web/packages/bnlearn/ [Accessed October 12, 2012].

Scutari, M., 2010. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35(3). Available at: http://www.jstatsoft.org/v35/i03/ [Accessed October 12, 2012].

Scutari, M. & Strimmer, K., 2010. Introduction to Graphical Modelling. *arXiv preprint arXiv:1005.1036*. Available at: http://arxiv.org/abs/1005.1036 [Accessed October 20, 2012].

Sloan, A., 1990. *My years with general motors*, Crown Business.

Soliman, M.T., 2008. The Use of DuPont Analysis by Market Participants. *Accounting Review*, 83(3), pp.823–853.

Spirtes, P. & Glymour, C., 1991. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1), pp.62–72.

Spirtes, P., Glymour, C. & Scheines, R., 1993. *Causation, Prediction, and Search*, Springer-Verlag.

Tsamardinos, I. et al., 2003. Algorithms for large scale Markov blanket discovery. In *Proceedings of the sixteenth international Florida artificial intelligence research society conference*. pp. 376–381. Available at: http://www.aaai.org/Papers/FLAIRS/2003/Flairs03-073.pdf [Accessed October 12, 2012].

Tsamardinos, I., Brown, L.E. & Aliferis, C.F., 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning*, 65(1), pp.31–78.

US Census Bureau, 2012. Quarterly Financial Report: Manufacturing, Mining, Trade, and Selected Service Industries. *United States Census Bureau*. Available at: http://www.census.gov/econ/qfr/index.html [Accessed October 24, 2012].

Wolfram, S., 2011. *Mathematica 8.0.4 © 1998 – 2011*, Champaign, IL.: Wolfram Research, Inc. Available at: http://www.wolfram.com/mathematica/.