MSc in Applied Economics and Finance

Master’s Thesis

Causality Analysis of Drivers for Test Market Investments

Practical Implications for Copenhagen Capacity

Dilovan Deniz Celik

Supervisor: Tim Mondorf

79 pages & 112,172 Character

Copenhagen Business School

06-01-2014

Abstract

Traditionally the arguments used by various Investment Promotion agencies, for attracting foreign direct investment in test markets have not been sufficiently backed up by empirical data. Instead, these arguments have been based on case studies and different types of anecdotal evidence, and there has yet to be conducted an in depth analysis on the drivers of test market investments.

However before it is possible for Investment Promotion agencies, such as Copenhagen Capacity, to develop arguments to attract Foreign Direct Investments in test markets, it is necessary to create a proper empirical foundation for them to use.

In this thesis I have chosen to create this foundation by using the theory of Bayesian Networks. Bayesian Networks allows me to investigate conditional dependencies, and together with the causal sufficiency assumptions it is possible to determine causal relationships between different parameters.

In the Bayesian network different causal relationships are suggested, however it is not possible to confidently assume causal sufficiency. This means that only conditional probabilities were found.

Nonetheless, the data does give Copenhagen Capacity, and other I. P. agencies, a better foundation to build better arguments, which can help attract Foreign Direct Investments in test markets.

The analysis concludes that 1) the perception of particular key parameters might be more relevant than the actual state of these parameters, 2) Foreign Direct Investment in a specific industry does not diverge largely from the overall Investment flows and, 3) The attraction of Foreign Direct Investments resembles a "winner takes it all" game.

Table of Contents

Introduction

This thesis is the result of a project proposal from Copenhagen Capacity, which is the Investment Promotion Agency of the Capital Region of Denmark. Copenhagen Capacity is interested in knowing what drives test market investments. Knowing these drivers, would help Copenhagen Capacity optimise their investment promotion efforts.

If the thesis shows that qualities which Denmark and the Capital Region have are important when business decide to invest in a test market, they can use this information to target their sales process to focus on these qualities.

On the other hand, if the thesis finds results which are not present in the Region, Copenhagen Capacity can use my results to help lobby for a better environment for foreign companies who wish to invest in test facilities in the Capital Region.

Based on my initial research I decided to focus on cause of test market investments. There are several reasons why I felt that this was the most relevant part of the project.

In my initial interviews with different agents in the industry, I quickly realised that there was a lag of data to support the different arguments which were used in the sales process. Because of this, many consultants in the industry have decided to use the test market argument as a side argument when convincing companies to invest in Denmark. Furthermore, I felt that a methodology which could be used in an analysis of causality between different parameters and investments in test markets can be applied in other economic problems.

Copenhagen Capacity

As mentioned in the previous section Copenhagen Capacity is the Investment Promotion Agency of the Capital Region. The organisation was established in 1996 by the Capital Region and developed from *Fonden for investeringsfremmende*, to increase the competitiveness of the region.

In its infancy the Copenhagen Capacity primarily focused on attracting Foreign Direct Investments, and was therefore measured on its ability to create jobs from these investments. In this way it acted as a classic Investment Promotion Agency. However over the course of the last five to eight years, the organisation has gone through and organisational change. It has gone from having a focus on attraction of Foreign Direct Investment to a broader competiveness focus. Because of this the organisation is now be split up in 3 main branches.

The Investment Promotion Branch, this is the core of Copenhagen Capacities efforts and what drives its main results. The Investment Promotion branch is focused on the attraction of Foreign Direct Investment from the following Markets:

- Geographic
    - China
    - Japan
    - Germany
    - United Kingdom

- Industries
    - Life Science
    - Cleantech
    - Information Communication Technology
    - Logistics

Other than focusing on the attraction of Foreign Direct Investment, Copenhagen Capacity has also had a focus on the Expansion and Retainment of current foreign companies located in the region. This has been a successful strategy, which has historically contributed positively to the reputation of the organisation.

The second branch is the Cluster Creation Branch, which primarily have focused on Cleantech. The Cluster department biggest project was the creation of the Copenhagen Cleantech Cluster and its membership of the International Cleantech Network.

The motivation for creating the Copenhagen Cleantech Cluster, was the ideas based on Michael Porters paper *The Competitive Advantage of Nations*[1]. The paper states that competitive companies tend to be locate near other companies in the same industry, and that this has a beneficial effect for both the companies and for the competitiveness for the regions or nations which contains these clusters. Because of this it was believed that attracting companies within the Cleantech industry would become easier if we had a strong cluster.

The Cleantech industry was chosen because of two main reasons, 1) it was believed that the Copenhagen Region and Denmark in general had a historical advantage in attracting these types of companies. This was believed to be true both because Denmark was early adopters of wind energy and because it was believed that the region had a competitive brand when it comes to this industry.

The Copenhagen Cleantech Cluster proved to be a very successful project. However because of its success it will have to depart from Copenhagen Capacity (a process which will be completed in June 2014). And the cluster department will have to find new projects which in can work with.

The third Branch of Copenhagen Capacity is the Talent Attraction Branch. This is the newest part of the organisation which has the main purpose of attracting talented foreign individuals, which could benefit the Danish economy.

The motivation for doing so is that if Denmark becomes a hub for talented internationals, companies which require this talent will relocate to Denmark to get the right employees.

[1] The Competitive Advantage of Nations, Michael E. Porter, 1990

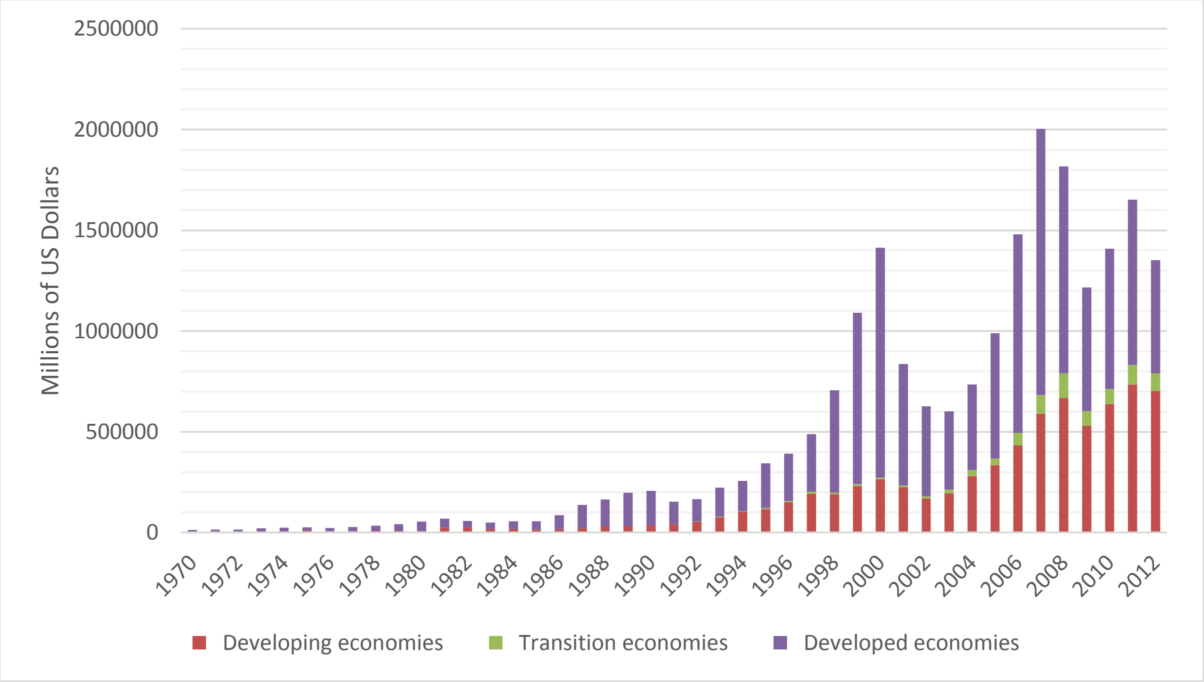
The Investment Promotion Industry

Because of the need to maintain a competitive advantage and because of the benefits from Foreign Direct Investments; many countries, regions and cities have established Investment Promotion Agencies such as Copenhagen Capacity.

This has created a highly competitive Investment Promotion Industry, which all try to sell their particular Country, Region or city as the most attractive location for Investments.

However over the last couple of years the financial crisis, the growth in third world countries and the lack of demand from the European Market, has changed the flows of Foreign Direct Investments around the world. Especially Europe have been hit hard, and is losing an increasingly larger share of the global Foreign Direct Investments to developing countries such as India and China.

Figure 1 Share of FDI Inflows Developing vs Developed countries



Source: UNCTAD Statistical database, http://unctadstat.unctad.org/TableViewer/tableView.aspx?ReportId=88

Because of this it has become necessary for the regions and thereby the Investment Promotion Agencies to become even more competitive.

This is done in many ways. Regions have created special economic zones such as in Mexico or Russia[2], others brand their regions as having the best talent, being cheapest or having the right customers. Most of these arguments are usually backed up with statistics, or other concrete evidence.

One argument which is increasingly used by the Investment Promotion Agencies is that their particular regions would serve as the perfect test market for foreign companies, however it is rarely backed up this any type of evidence. This becomes problematic because it becomes increasingly harder for the Investment Promotion Agencies to use this argument.

[2] http://blogs.law.uiowa.edu/ebook/faqs/what-are-special-economic-zones

Research Questions and Limitation

It will be the main focus of this thesis to work with the issue of what constitutes a good and competitive test market. However this can be done in several ways.

One way of doing so is to choose a qualitative approach where one could use knowledge from individual cases where a companies did invest in test markets, and draw on their experiences.

This is an approach which have been quite popular in the industry. What is usually done is to find companies which have chosen to make a test market investment in their particular regions, and use their case to prove that their region is the perfect test market.

Even though this can be a very successful approach when convincing potential investors to locate their test market functions in their particular region, it is also highly biased, it does not provide any true evidence for that particular region.

Because of this I have chosen to take a more quantitative approach in this thesis. I have chosen to research if it is possible to find general causal relationships with investments in test markets. This is done to try and move beyond the biased approach which is currently being used.

However it is still essential that my research can be applied by the industry and particularly Copenhagen Capacity, therefore I have chosen the following research questions.

- Is it possible to find any causal relationships with test market investments?
  - If yes, what are these causal relationships?

- How can these results be applied for Copenhagen Capacity to create better results for them as an organisation?

Theoretical Framework

The next section will provide the reader with a theoretical foundation from which the analysis can be understood. I will focus on different theories and methodological approaches which can determine causal relationships.

Introduction to Causality and Correlation

It is in our human nature to assume that because to variables are correlated, they must be dependent on one another in some way, especially in cases of high correlation. Unfortunately this reasoning simply isn't true. If one decided to use such an approach, one could easily fall victim to random correlations.

Although the correlation is highly significant[3], it would be wrong to assume that sunshine has had a negative effect on FDI in Design, Development and Testing. Taking a further look into the data demonstrates why this might be the case.

The mathematical formula for correlations is defined as:

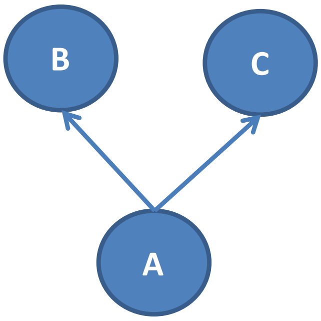
$$\rho_{xy} = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{E[(X - E[x])(Y - E[y])]}{\sigma_x \sigma_y}$$

This shows me that the variations in X resemble the variations in Y (ex. that X increases when Y increases). There are two situations in which this type of function would show correlations between otherwise independent variables. The first is when, as previously mentioned, that the variations are identical by chance, whereas the second is when both variables are influenced by another third variable[4]. In this case the variations in both would happen simultaneously, which would also be when variations in the third variable would happen. Such a relationship is shown in the graph below.

[3] In fact it is the 7th most significant amongst the 696 variables I tested

[4] Such variables will henceforth be mentioned as parent variables.

Figure 2 Relationship between variables



Another way one might use correlation to determine causation of variable X on variable Y is to correlate a lagged value of X with the present value of Y.I In this case the formula for correlation would look like this:

$$\rho_{xt-n,y} = \frac{cov(X_{t-n}, Y)}{\sigma_{xt-n}\sigma_y} = \frac{E[(X_{t-n} - E[x_{t-n}])(Y - E[y])]}{\sigma_{xt-n}\sigma_y}$$

However, unfortunately this kind of test would also be wrong as it is based on a logical fallacy known by its Latin name of *Post hoc ergo propter hoc* (after this, therefore because of this). It simply states that since the latter event followed the former event, the latter event must be caused by the former.

The reason this type of reasoning is flawed is because it is purely based on the temporal priority principle, and although the principle is important when testing for causality, it is inadequate if not subject to other types of testing. This is because it is still subject to the same errors as the simple correlation mentioned above, since both random correlations and the effect of a third unknown variable would still lead to correlations between otherwise independent variables.

Due to the aforementioned limitations of correlation, it is essential for this analysis is that I choose to include causality tests in methodology in order to determine which variables are causing FDI in Design, Development and Testing, and which are merely correlated. To do this, several statistical tests have been proposed.

Granger Causality

Most of these tests rely on time series data, such as the Granger Causality Test proposed by Clive Granger. The intuition behind the test is that economic causality (or in this case Granger Causality) is present[5]. It is not enough to show the correlation between the lagged value of the cause on the dependent variable (as discussed above), but it is essential that if true Granger Causality is present, then a sudden temporary spike in the cause would lead to a sudden temporary spike in the dependent variable. One example of this could be that it is not enough to show that investments in infrastructure are positively correlated with future FDI in Design, Development and Testing, however, if there were true Granger Causality a sudden spike in Infrastructure Investments would lead to a sudden spike in FDI in Design, Development and Testing. The test can be performed in the following way:

Step 1 is to determine the proper amount of lagged values in a univariate autoregressive model of Y so that I get:

$$Y_t = \propto_0 + \beta_0 Y_{t-1} + \beta_1 Y_{t-2} + \cdots + \beta_n Y_{t-k} + \varepsilon_t, \text{where } \mathrm{\varepsilon_t} \text{ is the error term}$$

Step 2 is to augment the regression by including the lagged values of X to the model so that I get:

$$Y_t = \propto_0 + \beta_0 Y_{t-1} + \beta_1 Y_{t-2} + \cdots + \beta_n Y_{t-k} + \gamma_0 X_{t-1} + \gamma_1 X_{t-2} + \cdots + \gamma_n X_{t-k} + \varepsilon_t$$

After including the lagged values of x, one determines which parameters are individually significant be performing a t-test. The variables which are individually significant are then retained in the model only if they collectively provide additional explanatory power. Whether or not additional explanatory power is provided is determined by an F-test.

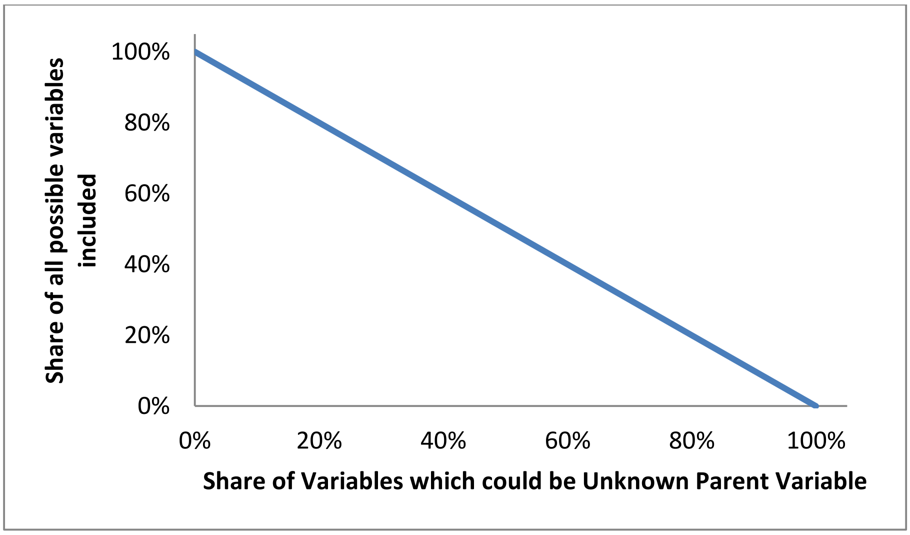
The null hypothesis (which is that measured values does not Granger-cause the dependent variable) is only confirmed when no lagged values of x is kept in the regression.

[5] *Basic Econometrics*, D. Gujarati et al, 2008

The Granger Causality Test provides me with a simple quantitative method of determining causality between different variables, and would therefore be very tempting to use in my analysis, but they is two primary reasons why this is not possible.

Firstly Granger Causality is not defined as true causality. Although the Granger Causality Test significantly helps decrease the probability of two independent variables being deemed dependent because of variations which are randomly alike. It does not eliminate the problem of both variables being affected by a common parent variable. Conversely, such issues could be removed by including enough variables in my analysis. If I were to include a large number of variables, the probability a third parent variable being present is smaller, shown in the following figure:

Figure 3 Relationship between included and unknown variables



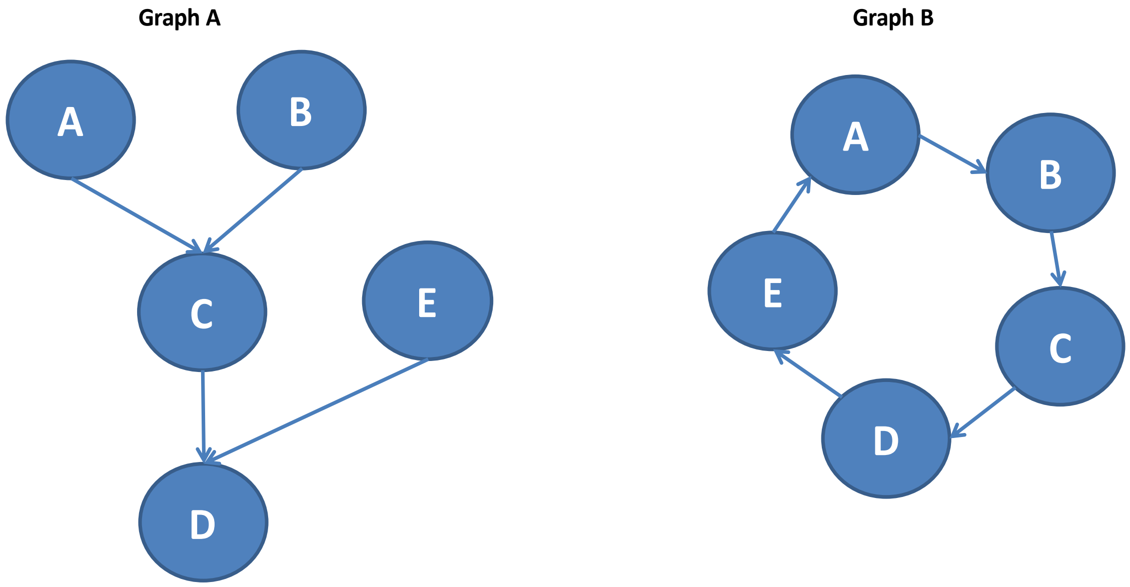
Although it would be impossible to include all possible variables, it is not impossible to include a large enough number so that I minimise the risk of third unknown parent variable to have an effect on my results. Yet this is not feasible, bringing me to my second point.

The Granger Causality test relies on the use of time series to conduct the analysis. This makes my data collection much more difficult, the data I have collected is cross-sectional, meaning it has been collected at one point in time. There would be a theoretical possibility of collecting the same sort of data as panel data (cross-sectional data collected over time, so that it could work as several time series). This would be highly unfeasible. The data collected so far has come from several different surveys which have been conducted over a long period of time as opposed to a short one, therefore finding time series data which is comparable to the existing data in both quantity and quality seems highly improbable. Even though the Granger Causality Test would be able to incorporate it, it seems highly unlikely that I would be able to acquire it.

The Bayesian Networks Approach

Another way of showing causation amongst different variables is through Causal Bayesian Networks. The Bayesian network is, in its essence, a probabilistic graphical model in which random variables are connected through their conditional probabilistic dependencies[6]. Two examples of such a model are shown in figure 3 below.

Figure 4 Probabilistic Graphical Model

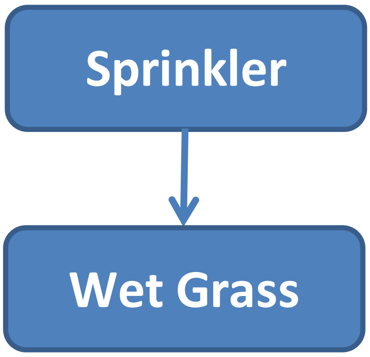


Both graphs are probabilistic graphical models, as well as directed graphs. The bubbles (henceforth denoted as nodes) show the different variables in the network. The arrows (henceforth be denoted as edges) show in which direction the dependencies go, so that in Graph A, variable D is dependent on variable C and E.

[6] Probabilistic Networks – An Introduction to Bayesian Networks and Influence Diagrams, Uffe B. Kjærulff & Anders L. Madsen, 2005

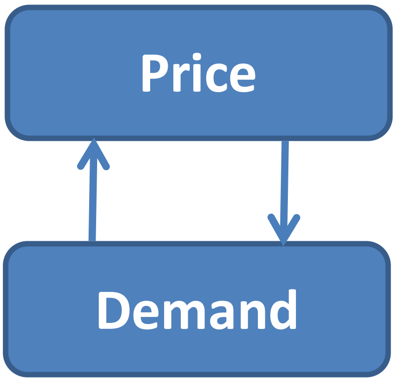
Furthermore graph A can be described as a Directed Acyclic Graph (DAG), and graph B can be described as a Directed Cyclic Graph (DCG). The difference is that an increase in A will not lead to a future increase in A in the DAG, whereas such an effect would happen in the DCG.

Figure 5 Example of DAG



An example of the DAG is the classic example of the sprinkler and wet grass. If the sprinkler is on, the grass will become wet, but the grass being wet does not turn the sprinkler on.

Figure 6 Example of DCG



On the other hand, the DCG can be explained by the example of price and demand. An increase in price will most likely lead to a decrease in demand, which would then lead to decrease in price. Although this is a very simple example of a DCG, it does prove the point that the variables in the DCG affect themselves.

This seemingly small difference in the graphs leads to different ways in which they can be applied. When you are working with DCGs it is necessary to measure the effect of variables on each other over time. It is therefore essential to have either time series or panel data from which you can calculate the probabilities given earlier actions. If you try to calculate the probabilities in this type of graph without time series or panel data, it could prove impossible to find which way the edges would be pointing. This issue does not exist with DAGs, since as a single variable does not have an effect on itself over time, it is not necessary to measure its effect over time. Because of this I can use cross-sectional data to calculate the conditional probabilities of the different nodes on each other.

In this thesis I will only use Bayesian Networks and DAG's, due to two reasons. The first reason is because of the feasibility criteria. As previously discussed, my data is collected as cross-sectional data, and it would be highly improbable to collect data of equal quantity and quality as panel data. Secondly and most importantly, it is against the assumptions of the Bayesian Network to include DCGs as the loops in a DCG would render it impossible to decompose the joint probabilities in a Bayesian Network.

Theory of Bayesian Networks

In the following section I will describe the theory of Bayesian Networks more indepth. I will firstly describe the notation that will be used, then going on to describe the principles of Bayesians Inference and probability theory which are both essential to understanding the intuition behind Bayesian networks. Following that, I will communicate the process of learning Bayesian networks. To conclude there will be an account of the assumptions and limitations encountered with the Bayesian Network Approach.

Notation

Before I start to describe the theory behind causality with Causal Bayesian Networks, I will spend some time describing the notation of a Bayesian Network. In the section above I outlined what nodes and edges are, therefore I will focus on the types of nodes and basic mathematic definitions[7].

- Parents: *A parent node is a node which causes carries information about another node. In the case of the sprinkler and the wet pavement, the sprinkler is the parent of the wet node.*

- Spouses: *Spouse nodes, are two parent nodes which have edges pointing towards the same node.*

- Children: *A child node is a node which has an edge directed towards it.*

[7] *Causality: Models, Reasoning, and Inference* 2nd Edition, Judea Pearl (2009)

- Descendants: *A node which is the child of a child.*

- Ancestor: *A node which is the parent of a parent.*

- Siblings: *Two nodes who have the same parent.*

- Family: *A family of nodes is a node and all its parents.*

In mye analysis I am primarily interested in the parents and ancestors of the Investment in Design, Development and Testing FDI.

Bayesian Networks

Before I define a Bayesian network, I need to define Bayes rule and conditional probabilities. Conditional probabilities are essential for building a Bayesian Network[8].

Using the example of the sprinkler and the wet pavement, one would state the conditional probability of having a wet pavement given the sprinkler being turned on. This can be formally expressed in the following way:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

Bayes rule expresses the same conditional probability, but without using joint probabilities which can simplify the calculation in certain situations. I can do so because of $P(B|A) * P(A) = P(A,B)$, therefore I get.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Using the following simple calculation, I can define conditional independence.

[8] Bayesian Methods: General Background, An Introductory Tutorial, E.T. Jaynes, 1996

Definition 1[9]: Conditional Independence

$X\ is\ conditionally\ independent\ from\ Y\ given\ Z\ whenever\ P(x|y,z) =$
$P(x|z), \qquad if\ P(y,z) > 0$

This means that X is only conditionally independent from Y if Y does not attribute with any additional information given that the variable Z is present.

An example of two variables which are conditionally independent of each other could be the hypothetical example of coffee and lung cancer. In this case I might conduct a survey where I find that people who drink coffee have a higher likelihood of having lung cancer. This type of data would suggest a dependency between coffee drinking and lung cancer, so that:

$$P(Lung\ Cancer) < P(Lung\ Cancer|Drinking\ Coffee)$$

However if one were to introduce a third variable, which represents whether or not the subject smokes, I would see another picture. This hypothetical survey would suggest that people who smoke are both more likely to drink more coffee as well as more likely to get lung cancer. On the other it shows that whether or not a person smokes, the likelihood of that person getting cancer does not increase with the intake of coffee. This can be expressed with definition 1 as:

$$P(Lung\ Cancer|Smoking, Drinking\ Coffee) = P(Lung\ Cancer|Smoking)$$

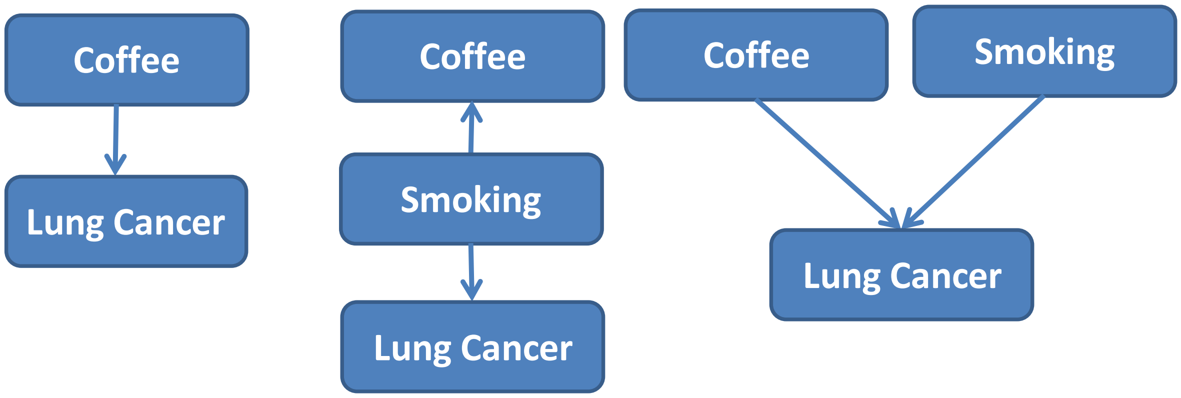
Alternatively if I were to find that even though smoking has an effect on the lung cancer, there is also evidence that coffee has an isolated effect on lung cancer. This would be shown formally in the following way:

$$P(Lung\ Cancer|Smoking, Drinking\ Coffee) \neq P(Lung\ Cancer|Smoking)$$

[9] *Causality: Models, Reasoning, and Inference* 2nd Edition, Judea Pearl (2009)

The intuition behind conditional independence will be essential for my analysis, and will be used to create DAGs and infer causality between variables.
In the case of the coffee and lung cancer survey I would get the following DAGs

Figure 7 Possible DAGs given different dependencies



Once I understand the intuition behind conditional dependencies and independencies in probability theory, I can proceed with developing my framework and are now ready to define a Bayesian Network Mathematically. There are a lot of equivalent definitions of Bayesian Networks, such as the definition from *Ben-Gal et al* (2007):

Definition 2: Bayesian Network[10]

A Bayesian network ***B*** *is an annotated acyclic graph that represents a Joint Probability Distributions over a set of random variables* ***V****. The network is defined by a pair B = (G, Θ), where G is the DAG whose nodes X_1, X_2, …, X_N represents random variables, and whose edges represent the direct dependencies between these variables. The Graph G encodes independence assumptions, by which each variable X_i is independent of its nondescendents given its parents in G. The second component Θ denotes the set of parameters of the network. This set contains the parameter $\Theta_{xi|\pi i} = P_B(x_i|\pi_i)$ for each realization x_i of X_i conditioned on $\pi_{i,}$, the set of parents of X_i in G. Accordingly, B defines a unique Joint Probability Distributions over* ***V****, namely:*

[10] *Causality: Models, Reasoning, and Inference* 2nd Edition, Judea Pearl (2009)

$$P_B(X_1, X_2, \dots, X_N) = \prod_{i=1}^{n} P_B(X_i|\pi_i 0) = \prod_{i=1}^{n} \theta_{xi|\pi i}$$

Using the above definition I can further proceed in my analysis. The next step will be to uncover an undirected graph, which is also known as a Markov Network. The Markov Network, can be seen as the skeleton of the Bayesian Network.

Definition 3.1: Markovian Parents[11]

*If **V** denotes the different random variables in a DAG and P (v) denotes the joint probability of these variables. I use PA$_j$ as the minimal set of parents to X$_j$; I call this set of parents for Markovian parents in the case where it satisfies:*

$$P(x_j|pa_j) = P(x_j|x_1, x_2, \dots, x_{j-1})$$

What this definition tells me is that I will always try to minimise the set of parents which contains information about a given node. It helps me minimise the amount of edges in my Markov Network, and helps me exclude parents who are in reality conditionally independent. I can now use the definition of Markovian Parents to define Markov Compatibility.

Definition 3.1: Markov Compatibility[12]

If the sum of the conditional probabilities of X$_j$ given its parents is equal to the joint probability of all variables, such that it satisfies:

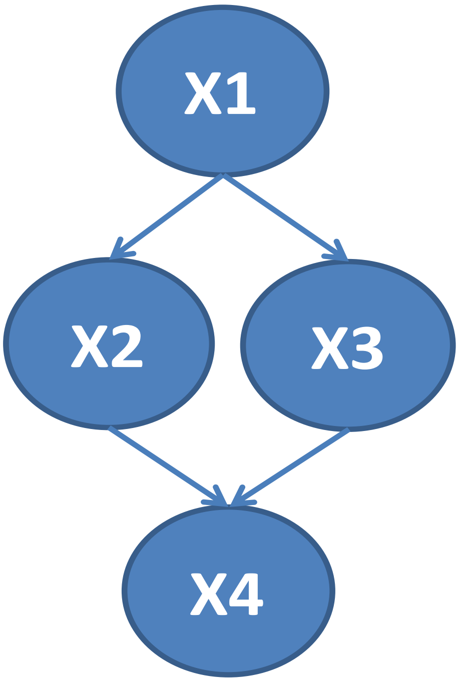
$$P(x_1, x_2, \dots, x_n) = \prod_{j} P(x_j|pa_j)$$

I have factorization of P. If this factorization is relative to DAG G, I can call it them Markov Compatible.

To exemplify the concept of factorisation I can use the following DAG:

[11] *Causality: Models, Reasoning, and Inference* 2nd Edition, Judea Pearl (2009)
[12] *Causality: Models, Reasoning, and Inference* 2nd Edition, Judea Pearl (2009)

Figure 8 Example of DAG for Markov Compatibility



The Joint Probability of the DAG can be expressed as:

$$P(X_1, X_2, X_3, X_4)$$

Having this joint probability function, I can now test that against a probability function of the given parameters and its parents to see if I have Markov Compatible with the following function:

$$P(X_1, X_2, X_3, X_4) \\ = P(X_1) * P(X_2|X_1) * P(X_3|X_1) * P(X_4|X_2, X_3)$$

The reason why it is important to know if the probability function is Markov Compatible with the DAG is because this compatibility can be used to check whether or not I have the right DAG for a given dataset.

One way of describing this set of probability distributions with its set of conditional independencies, is by using d-separations. The d-separation criterion is defined in the following way:

Definition 4: d-Separation[13]

A path between two nodes can be said to be d-separated when it is blocked by another node. So that the nodes X and Y are independent given node Z, for example given X - Z - Y.

D-separation is fairly simple to understand once I understand the concept of conditional independencies. If two nodes are conditionally independent, they must be d-separated to some degree. One can then say that d-separations are merely the graphical counterpart to the statistical conditional independencies. Furthermore it is possible to say that DAG is Markov

[13] *Causality: Models, Reasoning, and Inference* 2nd Edition, Judea Pearl (2009)

Compatible with the data when the d-separations are equivalent to the conditional independencies.

Using this knowledge of d-separations in DAGs leads me to an important fact about my Markov Network; if I know that two nodes are d-separated I also know that they are conditionally independent, and that it is highly unlikely that there is causal relationship between the two.

Now that I know the basic properties of Markov Networks, I can learn how to go from undirected graph to DAGs and I need to find out how one can learn the structure of optimal DAG.

Learning the Structure of the Bayesian Network

Before I can learn the structure of a DAG, I need to introduce two conditions on which I rank different DAGs, these two conditions are Minimality and Stability.

Condition 1: Minimality[14]

The structure of a DAG G is minimal, when G could not be presented with fewer edges.

The minimality condition ensures that I always choose the Markov Parents rather than any other set of parents, and in that way always describe the conditional dependencies in the most efficient way possible, in addition to not making the different parameters dependent on parents, from which they would have otherwise been conditionally independent from. In some situations, the minimality condition is enough to uncover the correct structures. However, in order to ensure the quality of my networks I need to introduce the stability condition.

[14] *Causality: Models, Reasoning, and Inference* 2nd Edition, Judea Pearl (2009)

Condition 2: Stability[15]

The structure of a DAG G is stable when the structure does not change, when subject to small changes in the observed data.

These two conditions help me ensure that I choose the best DAG, which is also Markov Compatible, given the data.

An example of this could be to consider a binary variable C that takes the value 1 whenever the outcomes of two coins (X and Y) are the same and takes the value 0 if not. In the trivariate distribution, by this parameterisation, each pair of variables is marginally independent yet conditionally independent on a third variable. Such a dependence pattern may in fact be generated by three minimal causal structures, each depicting one of the variables as causally dependent on the two others, but there is no way of deciding among the three. In order to rule out such pathological parameterisations, I impose a restriction on the distribution of stability. This restriction conveys the assumption that all the independencies embedded in P are Stable; that is, they are entailed by the structure of the model D and hence remain invariant to any change in the parameters[16].

Knowing these conditions, and the properties of a Bayesian Network which I defined in the previous chapter, makes it simple (however tedious) to learn the structure of the Bayesian Network.

- Step 1 is to find the joint probability distribution of each node.
- Step 2 is to learn the structure of all edges in the graph using an IC algorithm, as described below.

The IC Algorithm builds on the intuition from Rebane and Pearl (1987)[17]. In essence, one can discover the direction of the edges of X and Y by finding a third variable which correlates

[15] *Causality: Models, Reasoning, and Inference* 2nd Edition, Judea Pearl (2009)
[16] Causality: Models, Reasoning and Inference, 2nd Edition, Chapter 2.4, Judea Peal, 2009
[17] The recovery of causal poly-trees from statistical data, G. Rebane and J. Pearl, 1987

with Y, but not Z. In this situation one can build an algorithm which directs the edges. This algorithm is called the Inductive Causation (IC) Algorithm.

The IC Algorithm essentially works as an "IF[18]" function. When the third variable is not correlated on X or Y, you move on to find a new parameter. If however the parameter is correlated with both, and renders X and Y conditional independent given the new variable, you must draw arrows from X and Y to the new parameter[19]. It should also be noted that the IC algorithm is not the only algorithm which can be used to learn the orientation of the edges, but it has been included here to give a theoretical example of how such algorithms work.

Although this problem sounds relatively simple, it has been proven to be NP-Hard to solve in reality. This is mainly because of the many possible DAGs[20], when you have many different parameters which you want to include. The number of possible Bayesian can be expressed as 2^{N}[21] where N is the number of nodes in your DAG. This is illustrated in the below table:

Table 1 Number of DAGs given number of parameters.

Parameters	Number of possible DAGs
2	4
3	8
4	16
5	32
10	1.024
20	1.048.576
50	1.125.899.906.842.620
100	1,26765E+30
1000	1,0715E+301

Because of this it is practically impossible to estimate the perfect DAG to fit the data. However, several search algorithms have been developed to deal with this problem so that I can approximate a Bayesian Network which fits the data. These algorithms will be discussed in greater detail later.

[18] Also known as Case or When given the programming language you use.

[19] A theory of Inferred causation, J. Pearl & T. Verma, 1991

[20] Learning Bayesian Networks is NP-Complete, David Maxwell Chickering

[21] Given the Bayesian Network consist of Boolean values, the amount is much higher with non-Boolean discrete and continuous values.

In the above section, I have described the basics of Bayesian Networks, and how such structures are learned.

The first thing I outlined was the concept of conditional independencies. This was crucial to understanding the intuition behind Bayesian Networks. If two parameters are conditionally independent, I say that the two cannot directly cause each other, which means that there is a third parameter which better describe the parameters. I called this parameter the Markov Parent.

I also described the concepts of Markov Compatibility and d-separation. With these definitions, I moved from the algebraic space in to the graphical space. This will help by giving me a better overview of the different causal effect needed.

Furthermore, I introduced the conditions of Minimality and Stability. These conditions help me find the most solid and correct DAGs. By including them I make sure that the parent nodes presented in the DAG are always the Markov Parents and that they are not sensitive to small changes in the data.

In the end I went through the learning process for the structure of a Bayesian Network, introduced the IC Algorithm, to turn the errors, and that the problem is NP-Hard

Knowing this about Bayesian Networks is very helpful, but I have only shown probabilistic relationships through the graph. Although probabilistic relationships can have a very useful application in business and other real life situations, it is not enough to state any causal dependencies. I will spend the next chapter attempting to define such causal relationships.

From Bayesian Network to Causal Bayesian Networks

To understand the relationship between Bayesian Networks and Causal Bayesian Networks, one must first understand the *Common Cause Principle*, stated by Reichenbach in 1956[22]. The *Common Cause Principle* is as follows:

Definition 5: Common Cause Principle

If the two parameters A and B are probabilistically correlated, then either there is a causal dependency between A and B which causes the correlation, or there is a third parameter (the common cause) which is causally dependent on both A and B.

Reichenbach formalised this idea in the following way:

$$If\ P(A,B) > P(A) * P(B),$$
$$then\ there\ must\ be\ a\ common\ cause\ C\ so\ that$$
$$P\left(\frac{A,B}{C}\right) = P\left(\frac{A}{C}\right) * P\left(\frac{B}{C}\right)$$

If I choose to believe the intuition behind the *Common Cause Principle*, it can have great implications for whether or not causality is present. Using the above formula, I can quickly see if two parameters are causally dependent or if there is a third latent parameter which could be causing both. Although this seems like an easy solution to the causality problem, it does make a rather large assumption.

Assumption 1: No random correlation

Given the Common Cause Principle, two parameters cannot be correlated by coincidence, but must be directly dependent or through one or more common causes.

The above assumption can prove to be problematic in reality, since it cannot be mathematically proven. Although it is unlikely to find completely random correlation it is not

[22] The Direction of Time, Hans Reichenbach, 1956

impossible. This assumption first of all requires that I have a large sample[23], and even in the case of large samples, it is not completely impossible unless you have the entire population. Furthermore, it is not against the laws of nature for two events to be randomly correlated.

However because, with current knowledge, it would be practically impossible to disprove two variables from being merely randomly correlated, I would not be able to apply any test for causality if I do not believe in the *Common Cause Principle*.

Given the *Common Cause Principle*, I can determine causal relationship if one more condition is met.

Assumption 2: The Causal Sufficiency Condition[24]

In order to determine probabilistic dependencies to be causal, I must ensure that all data and all possible parameters is included in my sample, so that I with confidence can exclude any third variable being the cause of the probabilistic dependency.

Although it is theoretically possible to have data which includes all possible parameters, it is rarely the case in practice, especially when the data is passively observed. Even with very large data sets, it is possible that certain important parameters have been omitted, (for example, because of a prior selection bias).

But just as with the *Common Cause Principle*, if I choose not to assume that the *Causal Sufficiency Condition* holds, it makes my analysis practically impossible and so I must assume that it holds.

Given the *Common Cause Principle* and given the assumption that the *Causal Sufficiency Condition* holds for my data, I can assume causality between two parameters when there is a conditional dependency between the two.

I have now discussed the two important assumptions, which turn Bayesian Networks into Causal Bayesian Networks. The two where the *Common Cause Principle* and the *Causal*

[23] Given the law of large numbers, random samples tend to adjust towards the value of the population, when the sample size is increased. This means that random correlations becomes more likely when sample sizes are decreased

[24] An Introduction to Causal Inference, Richard Scheines

Sufficiency Condition, if both assumptions hold, Bayesian Networks become a very powerful tool in determining causal relationships between different nodes. Nevertheless, I found that these conditions can be very difficult to uphold in practice.

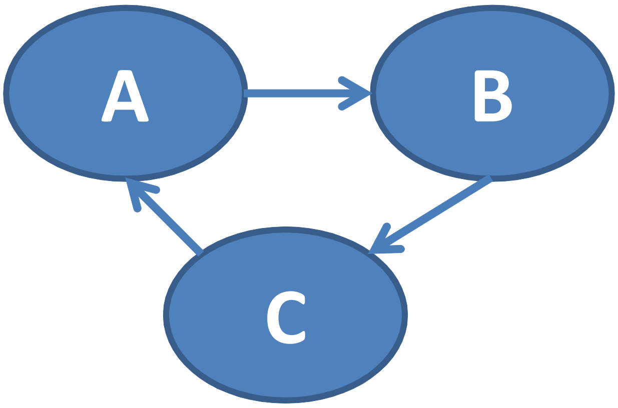
In the next section I will briefly discuss the limitations of Bayesian Networks.

Limitations of Causal Bayesian Networks

Other than the aforementioned assumptions not holding, Bayesian Networks have one major limitation. Because Bayesian Networks are represented as a DAG, it is not possible to present causality in form of *Feedback Loops*.

Feedback loops can be described as a complete causal path, where an initial parameter which begins the causal path is causally affected by the parameter at the "end" of the causal path. Such a feedback loop is shown in figure 8 below.

Figure 9 Example of Feedback loop



Feedback loops are very common in economics and data series containing temporal data.

One example is the cyclical nature of the financial markets. If I have bull market, stock prices will start increasing, and in turn these increases will lead investors to believe that stock prices will rises even further, making them buy more stocks and hence causing stock prices to rise.

Not being able to explain causality through feedback loops is major limitation with Bayesian Networks in two ways. Firstly, not being able to present data in causal feedback loops, simplifies the way I explain certain events. Although this might sound like a good thing, it can make my understanding of the cause and effect relationships I am trying to present incorrect.

Secondly, not being able to use feedback loops renders me unable to describe the data over time, and I am left with merely a snapshot of a chain of events.

However even with this limitation of Causal Bayesian Networks, it is still a powerful tool to determine causal networks, especially if I remain critical to my results.

Methodology

Methodological Paradigm

Before I start describing the methodological approach I will take throughout this analysis, I must first establish the paradigm through I perceive the world and my results.

I have chosen to perceive the world through the glasses of Neo Positivism. Neo positivism refers to the philosophical world view in which I believe that the truth is best obtained through objectivity[25]. This entails that I believe that I can provide an objective truth which holds for all observers. On the other hand I believe that the results I get are not the absolute truth and that they should be recognised as the best possible results only.

By using such a methodological paradigm, is still concrete enough to assert confidence in my results, and at the same time stress the importance of common sense when interpreting the results.

I believe that this methodology will help me benefit in my analysis because there is a great deal of unbacked claims from different countries claiming to be the perfect test market. By using the neo positivism, I am able to distance my from the use of anecdotal evidence and subjective statements.

[25] Stanford Encyclopaedia of Philosophy, plato.stanford.edu/entries/logical-empiricism

Data structure

Data Collection Motivation

The collection of data for this thesis has been inspired by the approached of Big Data[26]. In Big Data, one gathers very large datasets, and runs different correlations to see if one can find trends in the data, and to help build predictive models.

Other than using very large datasets, the big data approach, tries not to have any preconception about the data which it tries to analysis. It is this part of the approach which I wish to replicate in my analysis, because I that it will help me approach this analysis with as little bias as possible. I will try and do so by gathering as much data as possible, and let the data speak for itself.

However there are still substantial differences between my approach and that of data scientist working with Big Data.

This helps me in two ways, first of all it will helps minimise my bias because I do not pick the parameters I test myself. Second, by including as much data which is feasible, it becomes easier to fulfil the causal sufficiency condition.

With this in mind, I will now discuss, the data which I have collected and the sources from which it comes.

[26] Big Data, A Revolution that will transform how we live, work and think, Kenneth Cukier and Viktor Mayer-Schonberger, 2013

Data Sources

The three primary data sources which I have chosen to use is fDi Markets, fDi Benchmark and the IMD World Competiveness survey.

fDi Markets is part of the fDi Intelligence Portfolio, which is owned by the Financial Times Group. fDi Intelligence have specialised in Foreign Direct Investment (FDI) data and thereby in supporting the Investment Promotion industry.

fDi markets tracks international investments either in the form of Greenfield Investments or Expansions, which creates jobs. This means that joint ventures are only included in the sample if they lead to new physical investments, and that they exclude Mergers & Acquisitions and other pure equity investments.

This is restricting the analysis because I exclude domestic investments from the data. By doing so, I restrict my analysis to only defining the drivers of international test market investments, instead of focusing on all test market investments.

fDi markets have a team of in house analyst who searches the following sources of data on Greenfield Investments[27]:

- Information sources owned by the Financial Times Group
- External Media sources, such as press releases and other media corporations
- Project data from Industry Organisations and Investment Promotion Agencies
- Data from Market Research and Publication Companies
- Direct Company Sources

Looking through their sources it becomes evident, that fDi Markets heavily rely on the investments being publicised.

This can be the cause of concern because of two things. First it does not include projects which companies have successfully kept confidential. For instance investments which would damage the company's brand or have other harmful side effects if it was publicised.

[27] http://www.fdimarkets.com/about/

Furthermore the sample is heavily dependent on the qualifications of the in house analysis departments, and whether or not they possess any biases. One could easily imagine investment from smaller companies in smaller countries, where any public information was only publicised in the local media in the native language. In this case I am dependent on this department, to be aware of these media and to be able to understand them.

On their website fDi Markets states that they do not take responsibility of any inaccuracies which might exist in the data. I can therefore not trust the data as being the absolute truth. However, I do not have better sources to obtain this type of data, and I must trust that it gives a picture which is comparable to the truth.

The second source I have used is fDi Benchmark. Just as fDi Markets, fDi Benchmark is a part of the fDi Intelligence Portfolio owned by the Financial Times Group.

The main purpose of fDi Benchmark is to help different agents compare different locations for FDI (abbreviation for Foreign Direct Investments). This however is not the way I have chosen to use it. Other than having a Benchmarking part, fDi Benchmark contains a database function. This database has gathered information on a large number of different parameters for a more than 350 cities.

The data collected in fDi Benchmarks database comes mostly from reliable global sources such as Tower Watson, the Economist Intelligence Unit, The World Bank, etc. Local statistical agencies are used, where comparable data is available[28].

By extracting this data from fDi Benchmark, I make my data collection efforts a lot more efficient. I can now get data, which was created from different sources, from one source, which makes the data search more feasible. Furthermore I do not have access rights to a lot of the data sources directly even if I wanted to. This is because a lot of them require you pay for them before you can gain access.

The last data source I have used is the IMD World Competiveness Report. It is an annual report which is published by the World Competiveness Center at the IMD Business School in Lausanne Switzerland. The survey ranks the world 60 most competitive economies. The ranking consist of hard data (in this case hard data consist of statistically measured

[28] http://www.fdibenchmark.com/about/

parameters, such as unemployment, GDP, School Enrolment Rate etc.) as well as Survey results. The distribution is 2/3 hard data and 1/3 survey results. The Countries which participate are ranked on the four main factors which are:

- Business Efficiency
- Government Efficiency
- Economic Efficiency
- Infrastructure

In turn these factors are split in to further sub parameters, which in total leads to 300 different parameters from which they are ranked.
Although I see the World Competitiveness Report as a reliable data source, there are two attributes with the data which raises my concern.

The first one is that the data is reported on a country level, where the report from my two other sources are on a city level. This means that some of the granularity in my analysis will be lost.
There are two ways I can deal with this issue. The first one would be to aggregate the city level data in to a country level data points. The second is to treat the country level data as city level data, so that I say that all cities within the same country have the same score. I have chosen to do the latter because of one major reason, which is that the first would give me a false perspective. I do not have data for all cities in all countries and can therefore not properly aggregate the data on a country level.

The second attribute for the data which raises my concern is that 1/3 of it, is from survey data. Although this is not a problem in itself it does change the way I perceive the data. Because it is survey data, I must be aware that I am not looking at the truth of what is being measured, but merely the participant's perception of what is asked.

Having the discussed the Data sources I am using, I can now develop the framework which I will use.

Methodological framework

Correlation

The first step in my analysis will be to find the parameters which are significantly correlated with the investments in test markets. The reason I look at parameters which have a significant correlations, is because of Reichenbach's Common Cause Principle which were described in earlier in this report.

Because of the principle I can limit the possible parameters which could cause Investments in test markets to the parameters which are significantly correlated with the amount of investments.

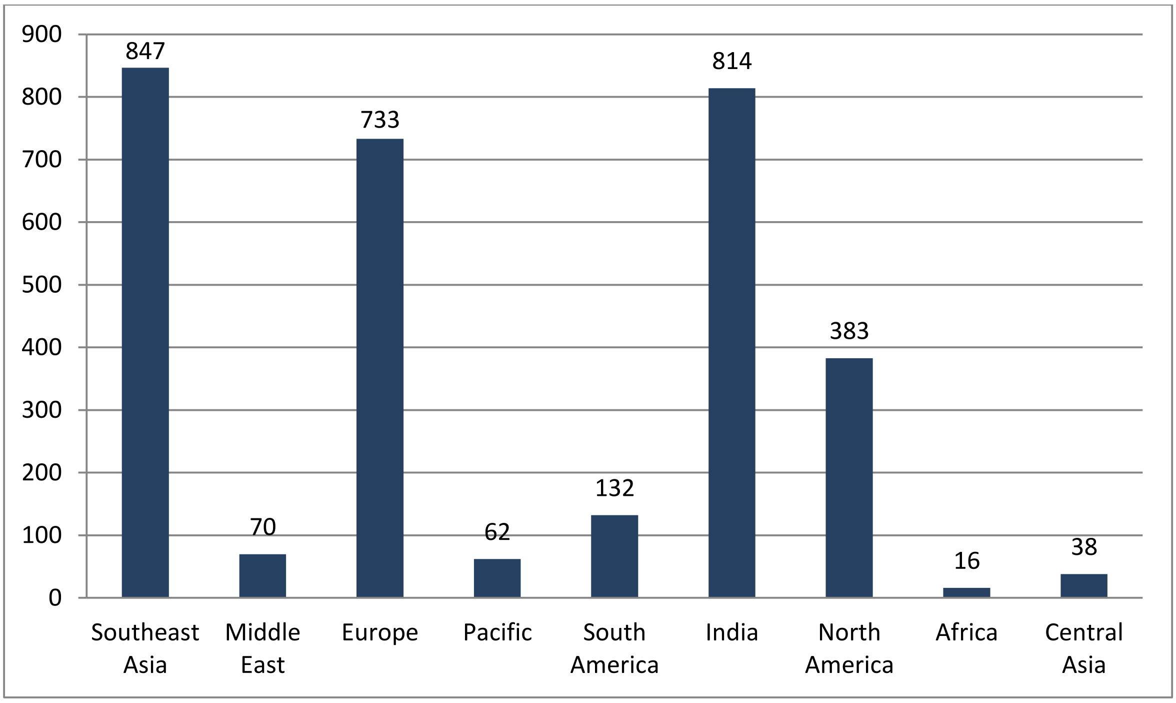
Unfortunately I do not have any parameter called investments in test markets, the closest I can come, is Investments in Design, Development and Testing, a discussion of this parameter will be conducted in the next section

Choice of measurement

The type of investments I am looking at are Foreign Direct Investment projects. That means I am working with quantity rather than quality. This is because I do not have enough data on the size of the projects. All though I could look at the overall FDI stock in a country, this does not give me the same opportunity to look into a specific type of investments, since it does not have the same granularity. In the end I have chosen to go with the ability to focus on a specific industry, since this is the only way to answer the problem statement, rather than focussing on size of the projects.

Before I can go any further, I need to describe the nature of the data, the first thing I wish to understand is if there is any characteristics which will make it difficult to create a model that can help understand the behaviour of test market investments. The way I do this is to look for outliers, which will affect the data in more than the rest. The first thing I look at is if there is any geographic region which dominates the dataset.

Figure 10 Amount of Greenfield investments in Design, Development and Testing per region



Source: fDi Markets

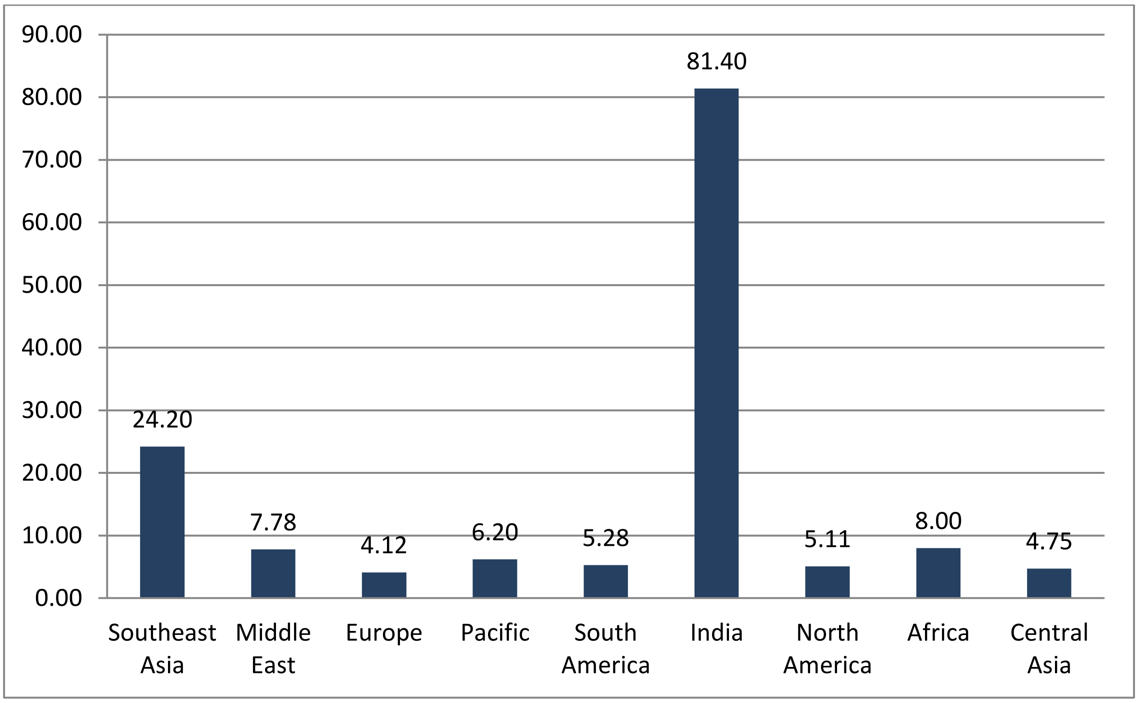
As I can see four regions are dominating the total amounts of test market investments, these are;

- Southeast Asia
- Europe
- India[29]
- North America

But looking at this data is not enough, because the regions are not equally represented in size. My data sets includes a lot more cities in the western world (especially the UK and the US), than the rest, and it is therefore important to look at the relative amounts of investment compared to number of cities in the region.

[29] I had to register India as an individual region, because the amount of investments in Design, Development and Testing were so large, that it would affect any other region to much, if it were included.

Figure 11 Amount of Greenfield investments in Design, Development and Testing per City



Source: fDi Markets

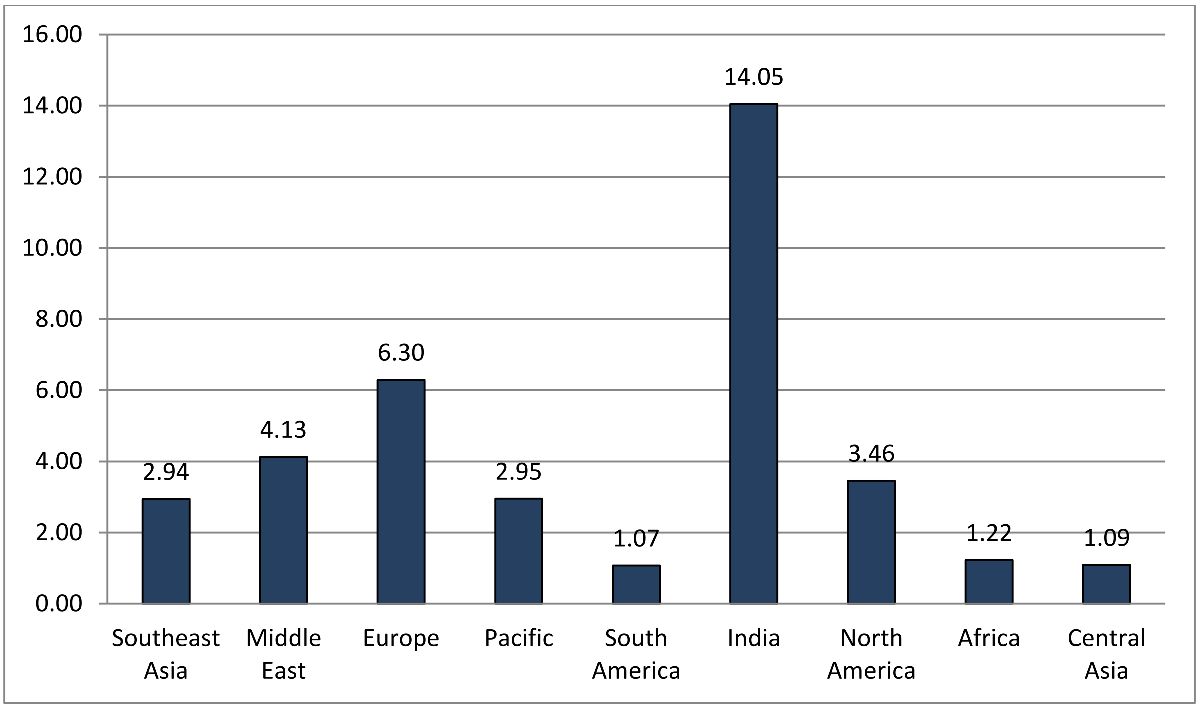
From the data I can see that India dominates the dataset, although Southeast Asia is still overrepresented compared to the other regions, it is evident that the Indian cities are dominating the dataset. I will therefore need to be careful, when I decide on parameters for my model, by using correlation. This is because random parameters which are common in the Indian cities (for example, number of Hindi speakers) might look like they are important factors when deciding on where to invest in Design, Development and Testing, when in reality there is no causation.

The easy way to deal with this problem is to exclude India from the dataset, but I believe that this would be a mistake. Omitting the country from the data because it has been successful at attracting investment in Design, Development and Testing could end up being counterproductive, because the country most likely possesses some factors which are important for investors who are making such locations decisions.

Another thing which is important to investigate is whether or not the cities in Southeast Asia and India are attracting more investments simply because they are larger than the ones in

Europe and the US. An easy way of determining whether or not this is the case is to take a closer look at investments per capita.

Figure 12 Amount of Greenfield investments in Design, Development and Testing per million capita per city



Source: fDi Markets

In the above graph, it is evident that India is still overrepresented, but a new patterns emerge. Southeast Asia is no longer dominating the other regions, but is now only the 6th largest receiver of test market investments. On the other hand, Europe and the Middle East seems to do relatively better compared to the amounts of investments per city.

I therefore believe that the best measurement for success as a test market will be to look at the investment per capita (or in my case million capita) than at absolute investments in cities. By using this measurement I ensure the validity of my results. I do not show that some cities receive a larger amount of investments simply because they are larger. This does not mean that population size will not be an attributing factor to an increase in test market investments, but it does eliminate the obvious bias.

Choice of Significance Level

The next thing I'm going to do is to start to look at my data. Since I have decided to use correlations to determine my model, it is important to determine at which level I find my correlations significant. I have to choose my significance level in a way which minimises the amount of type I and II errors. Since my null hypothesis is that the correlation is 0 a type I error would lead me to incorrectly accept that a correlation is true, when in fact it is not. On the other hand a type 2 error would leave me to reject that a correlation is true, when in fact it is.

Whether or not I make type I or II errors is determined by the level of significance which I choose. By choosing a significance level (p-value) which is very low I open up for type I errors on the contrary choosing a significance level which is too high I am vulnerable to type II errors. With this in mind I have chosen a significance level (p-value) of 5 %.

Having chosen both a significance level and a choice of measurement from which I correlate my data against, the correlations are relatively simple to calculate.

Bayesian Network

Ones I have found the variables which are significantly correlated with Investments in Design, Development and Testing per million capita, I will test the variables for causality. As described in the theoretical part of there is two ways I can approach this. I have the choice between Granger Causality and Causal Bayesian Networks. However I have chosen to use causal Bayesian Networks, because of the data requirements for the Granger Causality test.

This gives me some challenges such as what type of software to use to solve the problem, how I organize the data to fit Bayesian Networks and how do I solve a problem which is NP-Hard.

Choice of Software

There are several software packages which have been developed to deal with Bayesian Network statistics. Some of these are from the Open Source environment and others are protected under various copyright agreements. I have chosen to use open source software for two main reasons.

The first reason is that there is open access to the source code, which means that if I am forced to do some changes to the code, this would be allowed and fully legal. Furthermore open access to the source code, gives me an insight in how the different algorithms have been written and applied in the software.

The second reason is a matter of resources. Most of the Open Source software packages are made available free of charge, which eliminates the need to buy expensive software.

Ones I had decided to go with open source software, I had to decide which software package to use. I quickly narrowed my search to two different software packages. Both were add-on packages for the R environment. This meant that it was easy to apply and install. The two were, the packages deal[30] and bnlearn[31].

The packages had many similarities, but in the end I chose the bnlearn package. After trying both I was much more comfortable with using bnlearn, that as well as finding more support for the bnlearn package made me choose it.

[30] http://cran.r-project.org/web/packages/deal/deal.pdf

[31] http://cran.r-project.org/web/packages/bnlearn/bnlearn.pdf

Choosing a search algorithm

There are several types of algorithms which are able to help me solve my problem of finding the best fitting Bayesian Network. They are; the *Constraint Based Algorithms* and *Score Based Algorithms* (also called the Bayesian Approach). The two types of algorithms can both produce unique DAG's, but both have different approaches to the problems, and will therefore in some cases return different solutions.

The Constraint Based Algorithms[32] works by creating dependencies in the cases where it cannot significantly reject independency between two variables. That makes the Constraint Based approach very efficient and computational feasible.

The second approach is by using Score Based Algorithms[33]. In theory, score based algorithms works by ranking every possible DAG using a scoring function. The scoring is a function of, the amount of observations, nodes and the set of parents in the given DAG.

However in situations with larger datasets, the amount of possible DAG's becomes very large (as shown earlier) and such computations become unfeasible[34]. In this case, the best one can do is to use heuristic search algorithms. The heuristic search algorithm does not guarantee the optimal DAG, but it does return a solution which is "reasonably"[35] close to the optimal solution. Nonetheless they have the clear advantage of being a lot more computationally feasible.

In my research I have found three advantages which the Score Based Method holds, which are not applicable to the Constraint Based Approach[36]. They are:

1. The Score Based approach easily avoids making wrong decisions when it comes to the conditional independencies, where the Constraint Based approach is more susceptible to such errors in smaller data sets. That is because the Score Based approach uses model averaging in the case of small data sets.

[32] Learning Bayesian Networks, Chapter 10, Richard Neapolitan
[33] Learning Bayesian Networks, Chapter 8, Richard Neapolitan
[34] Learning Bayesian Networks, Chapter 9, Richard Neapolitan
[35] As expressed by R. Neapolitan in his book.
[36] Heckerman et al (1999)

2. The Constraint Based Method cannot handle missing data, where the Score Based can.

3. The Score Based Method can distinguish models which the Constraint Based cannot, because it cannot significantly reject independencies.

Although the Score Based approach is superior to the Constraint Based method, in many cases, the Constraint Based methodology has traditionally been used more in practice. This was mainly to do with it being more computationally feasible. However the rise in computing power, has made the use of Score Based Algorithms more popular.

Because of this I have chosen to use a Score Based Algorithm, and need only to decide which algorithm to use.

In *bnlearn* there are two available score based algorithms, they are; Hill-Climbing and Tabu Search.

Hill-Climbing is a technique, which starts with a random solution, and from there on tries to find a better solution by incrementally changing elements of the scoring algorithm[37]. The algorithm will then change solution, whenever the changes made, increases the score given by the scoring function. Yet because the Hill-Climbing algorithm only performs small incremental changes to the solution, it is susceptible to returning local optima as the solution.

On the other hand I have the Tabu Search, which like the Hill-Climbing algorithm also uses incremental changes to the current solution in order to search for a better solution[38]. However it deals with the local optima problem, by applying a memory function of previously visited solutions, and there by allows it to escape the local optima.

Because of this property, I have chosen to use the Tabu Search Algorithm, in my thesis. However it is important to stress that none of these algorithms are guaranteed to present the global optimum, but merely a results which is "reasonably" close.

[37] Hill-climbing Search, Bart Selman et al
[38] Tabu Search – Part 1, Fred Glover, 1989

Implementation of Analysis in R

Now that I have decided on how to conduct the analysis, all I have to do is to implement it in R. I have developed a small piece of code, which should enable any reader to easily perform the same analysis.

The first thing I need to do is to install the proper packages if they are not present. The first package I need to install is the bnlearn package, which is available from the CRAN (R repository), I can also install the package snow from CRAN. However I also need to get the packages grid and Rgraphviz (for better graphic opportunities), these packages comes from the bioConductor repository and are not official R packages. The packages are installed and loaded with the following code.

```
Install.packages(bnlearn)
Install.packages(snow)
Install.packages(graph)

source("http://bioconductor.org/biocLite.R")
biocLite()

biocLite(c(“grid”,”Rgraphviz”))

library(bnlearn)
library(snow)
library("Rgraphviz")
```

The next thing I do is to load and diagnose the data, I do that using the read.csv() function for loading and str() for diagnostics. In the first round I want to load the full data set

```
fdidata <- read.csv("FilePathForFullDataSet", header=TRUE,
colClasses=c("factor"))
str(fdidata)
```

Once the full data set has been loaded and diagnosed, and if the diagnose have returned all values as numeric values, I am are able to find the parameters which are relevant for my analysis, this is done by choosing FDI in Design, Development as the target variable, and determining all parameters which shows dependence on it.

I this by using the relevant() function which is available in bnlearn. The relevant function determines all relevant nodes, even if those nodes are not significantly so. Therefore for a node to be discarded it is not enough that the node is just weakly dependent to be omitted by the analysis.

This means that I am more exposed to type I errors, however since I only use this function to minimise my data set so that it becomes computational feasible, it does not impose a significant risk to my result. The code is written in the following way.

```
relevant("FDI in Design, Development and Testing", data = fdidata)
```

Having determined the relevant parameters which are contained in the Markov Blanket containing FDI in Design, Development and Testing, I rebuild my data set to only contain these parameters. Once this is done I reload it and do the diagnostics again, like before:

```
fdidata <- read.csv("FilePathForDecreasedDataSet", header=TRUE,
colClasses=c("factor"))
str(fdidata)
```

The next step is when I learn the structure of a Bayesian Network containing only the relevant parameters. As determined before I will use the Tabu Search algorithm to do so. I call the Tabu Search algorithm for the data in the following way:

```
bn.ts <- tabu(fdidata)
bn.ts
```

Now that I have determined the structure, I just need to plot it. To give the best graphical overview, I use Rgraphviz to do so in the following:

```
graphviz.plot(bn.ts)
```

Having developed this code, I can now find the right parameters, for my analysis.

Regression

Once I have the result from my Bayesian Network, it is important for me to find out how much of the variations in FDI in Design, Development and Testing. I have chosen a reasonably simple way to test for this. The way I do it is by conducting a linear regression of a model consisting of the parameters from the Bayesian Network. I will then test for significance of the individually parameters and for the r^2 of the entire regression.

By doing so, I can either reject or accept the hypothesis that the parameters in the model can predict the variations in FDI in Design, Development and Testing.

Methodological Interim Conclusion

In this part I have decided to look at the amount of FDI projects per millionth capita; I did this for two main reasons. First of all I noticed that India was extremely over represented, this was partly because of the large populations in Indian cities. Second the data is structured in a way that FDI projects is the most relevant approach to take.

Furthermore I decided to use correlations and the relevant() function in bnlearn to restrict the amount of parameters from which I then built a Bayesian Network. The Bayesian Network is built to show which parameters that might show causality with FDI in Design, Development and Testing.

In the end I will perform a regression to test the validity of my results.

Results

Correlations

The first thing I had to do was to calculate the correlations of all the different parameters, so that I can narrow my search for causality. But it was not just important to determine whether or not there is a correlation, it is perhaps more relevant to test whether or not this correlation is statistically significant.

In the two tables below I have listed all correlation which is significant at the 5% level.

Category	Correlation	P Value
Shared Services Centre (inward FDI)	0.4638	0.0000
Technical Support Centre (inward FDI)	0.4515	0.0000
Design Development and Testing (inward FDI)	0.4328	0.0000
Research and Development (inward FDI)	0.3481	0.0000
Exports of commercial services (%)	0.3387	0.0000
Customer Contact Centres (inward FDI)	0.3209	0.0000
ICT & Electronics (inward FDI)	0.2719	0.0000
Life Sciences (inward FDI)	0.2681	0.0000
Real corporate taxes do not discourage entrepreneurial activity	0.2510	0.0000
Corporate tax rate on profit	-0.2398	0.0000
Software for life sciences (inward FDI)	0.2361	0.0000
Business support services (inward FDI)	0.2333	0.0000
International Trade to GDP ratio	0.2198	0.0001
South Asia	0.2193	0.0001
Chemicals (% of value added in manufacturing)	0.2079	0.0002
Urban population (%)	-0.2064	0.0002
Imports of goods & commercial services (% of GDP)	0.2025	0.0003
National culture is open to foreign ideas	0.2015	0.0003
Business impact of rules on FDI	0.1984	0.0004
Exports of goods and services (% of GDP)	0.1975	0.0004
Investment incentives are attractive to foreign investors	0.1967	0.0004
Trade (% of GDP)	0.1931	0.0005
Forest area (% of land area)	-0.1915	0.0006
Transport Equipment (inward FDI)	0.1889	0.0007
No. days to enforce a contract	0.1878	0.0007
R&D (inward FDI)	0.1867	0.0008
Public sector contracts are sufficiently open to foreign bidders	0.1848	0.0009
Aerospace (inward FDI)	0.1823	0.0011
Specialisation in hotels and tourism	0.1803	0.0012
Manufacturing value added (annual % growth)	0.1761	0.0016
State ownership of enterprises is not a threat to business activities	0.1731	0.0019
Travel services (% of commercial service exports)	-0.1716	0.0021
Computer communications and other services (% of commercial service exports)	0.1712	0.0022
Headquarters (inward FDI)	0.1712	0.0022
Subsidies do not distort fair competition and economic development	0.1711	0.0022
Flexibility and adaptability	0.1691	0.0025
Wind power as a percentage of renewable electricity generated	0.1680	0.0027
Travel services (% of commercial service imports)	-0.1674	0.0028
Construction (inward FDI)	0.1670	0.0028
Immigration laws do not prevent your company fr...	0.1648	0.0032
Waste management & remediation services (inward ...	0.1616	0.0039
Exports of goods per capita	0.1606	0.0041
Environmental Technology (inward FDI)	0.1595	0.0044
Exports of goods (% of GDP)	0.1593	0.0044
Science in schools is sufficiently emphasized	0.1566	0.0052
Healthcare (inward FDI)	0.1564	0.0052
Protectionism does not impair the conduct of your...	0.1548	0.0057
Skilled labour	0.1495	0.0076
Software & IT services (inward FDI)	0.1480	0.0083
Exports of goods and services (annual % growth)	0.1475	0.0085
Arable land (% of land area)	0.1459	0.0092
Space and Defence (inward FDI)	0.1454	0.0095
Relocation of services is not a threat to the future...	-0.1447	0.0098
Agricultural productivity (PPP)	-0.1446	0.0099
Productivity in industry (PPP)	-0.1441	0.0102
Senior Scientist	-0.1436	0.0104
R&D Team Leader	-0.1434	0.0105
Civil engineer	-0.1434	0.0105
Geologist	-0.1434	0.0105
Industrial engineer	-0.1434	0.0105
Mining Engineer	-0.1434	0.0105
Soundness of banks	-0.1432	0.0106
Top corporate tax rate	-0.1430	0.0107
Accounting tax preparation bookkeeping & payroll...	0.1421	0.0112
Economic sectors / Services (% of GDP)	-0.1415	0.0116
Availability of Competent senior managers	0.1395	0.0128
Engineering Manager	-0.1390	0.0131
Chief Engineer/Technical Mgr	-0.1390	0.0131
Chief Scientist/Technologist	-0.1390	0.0131
Food Beverages and Tobacco (inward FDI)	0.1380	0.0138
International experience of Senior Manager is impo...	0.1370	0.0145
Electrical Engineer	-0.1367	0.0147
Laboratory Specialist	-0.1367	0.0147
Senior Engineer	-0.1367	0.0147
Senior Technical Drawer	-0.1367	0.0147
Laboratory Manager	-0.1347	0.0162

Category	Correlation	P Value
Principal Engineer	-0.1347	0.0162
Chemical Engineer	-0.1345	0.0163
Productivity in services (PPP)	-0.1345	0.0163
Solar photovoltaics and solar thermal electricity g…	-0.1332	0.0174
Ease of doing business	0.1331	0.0175
Creation of firms	0.1306	0.0196
The legal and regulatory framework encourages th…	0.1286	0.0216
Specialisation in coal oil and gas	0.1285	0.0216
Forest area (sq. km)	-0.1277	0.0225
Financial cards in circulation pr capita	-0.1275	0.0227
Companies in aircraft parts and auxilliary equipment	0.1260	0.0242
Equal opportunity legislation in your economy enco…	0.1260	0.0242
Bureaucracy does not hinder business activity	0.1249	0.0255
Specialisation in engines and turbines	0.1243	0.0261
Biological products (except diagnostics) (inward…	0.1243	0.0261
The educational system meets the needs of a com…	0.1243	0.0262
Imports of electricity	-0.1241	0.0263
Population over 65 years (%)	-0.1241	0.0264
Adaptability of government policy	0.1240	0.0265
Architect	-0.1234	0.0271
Exports of goods ($bn)	-0.1225	0.0283
Consumer Goods (inward FDI)	0.1224	0.0284
Financial Services (inward FDI)	0.1221	0.0288
Transport services (% of commercial service imports)	-0.1214	0.0297
Cost-of-living index	-0.1203	0.0311
Railways passengers carried	0.1202	0.0312
South America	-0.1201	0.0314
Size of Labor Force	0.1192	0.0325
Attitudes toward globalization are possitive	0.1191	0.0327
Head of Research and Development	-0.1190	0.0328
Cyber security is being adequately addressed by c…	0.1190	0.0329
Average hours of sunshine per day	-0.1183	0.0339
Government consumption expenditure ($bn)	-0.1177	0.0347
Cost of establishing a business	-0.1173	0.0354
Employment services (inward FDI)	0.1172	0.0354
Pharmaceutical preparations (inward FDI)	0.1171	0.0356
Patent applications residents	-0.1167	0.0362
Aerospace MRO (inward FDI)	0.1165	0.0365
Cost to import	-0.1161	0.0371
Rigidity of Employment Index	-0.1157	0.0378
Video games applications and digital content (inwa…	0.1156	0.0378
Total general government debt ($bn)	-0.1154	0.0382
Senior Technician	-0.1148	0.0392
Communications equipment (inward FDI)	0.1147	0.0393
Specialisation in engineering services	0.1147	0.0393
Environmental Consultant	-0.1143	0.0399
Lighting Technician	-0.1143	0.0399
Production Sound Mixer	-0.1143	0.0399
Quantity Surveyor	-0.1143	0.0399
Clinical Research Associate	-0.1141	0.0402
Engineer	-0.1141	0.0402
Scientist	-0.1141	0.0402
Technical Drawer	-0.1141	0.0402
Management education meets the needs of the busi…	0.1141	0.0403
Irrigated land	0.1131	0.0418
Europe	0.1130	0.0420
Merchandise exports	-0.1127	0.0426
Communications (inward FDI)	0.1126	0.0427
Exports of goods & commercial services ($bn)	-0.1124	0.0432
Need for economic and social reforms are well und…	0.1120	0.0438
Assistant Engineer	-0.1118	0.0440
Laboratory Technician	-0.1118	0.0440
Assistant Scientist	-0.1112	0.0453
Investment management (inward FDI)	0.1106	0.0463
World exports contribution (%)	-0.1101	0.0472
Gross fixed capital formation ($bn)	-0.1095	0.0482
Food production index	-0.1095	0.0484
Employer's social security contribution rate	-0.1088	0.0497

One of the things which is very evident from the results is that one of the biggest things which are correlated in FDI in Design, Development and Testing is FDI in other sectors. This suggests, that the attraction of FDI in Design, Development and Testing, is not that different from attraction of FDI in general.

However given the Common Cause Principle I cannot determine whether or not investments drives other investments (like it would in the case of a positive feedback loop) or that there is common causes for all types of investments.

Another thing which seems to have an effect is both the perception of taxes and the corporate tax level. Both are significantly negatively correlated with amount of investments in Design, Development and Testing. That could suggest that Taxation does have a significant effect on the quantity of Investments in Design Development and Testing.

Again this could be due to a third common cause, such as a business friendly government. However a significant effect on investments from taxation would be coherent with economic theory, which argues that a cost benefit analysis (Practically represented as a Net Present Value Analysis). Such an analysis would be affected by the level of taxation, because it affects the cash flows which are measured.

Nevertheless, evidence for taxation having a significant standalone effect on FDI in Design, Development and Testing has not been convincingly empirically proved so far[39].

Another thing I can see from the table is that being open to foreign investors and having an open economy, is significantly positively correlated with FDI in Design, Development and Testing. This is coherent with basic economic intuition. It simply makes sense that companies prefer to invest in economies which are open. Furthermore, it is most likely practically less demanding for them to make investments in such countries.

The easiness of investing also seems to be significantly positively correlated. This supports my previous statement. What I see is that the ease of doing business and a flexible environment is significantly correlated.

[39] How tax policy and incentives affect foreign Direct Investment a review, Morisset et al, 1999

Something I also notice is that the cost of doing business is significantly negatively correlated with FDI in Design, Development and Testing. I see that both salaries and cost index' are negatively correlated with investments.

The above mentioned observations are all very intuitive; however there is also counterfactual and illogical observations. For instance I see that high exports in Commercial Services are positively correlated with FDI in Design, Development and Testing, however I find that having a large services sector is negatively correlated with the same type of FDI.

I also see that the soundness banks are negatively correlated with FDI in Design, Development and Testing. This is not intuitive on the surface, because one would in general relate the soundness of banks with a less risky economy.

Carrying less market risk is normally good for investors which carry high business risk which is normally present with this type of FDI. However my preliminary results show the opposite to be true.

Having shown the different correlations is not enough to show causation, as discussed in the theoretical framework. It is therefore necessary to conduct a Causal Bayesian Network analysis to show if there is any causal relationship between the above parameters and Investments in Design, Development and Testing per million capita.

Bayesian Network

Having determined the parameters which are significantly correlated with FDI in Design, Development and Testing, I can proceed to find the Markov Blanket, as mentioned in the in the methodology I can do this by using the relevant() function which is included in the bnlearn package. Calling the function I get the following parameters:

Parameter	Correlation	Significance (P-value)
Chief Engineer/Technical Manager	-0.1390	0.0131
No. days to enforce a contract	0.1889	0.0007
Soundness of Banks	-0.1432	0.0106
Financial Cards in Circulation	-0.1275	0.0227
Imports of goods and Commercial services (% of GDP)	0.2025	0.0003
Science in schools is sufficiently emphasised	0.1566	0.0052
Management education meets the needs of the business community	0.1141	0.0403
Real corporate taxes do not discourage entrepreneurial activity	0.2510	$4.5366e^{-06}$
The legal and regulatory framework encourages the competitiveness of enterprises	0.1286	0.0216
Adaptability of government policy	0.1240	0.0265
Bureaucracy does not hinder business activity	0.1249	0.0255
Investment incentives are attractive to foreign investors	0.1984	0.0004
Ease of doing business	0.1331	0.0175
Equal opportunity legislation in your economy encourages economic development	0.1260	0.0242
Cyber security is being adequately addressed by corporations	0.1190	0.0329

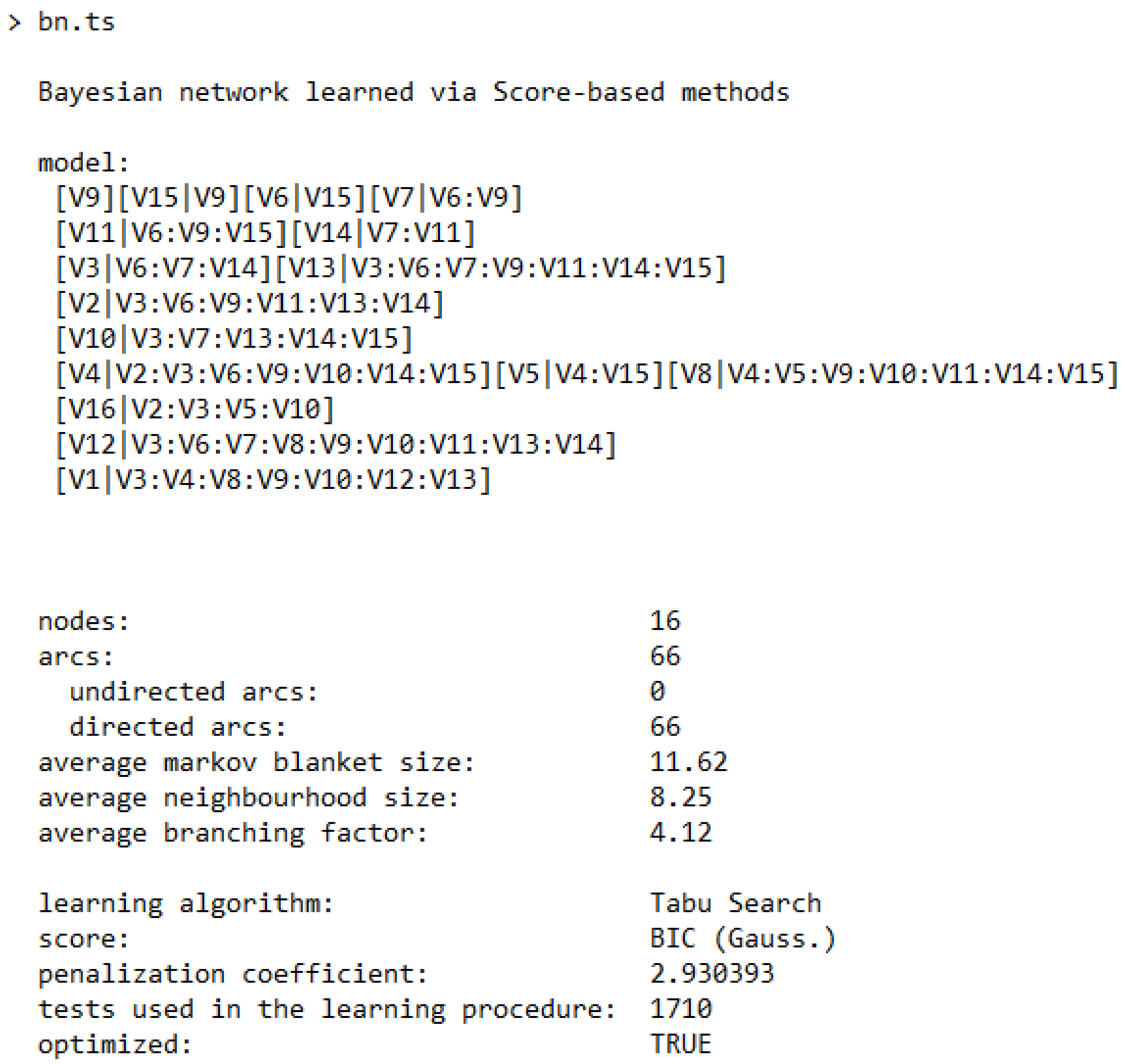
Having 16 (when including FDI in Design, Development and Testing) parameters from which I can recreate a smaller data set, which is more computational feasible, since I have gone from 1,038,459,371,769,700,000,000,000,000,000,000 (2^113) possible DAGS to 65,536 (2^16).

What is shown is that the relevant factors are Financial Incentives (Low Cost, Low Taxes and Subsidies), Prober Legal Framework, Technological Ability and Talent.

It is interesting to see, that the parameters on the list takes place in front of parameters who have a correlation with FDI in Design, Development and Testing per million capita which are higher. It is however in line with the idea that the correlations, however strong they might be, does not necessarily entail causality.

Having isolated the 16 parameters which are in a Markov Blanket with FDI in Design, Development and Testing, I can now calculate the structure of the important Bayesian Network.

Using, the tabu search algorithm in R I get the following output.



```
> bn.ts

  Bayesian network learned via Score-based methods

  model:
   [V9][V15|V9][V6|V15][V7|V6:V9]
   [V11|V6:V9:V15][V14|V7:V11]
   [V3|V6:V7:V14][V13|V3:V6:V7:V9:V11:V14:V15]
   [V2|V3:V6:V9:V11:V13:V14]
   [V10|V3:V7:V13:V14:V15]
   [V4|V2:V3:V6:V9:V10:V14:V15][V5|V4:V15][V8|V4:V5:V9:V10:V11:V14:V15]
   [V16|V2:V3:V5:V10]
   [V12|V3:V6:V7:V8:V9:V10:V11:V13:V14]
   [V1|V3:V4:V8:V9:V10:V12:V13]


  nodes:                                 16
  arcs:                                  66
    undirected arcs:                     0
    directed arcs:                       66
  average markov blanket size:           11.62
  average neighbourhood size:            8.25
  average branching factor:              4.12

  learning algorithm:                    Tabu Search
  score:                                 BIC (Gauss.)
  penalization coefficient:              2.930393
  tests used in the learning procedure:  1710
  optimized:                             TRUE
```

What I can see from the Bayesian Network Analysis is that FDI in Design, Development and Testing (Shown as V16) has four parents, which are:

- No. of days to enforce a contract (V2)
- Soundness of banks (V3)
- Imports of goods & commercial Services (% of GDP) (V5)
- Adaptability of government policy (V10)

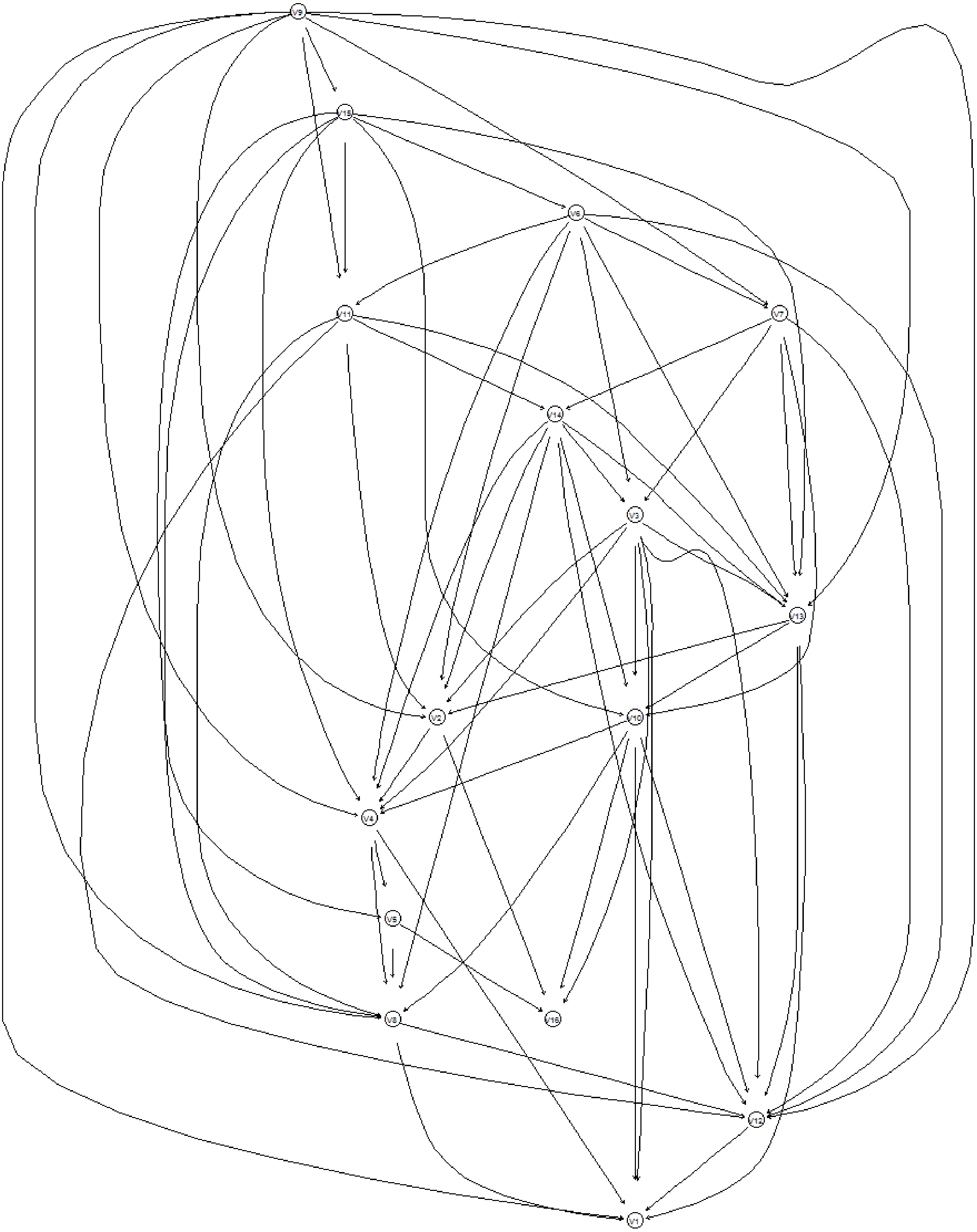
That means that I can calculate the probability for an increase in FDI in Design, Development and Testing using, only these 4 parameters.

I can show the whole network in graph, by using the RGraphviz package, as shown below.

The parameters have been renamed for aesthetic reasons, they are:

Old Name	**New name**
Chief Engineer/Technical Mgr	V1
No. days to enforce a contract	V2
Soundness of banks	V3
Financial cards in circulation pr capita	V4
Imports of goods & commercial services (% of GDP)	V5
Science in schools is sufficiently emphasized	V6
Management education meets the needs of the business community	V7
Real corporate taxes do not discourage entrepreneurial activity	V8
The legal and regulatory framework encourages the competitiveness of enterprises	V9
Adaptability of government policy	V10
Bureaucracy does not hinder business activity	V11
Investment incentives are attractive to foreign investors	V12
Ease of doing business	V13
Equal opportunity legislation in your economy encourages economic development	V14
Cyber security is being adequately addressed by corporations	V15
Investment in D, D and T per million person	V16

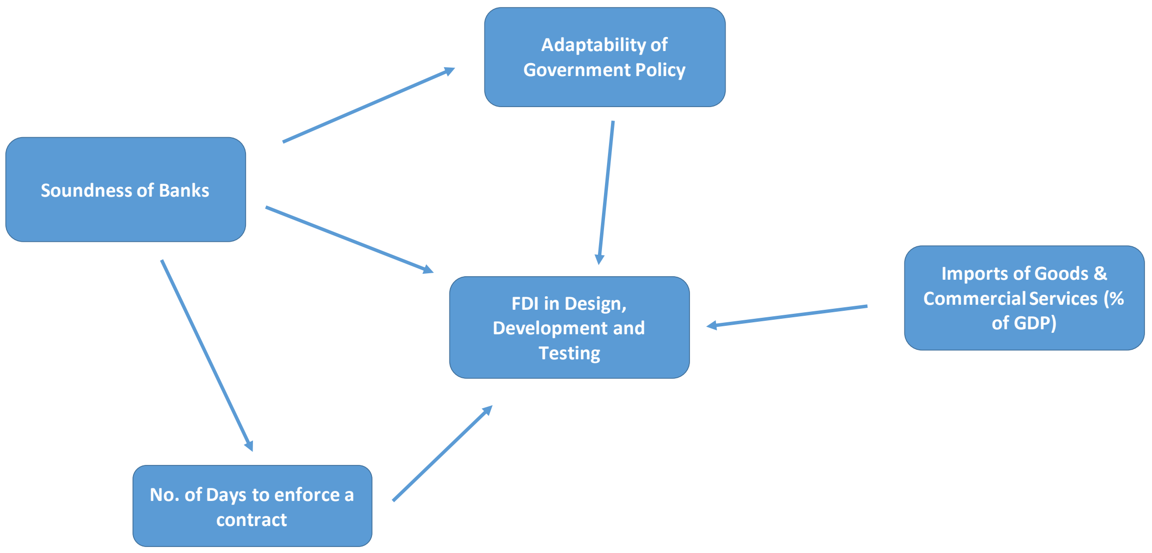
Figure 13 Bayesian Network depicting the Markov Blanket



The above DAG shows the whole Markov Blanket, which were calculated by the bnlearn package, however it can be a bit confusing to look at I have therefore decided to break it down into smaller pieces.

In the next DAG, I decided to only look at the parameters which are conditionally dependent with FDI in Design, Development and Testing.

Figure 14 DAG of FDI in D, D, T and Conditional Dependent Parameters


Adaptability of Government Policy
Soundness of Banks
FDI in Design, Development and Testing
Imports of Goods & Commercial Services (% of GDP)
No. of Days to enforce a contract

One thing which is evident, is that it is a complex network. The parameters which seems to cause FDI in Design, Development and Testing per million capita also seems to be affecting each other.

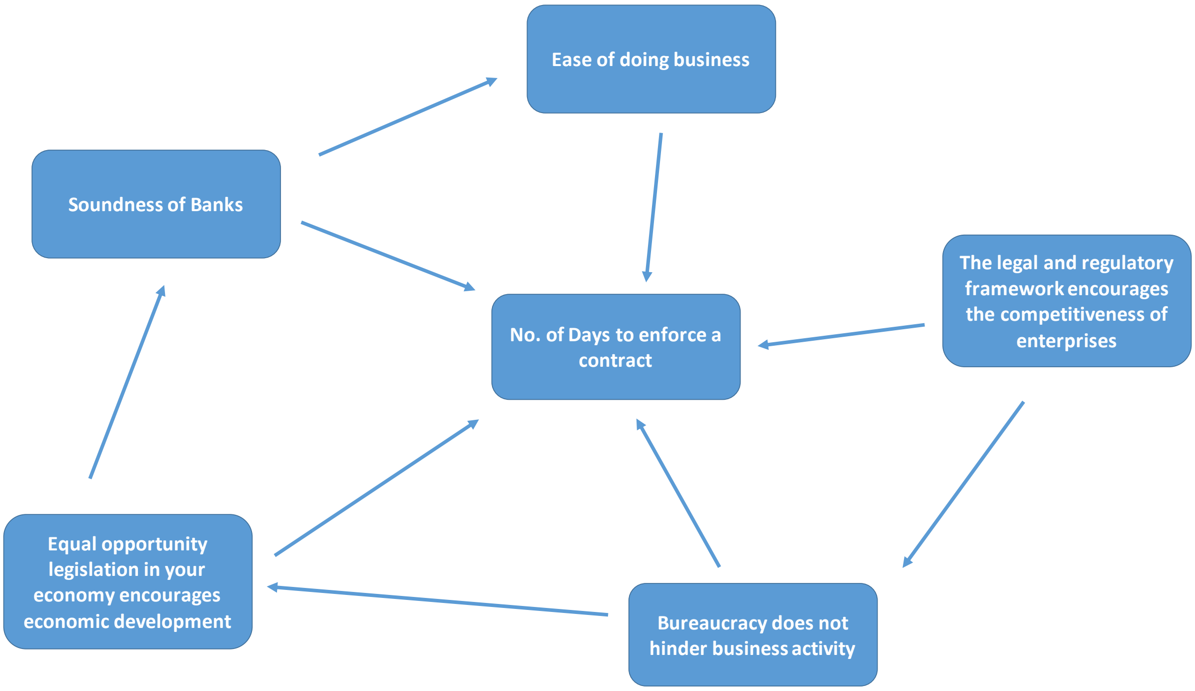
However other than that, this DAG does not give me much more information than what I learned before. Something which could give me more information about the structure of the Network is to look at the conditional dependencies of each of those four parameters.

If I look at the No. of Days to enforce a contract, I find that there are six parameters which are effecting it, they are:

- Soundness of Banks (V3)
- The legal and regulatory framework encourages the competitiveness of enterprises (V9)
- Bureaucracy does not hinder business activity (V11)
- Ease of doing business (V13)
- Equal opportunity legislation in your economy encourages economic development (V14)

Knowing the different potential causes of No. of Days to enforce a Contract, is however not enough to gain insight in the causality structure. To gain this insight it is necessary to present the data in a DAG, this is done like so:

Figure 15 DAG of conditional dependencies with No. of days to enforce a contract


Ease of doing business
Soundness of Banks
No. of Days to enforce a contract
The legal and regulatory framework encourages the competitiveness of enterprises
Equal opportunity legislation in your economy encourages economic development
Bureaucracy does not hinder business activity

What I quickly see is a more complex structure than the one I get from just observing the dependencies of No. of days to enforce a contract. It is evident that the different parameters don't just have an influence the No. of days to enforce a contract.

I see the same pattern with the other parameters which has an effect on FDI in Design, Development and Testing.

This indicates that looking for simple causes to what might drive investments in Design, Development and Testing, might not be the right approach. However it does look like, the causes of FDI in Design, Development and Testing are all part of a complicated network.

For instance, even though I can show, that a parameters such as *The legal and regulatory framework encourages the competitiveness of enterprises*, does not help me calculate the probability of an increase in FDI in Design, Development and Testing.

I do not believe that the parameter should be ignored. It is most likely the case that countries which have an efficient structure to deal with the enforcement of contracts, are the same countries which have regulatory and legal framework which encourages the competiveness of businesses.

Another example of the same thing could be the *ease of doing business* parameter (V13). Although the ease of business is conditionally independent from FDI in Design, Development and Testing, I see that it influence both *No. of days to enforce a contract* (V2) as well as *Adaptability of Government Policy* (V10).

Knowing this I cannot suggest that the ease of doing business in a country, does not have an effect on the amount of FDI in Design, Development and Testing. Even though it does not directly cause the FDI, it does create an environment where companies are more likely to invest.

With this in mind two conclusions can be made.

1. There are certain parameters which increases the likelihood of FDI in Design, Development and Testing. Therefore if a country works on improving these parameters, it should have a positive influence on the FDI inflow.

2. Although certain parameters can be determined as conditionally dependent, it is not enough to only work on them. The parameters which I have identified are all interconnected in a complex network. I found that the parameters which are conditionally dependent on FDI in Design, Development and Testing, themselves are dependent on a lot of different parameters. I believe that there might be two reasons for this. The first is that it is necessary to be strong in a lot of different parameters, which therefor all have an effect on the attraction of FDI. The second would be that different cities have chosen different strategies which are all successful. Since a Bayesian Network is not able to make this distinction the network starts looking very complex.

I believe that understanding the fact that investment decisions are not made because of a few parameters but rather has several interconnected causes is key to understanding what drives investments.

Because of this I have chosen to include all parameters, from the network in my regression. I have done this because I believe that is necessary to examine all variables in this way.

Regression

I have chosen to run the Regression in Microsoft Excel. Although there is other statistical packages which are both more advanced and sophisticated, Microsoft Excel was chosen because of simplicity.

I do not need a very advanced regression analysis in this case. I am mostly interested in the r^2 of the regression and of the individual significance levels of the different parameters.

This is because I do not intend to use the regression to do any type of predictive analysis. However it is relevant for me to see how much of the variations in FDI in Design, Development and Testing can be described by a model, building on the parameters which were identified in the Bayesian Network.

Running the regression gives me the following output:

Regression Statistics	
Multiple R	0.45133
R Square	0.20370
Adjusted R Square	0.16804
Standard Error	6.91403
Observations	351

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	15	4096.5636	273.1042	5.7130	0.0000
Residual	335	16014.2951	47.8039		
Total	350	20110.8587			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-2.8803	4.9418	-0.5828	0.5604
Chief Engineer/Technical Manager	0.0000	0.0000	-0.0891	0.9291
No. days to enforce a contract	0.0111	0.0025	4.5005	0.0000
Soundness of banks	-2.9265	0.5871	-4.9850	0.0000
Financial cards in circulation pr. capita	0.1227	0.3570	0.3436	0.7314
Imports of goods & commercial services (% of GDP)	0.0285	0.0205	1.3920	0.1649
Science in schools is sufficiently emphasized	0.9208	0.6457	1.4261	0.1548
Management education meets the needs of the business community	-1.6654	0.7856	-2.1199	0.0347
Real corporate taxes do not discourage entrepreneurial activity	0.9707	0.7894	1.2297	0.2197
The legal and regulatory framework encourages the competitiveness of enterprises	0.1921	1.0988	0.1749	0.8613
Adaptability of government policy	0.8972	0.8065	1.1126	0.2667
Bureaucracy does not hinder business activity	0.4999	0.9946	0.5026	0.6156
Investment incentives are attractive to foreign investors	0.5028	0.8370	0.6006	0.5485
Ease of doing business	-0.3154	1.2020	-0.2624	0.7932
Equal opportunity legislation in your economy encourages economic development	0.7980	1.1145	0.7160	0.4745
Cyber security is being adequately addressed by corporations	0.5013	0.8593	0.5834	0.5600

The parameters from the regression shows me that the model does not fit the data very well.

First of looking at the r^2 I can determine that the model only predicts 20 % of the variations in FDI in Design, Development and Testing. That means that 80 % of the variations can be attributed to variables which are not included in the model.

Another thing which is evident, is that the only 3 of the individual parameters are significant at the 5 percent level they are:

Parameter	**Coefficient**	**Significance (P-Value)**
No. days to enforce a contract	0.0025	0.0000
Soundness of banks	0.5871	0.0000
Management education meets the needs of the business community	-1.6654	0.0347

This suggests that I cannot say with confidence that any of the other variables have a significant influence on FDI in Design, Development and Testing.

This poses a problem, since the parameters *Imports of goods & commercial services (% of GDP)* and *Adaptability of government policy* are not shown to have a significant influence in the regression model, but is listed as a parent in the Bayesian Network. That does intuitively seem to contradict one another. But examining the fundamentals of both models shows me that this is not necessarily the case.

In the Bayesian Network I identify the parameters which contains information on the probability of an increase in FDI in Design, Development and Testing per million capita. This leads me to a number of parents which are enough to predict the probability, so that I can ignore the other parameters.

However in the regression analysis I build a model where I try to predict the amount of FDI in Design, Development and Testing per million capita, given several parameters. In this particular applications of the regression analysis I do not attempt to determine which parameters that contains the most information about FDI in Design, Development and Testing. I simply try to build a model which fits the data the best, and gives the highest r^2. Therefore the individual significance of the parameters are not important.

In the end the regression tells me that the combination of those particular parameters can predict FDI in Design, Development and Testing, and whether or not the coefficient of those parameters should be statistically significantly different from zero.

Robustness of Results

In this section I will describe the reliability of the analysis I have conducted. I will especially focus on the two main assumptions of Stability & Minimality and Causal Sufficiency

Stability & Minimality

As mentioned in the theoretical section, it is necessary for the Bayesian Network to be both stable and minimal. This is because of two different reasons

1. If the structure of the network changes, because of minor changes in the underlying data, it can be hard to prove any type of consistent causal relationship.

2. If I do not include the minimum amount of Markovian Parents, I have not found the true causes of the child.

I believe that these conditions have been tested sufficiently and proven to hold. This is because I have made changes in both the data and the learning algorithms, and have still gotten the same structure of the network.

Causal Sufficiency

The Causal sufficiency assumption state that all relevant variables must be included, before I can go from conditional probabilities to real causal relationship.

This assumption is quite intuitive, however it can be quite difficult to live up to in practise. The reason that an analysis of relationships is initiated in the first place is that one does not know which parameters that are relevant. Therefore the best one can do is to include as many parameters one would find relevant. Nonetheless in most situation there will still be a suspicion that other variables might have an influence on the parameter in question.

Because of this I cannot be completely assured that my results are depicting real causal relationship and not just conditional probabilities.

Practical Implications

The analysis presents different parameters which could have a causal relationship with FDI in Design, Development and Testing. However it can be difficult to see how it can be applied. I will use this section to show how my analysis can be applied and see what insights that can be gained from the data.

The first thing I want to look at is the top ten cities when it comes to Foreign Direct Investments in Design Development and Testing per million capita. So far in the analysis I have looked at the data in a global perspective, but in most cases Copenhagen Capacity does not compete on a global level. They are however mostly competing with other European cities. I have therefore decided to focus on Europe to make it more relevant for Copenhagen Capacity. Looking at the top 10, it is very evident that Ireland and the United Kingdom are dominating the list. What is even more interesting is that every city in the top five is located on the Irish Island. Another interesting observation is that England is not very well represented in the top 10. In fact Reading is the only English city.

Table 2 Top 10 Cities FDI in Design, Development and Testing

Rank	City	Country	Investments per million capita
1	Belfast	United Kingdom	57.14
2	Londonderry (Derry)	United Kingdom	54.55
3	Galway	Ireland	42.31
4	Dublin	Ireland	36.00
5	Newry	United Kingdom	33.33
6	Aberdeen	United Kingdom	31.58
7	Reading	United Kingdom	31.25
8	Edinburgh	United Kingdom	21.33
9	Cork	Ireland	21.15
10	Krakow	Poland	20

To make the analysis more relevant, I can compare these cities with the parameters from the analysis, to see how the cities do, and if there position is justified. In the following table I have depicted the top 10 cities and there individual rankings.

Table 3 Top 10 cities and their rankings

City	Chief Engineer	No. days to enforce a contract	Soundness of banks (Descending)	Financial cards in circulation per capita	Imports of goods & commercial services (% of GDP)	Science in schools is sufficiently emphasized	Management education meets the needs of the business community	Real corporate taxes do not discourage entrepreneurial activity	The legal and regulatory framework encourages the competitiveness of enterprises	Adaptability of government policy	Bureaucracy does not hinder business activity	Investment incentives are attractive to foreign investors	Ease of doing business	Equal opportunity legislation in your economy encourages economic development	Cyber security is being adequately addressed by corporations
Belfast	119	50	25	1	72	72	41	35	12	44	38	14	25	8	60
Londonderry	160	50	25	1	72	72	41	35	12	44	38	14	25	23	75
Galway	77	19	1	131	19	56	19	5	94	5	14	1	14	8	60
Dublin	46	19	1	131	19	56	19	5	94	5	14	1	14	8	60
Newry	139	50	25	1	72	72	41	35	12	44	38	14	25	23	75
Aberdeen	55	50	25	1	72	72	41	35	12	44	38	14	25	23	75
Reading	52	50	25	1	72	72	41	35	12	44	38	14	25	23	75
Edinburgh	56	50	25	1	72	72	41	35	12	44	38	14	25	23	75
Cork	69	19	1	131	19	56	19	5	94	5	14	1	14	8	60
Krakow	159	9	139	159	35	20	122	131	120	15	125	120	124	112	30

What can be seen from the table is that even though some cities do well in some rankings, none of the cities do well overall in the rankings.

For instance evident that the Irish cities are doing really at providing investments incentives for foreign companies which are attractive, as well as having a corporate tax rate which does not deter entrepreneurship.

Both these parameters are very interesting for two main reasons.

1. The parameters both provide concrete financial incentives which are easy to relate to.

2. The parameters are also both taken from the IMD survey of top managers. In fact 11 of 15 of the parameters which have been chosen to be significant are from the IMD survey.

What Copenhagen Capacity can take from this is that, it helps to provide financial incentives, which is a quite intuitive conclusion.

However this is not a very helpful conclusion for Copenhagen Capacity, since it is not within their power to provide such financial incentives. Although they could do some lobbying, it is a very long term strategy, and it does not play on their core competencies. Nonetheless if I look deeper into the type of parameter, it is evident that it is the idea of having strong incentives or a low tax rate which are the cause of FDI.

Copenhagen Capacity could in this case chose to brand Copenhagen as a place with strong investment incentives and a low tax rate. This could be done in several ways.

The Danish government are currently implementing new taxation rules for companies, and in depth analysis of why these rules might be beneficial for the potential investors. In the same way, one could use concrete examples where investors benefited from the investments incentives which are already present in the country.

I believe that the fact that 73 % of the parameters which are believed to be causing the FDI in Design, Development and Testing are from the IMD Survey is an important observation. Especially because it only constitutes approximately 15 % of all parameters which were included in the sample. This gives Copenhagen Capacity an opportunity, prioritizing the branding efforts, by telling the right stories, and giving the potential investors an idea of Copenhagen being an attractive place for them, could benefit their results.

One of the reasons I believe this to be true is because the data suggests that the Republic of Ireland and Northern Ireland (part of the United Kingdom) occupies the top five. Doing so could suggest that the branding efforts made by the Republic of Ireland, to be seen as a more business friendly country, have had a spill over effect in Northern Ireland.

Although this is in contrast to the rational decision theory I am familiar with. Irrational decision making have been known to affect the location of corporate headquarters, factors such as key decision makers being romantically involved have been known to drive both interest in a particular countries as well as lead to concrete investments. And although those examples are no more than anecdotal evidence, it is not hard to believe that the investors had a preconditioned opinion of Ireland being a country which are attractive to investors, effected the opinion of Northern Ireland, even though they are two completely different nations.

If I continue to look at Irelands ranking another thing pops to mind. It is shown that the soundness of the Irish banks are the lowest in all of Europe (Because the ranking is sorted in descending order, number one means it is the lowest).

It can be hard to find a logical reason for why a weak financial sector should drive investments. However in a historical perspective it gives me a bit more clarity, Ireland has had a very open financial sector, which was highly dependent on the global markets. So when the financial crisis occurred in 2007, the Irish banks were particularly vulnerable.

I do not believe that the weak financial sector in itself is parameter which drives FDI in Design, Development and Testing, however it is merely a dummy variable which provides insights in to how exposed to international financial risk the countries were. It would usually be the case that companies which are open to international investments are more exposed to this type of risk.

Although looking at the top 10 list does give me some insight, is also necessary to see how well Copenhagen does in the rankings.

The table below shows that Copenhagen does relatively well when it comes to how much FDI in Design, Development and Testing per million capita they receive.

Table 4 Ranking of Copenhagen

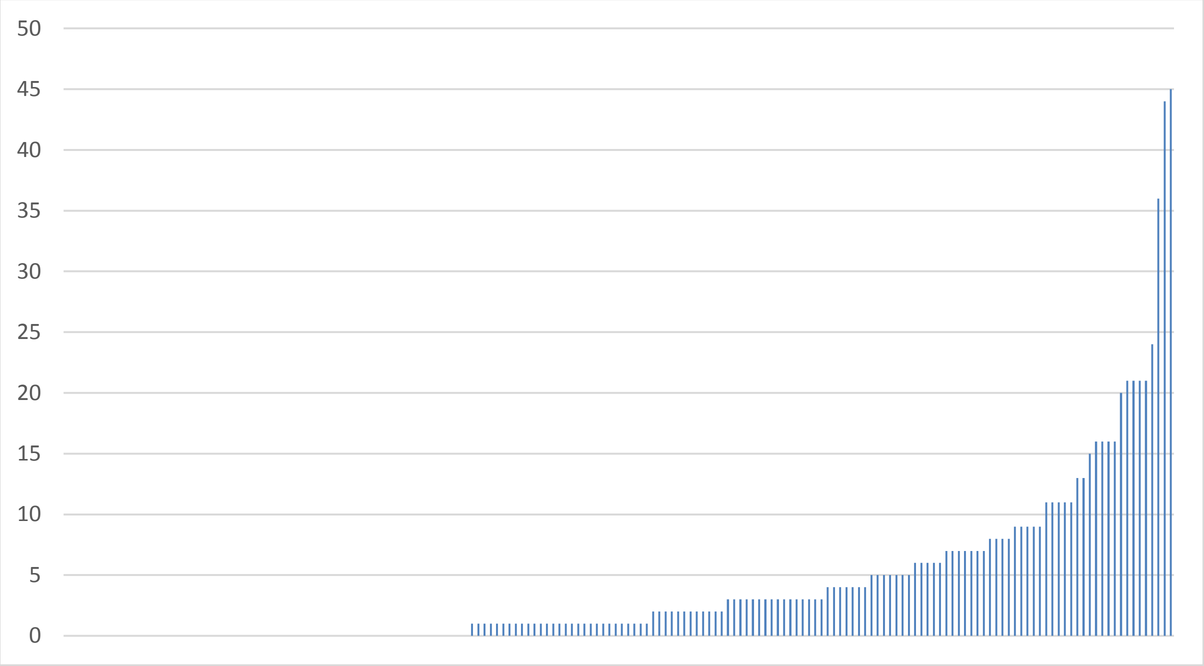
Rank	City	Country	Investments per million capita
33	Bangor	United Kingdom	8.33
34	Brasov	Romania	8.33
35	Budapest	Hungary	8.20
36	Szczecin	Poland	8.00
37	Copenhagen	Denmark	7.88
38	Toulouse	France	7.44
39	Zurich	Switzerland	7.41
40	Lodz	Poland	7.22
41	Newport	United Kingdom	7.14
42	Munich	Germany	6.93

I see that Copenhagen is ranked as the 38th (out of 176 cities) best European city to attract FDI in Design, Development and Testing.

One thing which arises from comparing table 4 with table 2 is that there is a lot larger difference between the cities in table 2 than the ones in table 4. This is an interesting point, which suggests that a few cities receives the majority of investments and the rest are competing for a relatively small part of the cake. Looking at the graph below it is evident that the data does contain large differences between the top and medium attractors.

This suggest that the competition for FDI could resemble a winner takes it all game. Being the best city at attracting FDI means that you are receiving the lion's share of investments.

Figure 16 Absolute number of FDI cases in Design, Development and Testing



Source: fDi Markets

However I do not see any of the cities which Copenhagen Capacity, normally compares Copenhagen with. Those cities are; Malmö, Stockholm, Oslo, Helsinki and Hamburg. If I however chose to compare Copenhagen to those cities I get the following table.

Table 5 Ranking of Copenhagen and its normal competitors

Rank	City	Country	Investments per million capita
31	Stockholm	Sweden	8.33
32	Malmo	Sweden	8.33
37	Copenhagen	Denmark	7.87
52	Oslo	Norway	5.17
59	Helsinki	Finland	4.65
80	Hamburg	Germany	2.14

As seen from the table above, Copenhagen is still performing relatively well compared to its competitors. What I see is that it is only beat by Malmö and Sweden, and therefore performs better than all Non Swedish traditional competing cities. Looking a bit deeper into the data, I get the following rankings for each individual category.

Table 6 Ranking of Copenhagen and its competitors

City	Stockholm	Malmo	Copenhagen	Oslo	Helsinki	Hamburg
Chief Engineer (descending)	144	133	120	152	124	163
No. days to enforce a contract (descending)	5	5	129	4	10	27
Soundness of banks (descending)	170	170	107	175	176	116
Financial cards in circulation per capita	103	103	135	82	107	117
Imports of goods & commercial services (% of GDP)	62	62	29	177	60	46
Science in schools is sufficiently emphasized	60	60	7	71	1	42
Management education meets the needs of the business community	7	7	5	11	12	23
Real corporate taxes do not discourage entrepreneurial activity	11	11	129	116	121	15
The legal and regulatory framework encourages the competitiveness of enterprises	6	6	98	1	10	106
Adaptability of government policy	1	1	25	13	27	30
Bureaucracy does not hinder business activity	1	1	5	23	5	24
Investment incentives are attractive to foreign investors	114	114	136	130	155	99
Ease of doing business	1	1	10	12	18	110
Equal opportunity legislation in your economy encourages economic development	2	2	12	1	6	144
Cyber security is being adequately addressed by corporations	1	1	8	59	10	40

As I can see from the graph Copenhagen has some categories where they seem to be doing better than its competitors, and some where it does not seem to be doing as well.

On the positive side I see that Copenhagen seems to have the cheapest salaries for Chief Engineers. This makes the city a lot more financially attractive to foreign investors, and could possibly sold as an investment incentive.

Copenhagen imports more as a % of GDP and have a lower soundness of banks. Both points to Copenhagen having a more open economy, which is important for global investors who wish to act on a global market.

Furthermore the management education in Copenhagen seems to be better aligned with the needs of business compared to the cities which I compare the city with. This could lead Copenhagen Capacity to argue, that Copenhagen have better business talent in Copenhagen or at least better possibilities to qualify ones workforce.

On the negative side, Copenhagen looks severely regulatory inefficient when it comes to its competitors. It does take significantly longer to enforce a contract in Copenhagen, at the only city which scores worse than Copenhagen on *The legal and regulatory framework encourages the competitiveness of enterprises* parameter is Hamburg.

Another negative for Copenhagen is how its tax rate is perceived by top managers. This could hurt Copenhagen when arguing for Copenhagen being a financially attractive location to place ones Design, Development and Testing facility.

As shown Copenhagen has both advantages and disadvantages to its competitors, and I believe that this information gives me two main conclusions.

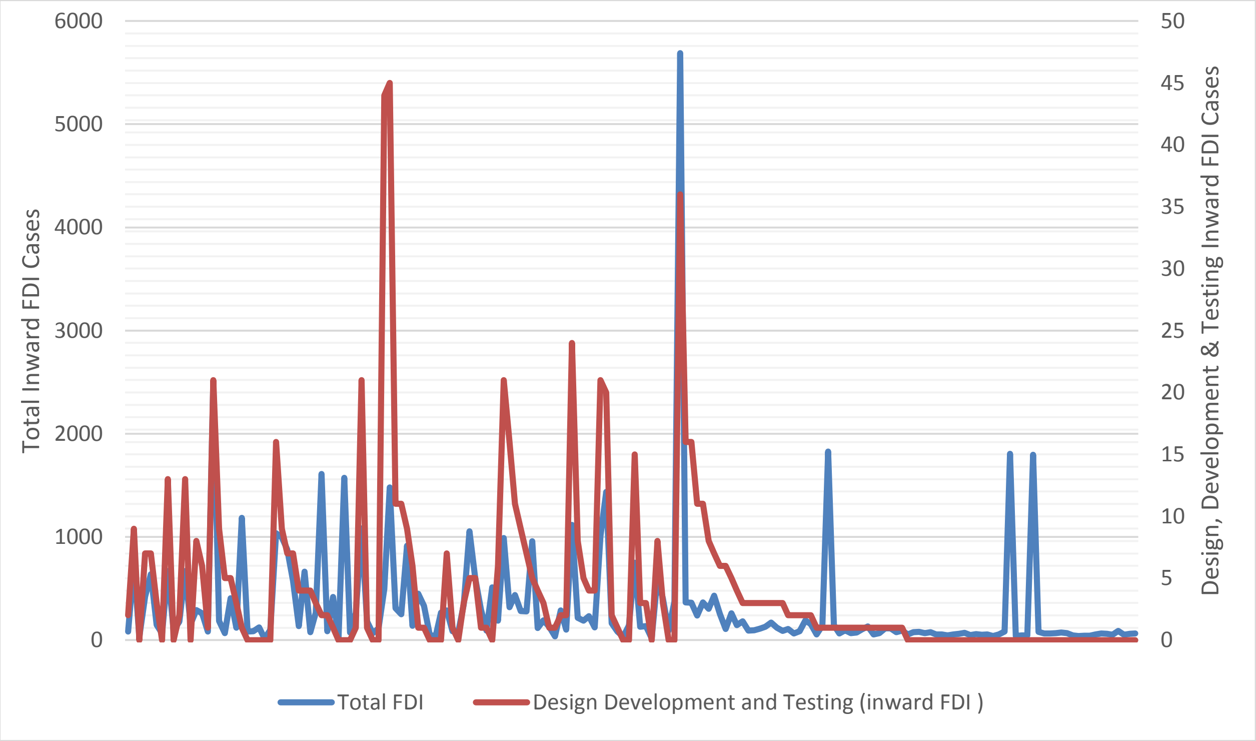
1. Copenhagen is in a position where it can compete with is competitors.
2. Copenhagen Capacity should focus on improving Copenhagen Capacities brand on key parameters which are important to the attraction of FDI in Design, Development and Testing.

The Impact of other types of FDI

One thing which I do not touch upon on in the analysis, is the effect of other types of FDI. From the correlations it is evident that other types of FDI flows are highly correlated with FDI in Design, Development and Testing. That suggests that cities which are generally good at attracting FDI, will increase the likely hood of also receiving FDI in Design, Development and Testing.

If one tries to calculate the relation between the total amount of inward FDI Cases and the inward FDI cases in Design, Development & Testing I found that they have a correlation of 0.4228 for the global sample and 0.5799 for the European sample. I have visualised the relationship for the European sample in the graph below.

Figure 17 Relationship between total FDI Cases and FDI in Design Development & Testing Cases



Source: fDi Markets

The effect of other types of FDI were omitted from the dataset before I initiated the Bayesian Network analysis. That was because an answer which told Copenhagen Capacity, the best approach to attract FDI in Design, Development and Testing is to attract other types of FDI.

However this observation is an interesting one, since it leads me to think about what kind of strategy an Investment Promotion Agency should take. Should it focus on attracting investment as specific as FDI in Design, Development and Testing, or should it rather focus its energy on communicating broader strengths to a broader audience.

On one side one could also argue that it is irrelevant what type of FDI one focuses on, since being strong enough to attract one type, would easily translate into being strong enough to attract other types of FDI.

However the data I have suggests that the attraction is a winner takes it all type of game. If this is the case it would not be beneficial to brand the city as a decent location for most types of investments. It would perhaps be more beneficial to isolate niche industries in which Copenhagen is performing significantly better than other cities. By doing so, it would take the absolute majority of such investments.

In the end the decision should be made with the resources which Copenhagen Capacity have in mind. If the strength of the organisation favours one approach, then that is the approach which should be chosen.

Conclusion

Throughout this analysis, I tried to answer my problem statement, which asked the following questions:

- Is it possible to find any causal relationships with test market investments?
  - If yes, what are these causal relationships?

- How can these results be applied for Copenhagen Capacity to create better results for them as an organisation?

To discover a causal relationship with I decided to you use Bayesian Networks, for a Bayesian Network to show causal relationships it must fulfil two conditions, the first is of Minimality & Stability and the second is the Causal Sufficiency Assumption. It was found that my model fulfilled the Minimality & Stability condition, however I could not confidently state that the Causal Sufficiency assumption held. Therefore it cannot without a doubt be said that causal relationships where found.

However the dataset did reveal interesting observation. I found that the following parameters are conditionally dependent on FDI in Design, Development and Testing.

Parameters
Chief Engineer/Technical Mgr
No. days to enforce a contract
Soundness of banks
Financial cards in circulation per capita
Imports of goods & commercial services (% of GDP)
Science in schools is sufficiently emphasized
Management education meets the needs of the business community
Real corporate taxes do not discourage entrepreneurial activity

The legal and regulatory framework encourages the competitiveness of enterprises
Adaptability of government policy
Bureaucracy does not hinder business activity
Investment incentives are attractive to foreign investors
Ease of doing business
Equal opportunity legislation in your economy encourages economic development
Cyber security is being adequately addressed by corporations
Investment in D, D and T per million person

After having shown the parameters which were conditionally dependent on FDI in Design, Development and Testing, a regression model was built to show the combined explanatory power of the parameters and there individual significance.

It was shown that the model only had an r^2 of 0.2037 and only 3 parameters were significant on the 5% level. This further proves that the causal sufficiency assumption did not hold.

After having shown the empirical evidence for parameters which affects the attraction of FDI in Design Development and Testing. I showed how these findings can be used in practise for Copenhagen Capacity. I found three main conclusions which are relevant for Copenhagen Capacity:

1. Since the IMD World Competitiveness parameters are significantly overrepresented amongst the parameters which are conditionally dependent on the attraction of FDI Design, Development and Testing. It suggest that branding of a city as being competitive in key areas are more efficient than actually improving those key areas.

2. Attraction of FDI in Design, Development and Testing is highly positively correlated with attraction of FDI in general.

3. It was shown that the attraction of FDI resembles a winner takes it all type game and Copenhagen Capacity should therefore focus on competing in niche industries where they are more likely to be ahead of other cities.

Using these conclusions I believe that Copenhagen Capacity can optimise their efforts in attracting FDI.

However in the end the main conclusion for this thesis must be, that the factors which drives Foreign Direct Investments are not easily identified.

Different cities have used different investment attraction strategies which have both been successful and failed. In the end it is important for Copenhagen Capacity that they study the data, and chose to compete on the parameters which fits their core strengths.

Bibliography

Books

Causality: Models, Reasoning, and *Inference* 2nd Edition, Judea Pearl, 2009

Basic Econometrics 5th Edition, Damodar N. Gujarati & Dawn C. Porter, 2008

The Direction of Time, Hans Reichenbach, 1956

Big Data, A Revolution that will transform how we live, work and think, Kenneth Cukier and Viktor Mayer-Schonberger, 2013

Learning Bayesian Networks, Richard Neapolitan, 2003

A Bayesian Approach to Causal Discovery, D. Heckerman C. Meek & G. Cooper, 1999

Probabilistic Networks – An Introduction to Bayesian Networks and Influence Diagrams, Uffe B. Kjærulff & Anders L. Madsen, 2005

Tabu Search – Part 1, Fred Glover, 1989

Articles

The Competitive Advantage of Nations, Michael E. Porter, 1990

The recovery of causal poly-trees from statistical data, G. Rebane and J. Pearl, 1987

A theory of Inferred causation, J. Pearl & T. Verma, 1991

Learning Bayesian Networks is NP-Complete, David Maxwell Chickering, 1994

An Introduction to Causal Inference, Richard Scheines, 1997

Hill-climbing Search, Bart Selman & Carla P Gomes, 2006

How tax policy and incentives affect foreign Direct Investment a review, Jacques P. Morisset & Nede Pirnia, 1999

Learning Causal Bayesian Network Structures from Experimental Data, Byron Ellis & Wing Hung Wong, 2008

Causal Analysis in Economics: Methods and Applications, Pu Chen, Chihying Hsiao, Peter Flaschel & Willi Semmler, 2007

Structure Learning of Bayesian Networks using Constraints, Cassio P. de Campos, Zhi Zeng & Qiang Ji, 2011

The max-min hill-climbing Bayesian network structure learning algorithm, Ioannis Tsarmardinos, Laura E. Brown & Constantin F. Aliferis, 2005

The Direction of Causation: Ramsey's Ultimate Contingency, Huw Price, 1992

Acyclic Orientations of Graphs, Richard P. Stanley, 1972

Using Markov Blankets for Causal Structure Learning, Jean-Philippe Pellet & André Elisseeff, 2008

Package "deal", Susanne Gammelgaard Bottcher & Claus Dethlefsen, 2013

Deal: A Package for Learning Bayesian Networks, Susanne Gammelgaard Bottcher & Claus Dethlefsen, 2003

Package "bnlearn", Marco Scutari, 2013

Learning Bayesian Networks with the bnlearn R Package, Marco Scutari, 2010

Bayesian Methods: General Background, An Introductory Tutorial, E.T. Jaynes, 1996

Websites

http://blogs.law.uiowa.edu/ebook/faqs/what-are-special-economic-zones

Stanford Encyclopaedia of Philosophy, plato.stanford.edu/entries/logical-empiricism

http://www.fdimarkets.com/about/

http://www.fdibenchmark.com/about/