Causal Structure Learning for Famine Prediction

Ernest Mwebaze and **Washington Okori** and **John A. Quinn**
Faculty of Computing & IT
Makerere University
P.O. Box 7062, Kampala, Uganda
{*emwebaze, wokori, jquinn*}*@cit.mak.ac.ug*

Abstract

Food shortages are increasing in many areas of the world. In this paper, we consider the problem of understanding the causal relationships between socioeconomic factors in a developing-world household and their risk of experiencing famine. We analyse the extent to which it is possible to predict famine in a household based on these factors, looking at a data collected from 5404 households in Uganda. To do this we use a set of causal structure learning algorithms, employed as a committee that votes on the causal relationships between the variables. We contrast prediction accuracy of famine based on feature sets suggested by our prior knowledge and by the models we learn.

Introduction

Many inhabitants of developing countries are at risk of famine, and quantifying the risk under different circumstances is an important problem. One clear indicator of impending famine is the early stage of food shortage at the household level, known as 'food insecurity'.

Household food security may be related to several socioeconomic factors such as age, sex, marital status and education level of the household head, location, size of land, size of household, amount of labour available, possession of livestock, distance from the household to the main road, income and presence of agricultural shock (Okori, Obua, and Baryamureeba 2009). In this study we analyze an extensive dataset of such variables collected from households in Uganda, and apply structure learning techniques in order to understand the causes of famine and predict which households are at risk. A similar approach based on Bayesian networks has been employed in the study of the socioeconomic factors affecting technology penetration (Nedevschi et al. 2006).

To understand the causative relationships in the data, we employ a set of causal structure learning methods that are combined in a committee and vote on the causal relationships between the different variables. We present three candidate models for the causal relationships between the variables under study: 1) an *a priori* model, derived using our knowledge of the domain without reference to the data, 2) a model derived using statistical inference only, and 3) a model obtained by using both prior domain knowledge and statistical inference. We note that causal structure learning on natural datasets often makes most sense as an iterative process, combined with a structure prior from domain knowledge. Given the three models, we compare prediction accuracy of famine risk using corresponding feature sets.

Our work has implications for policy and intervention planning in food security, indicates how predictable food shortages are at the household level, and provides evidence that a combination of recently developed structure learning methods (which have mainly been analysed with synthetic data) are capable of producing intuitively plausible results from natural datasets.

We now review first the structure learning algorithms and then the dataset used in this analysis. In subsequent sections we present our inferred structures and findings in predicting famine at the household level.

Causal Structure Discovery

In recent years causal structure discovery has emerged as a branch of machine learning. The goal is to learn causal relationships from purely observational data, without being able to perform manipulations.

The first benefit of such techniques is the ability they provide to understand causal mechanisms in domains such as food security, epidemiology or genetics where it is either expensive, unethical or impossible to carry out certain experiments. Other advantages of causal analysis include prediction under interventions, manipulation and counterfactual.

Causal analysis is of an entirely different nature to correlation-based analysis (density modelling) of a dataset, as covered in (Pearl 2000). A causal relationship implies a change in one variable will have a corresponding change in any variable that is causally dependent on it, and this must hold even when the variables are subjected to external interventions. This is important when predictions are to be made on data is subject to external interventions.

A probabilistic (Bayesian) network is a graph in which nodes represent random variables and directed edges between the nodes encode dependence and independence information between the different nodes. Causal networks – which we are interested in learning here – are a subset of directed Bayesian networks, where we interpret the arcs as

direct causal relationships. In a directed Bayesian network, an arc from A to B determines the conditional independence class and does not necessarily mean that A is a cause of B. There can be many different causal networks which have the same conditional independence properties and are therefore indistinguishable as Bayesian networks.

Most classification and regression algorithms presume that test data has the same distribution as training data. When the distribution of test data may be altered due to interventions, only methods which incorporate an accurate causal model of the data that are able to make robust inferences.

Two broad categories of methods have been employed to date in structure discovery: search-and-score based methods which use probabilistic methods to score a causal directed acyclic graph (DAG) based on the data, and constraint-based methods that use the conditional independence properties between the variables in the data to ascertain the causal relationships. There are several variations on these themes, for example methods that focus on discovering latent variables (Spirtes, Glymour, and Scheines 2000; Elidan et al. 2000) and on learning linear non-Gaussian models (Hoyer, Shimizu, and Kerminen 2006b). These methods have been tried extensively on synthetic datasets, but less is known about their performance on real-world data.

Learning causal relationships in a network of variables is not trivial: a naïve search for all possible DAGs in a dataset with n random variables results in a hyperexponential number of possible graphs $r(n)$ where for example $r(2) = 3, r(3) = 25, r(5) = 29281, r(10) \approx 42 \times 10^{18}$ (Robinson 1977). In the case of constraint-based testing for conditional independence relationships, the problem is that to establish an independence relationship between any two variables we would in principle need to test this for all possible conditioning sets.

In this paper we apply a committee of causal structure learning methods. We simplify the task of searching through the large number of possible models by first finding an undirected graph between the variables (the skeleton), and then orientating the edges using the structure learning committee.

The Causal Committee

The committee method used (Mwebaze and Quinn 2009) attempts to harness the combined power of differently motivated structure discovery algorithms. In order to optimize the economics of execution time and accuracy, the committee method is initialized by skeleton discovery with feature reduction. We do this by first discovering the sets of parents and children for each variable in order to break the problem up into sets of tractable local neighbourhoods (skeleton discovery). We then apply a structure learning committee for orientating edges between the variables.

Skeleton Discovery

We obtain the local neighborhoods of the variables by use of a pseudo-committee of extensions of HITON-PC(max-k = 3, alpha = 0.01); (i) HITON applied to the full dataset, (ii) HITON initialized by sets of relevant features for the different variables.

Algorithm 1 Localized causal discovery committee.

1: **input**: $\mathbf{c}_1 \ldots \mathbf{c}_N$, data vectors for variables $C_1, \ldots, C_N$
   $PC(i)$, set of parents and children for each variable C_i.
2: **for each** variable C_i, $i = 1 \ldots N$ **do**
3: **for each** algorithm $Algo_j$ **do**
4: $causes(C_i, j), effects(C_i, j) \leftarrow Algo_j(C_i, PC(i))$
5: $\mathbf{C_i} \leftarrow$ majorityVote($causes(C_i, :)$)
6: $\mathbf{E_i} \leftarrow$ majorityVote($effects(C_i, :)$)
7: **return**: {**C,E**}, causes and effects of $C_1, \ldots, C_N$.

Relevant features for the variables are obtained using Relevance Learning Vector Quantization (RLVQ) (Bojer et al. 2001), an extension of Kohonens LVQ. LVQ and all methods based on it are essentially prototype-based classification methods applied in supervised machine learning. They employ a distance measure (typically Manhattan distance or quadratic Euclidean distance) that quantifies the similarity of a given feature vector with a prototype (representative) of any particular class. RLVQ employs adaptive scaling factors that scale the features based on their relevance for classification.

HITON is a standard algorithm for feature selection. It has been shown to have two main advantages over other feature selection algorithms. Firstly, it reduces the number of variables in the prediction models roughly by three orders of magnitude relative to the original variable set while improving or maintaining accuracy. Secondly, it outperforms the baseline algorithms by selecting smaller variable sets than the baselines (Aliferis, Tsamardinos, and Statnikov 2003).

For each variable the union of the results of each of the members of this pseudo-committee forms its skeleton of directly connected variables. Individual feature/variable skeletons combined form the skeleton graph of our famine network.

Committee Members

Once the skeleton is obtained, we take each variable C_i individually with its direct neighbours (the set $PC(i)$ of parents and children) and try to orientate the edges to find which are the parents and which are the children. Rather than use a single method to orientate the edges, we use a committee of causal discovery algorithms, some standard and some relatively novel. Each method is applied to the skeleton to orient the edges and then voting amongst the individual methods about the causal relationships between the variables is done. This is advantageous in two ways (i) because some of these algorithms have mainly been analysed on synthetic data in the literature, it is difficult to predict how they will perform on real data, hence averaging the performance over several of these methods results in better accuracy, and (ii) we use methods that have varied strengths and different assumptions based on the data they are applied to, hence application of a committee of these methods on a real dataset and voting is likely to give a more robust result than using any one of these methods individually. The committee algorithm is described in Algorithm 1.

Standard algorithms were used to form the structure learn-

ing committee. They were selected to include methods based on different principles. We use PC (Spirtes, Glymour, and Scheines 1993), MWST (Chow and Liu 1968), GES (Munteanu and Bendou 2002), K2 (Cooper and Herskovits 1992), LiNGAM (Shimizu et al. 2006) and EPC (Mwebaze and Quinn 2009).

Famine Data

The dataset used in this paper comprised of information collected by the Uganda Bureau of Statistics from 5404 households in four regions of Uganda (Central, Eastern, Northern and Western), spanning 57 districts of the approximately 80 districts of Uganda. The data collected was aimed at determining the degree to which households are susceptible to famine or food insecurity. The original dataset has 24 features. For this study several of the features were removed including household size, region, district, and income. Income was removed because it had several missing values. In all for this study a total of 13 features were used, listed in Table 1. Preprocessing reduced the data to 3094 households after deleting all the households with missing values in any of the 13 features. The preprocessed data was used in its entirety to discover the underlying causal structure around the target variable and later split into training and test data to test the prediction accuracy based on the whole dataset, and contrasted with prediction accuracy based on the derivative datasets from the derived causal graphs.

For individuals who are faced with food shortages, the caloric intake in their diet is reduced and this measure can be used as a proxy for food insecurity (Salih 1994). The net caloric intake is calculated from the difference between the energy content (calories) of each agricultural food crop that is produced by a household and that utilized for different purposes like animal feeds and wastage within an agricultural year. For each household this net value is divided by the number of people in a household and number of days in a year to derive the dietary energy intake per person per day. Those households that fall short of a value of 1800 kilocalories per person per day in this study are considered as being food insecure.

Sex of the household head	male/female
Age of the household head	years
Marital status of the household head	married/ divorced/ single/ widowed
Size of household	number of people
Size of land available to the household for farming	acres
Amount of labour available for cultivation per year	person-years
Distance from household residence to the nearest main road	km
Distance from household residence to farm land	km
Total annual production of crops available for consumption by the household (excluding crops which are sold)	kg
Agricultural shock (e.g. presence of flooding, drought, market fluctuation)	true/false
Crops attacked by pests	true/false
Ownership of livestock	true/false
Household famine status (whether daily calorie intake per person in the household is above 1800 kCal)	famine/ not famine

Table 1: Variables in the famine dataset describing each household surveyed.

Experiments and Results

Experiments were carried out on three different configurations of the famine network.

A Priori Graph

A graph was first constructed based on the authors' knowledge of the domain, without reference to the data. The structure is shown in Figure 1.

The intuition behind the causal relationships in Figure 1 is as follows:

TP → Fa : The more food that a family produces, the more is available for direct consumption.

Li → Fa : Owning livestock is a direct mitigation of famine risk, e.g. consumption of milk and eggs increases calorie intake.

HS → Fa : The greater the size of the household, the smaller the share of consumable produce per person.

La → TP : The greater the size of the household, the more manpower is available for raising crops.

LS → TP : The more land available to a family for farming, the more potential they have for growing crops.

DR → TP : Households closer to the road are more likely to sell produce rather than keep it for consumption.

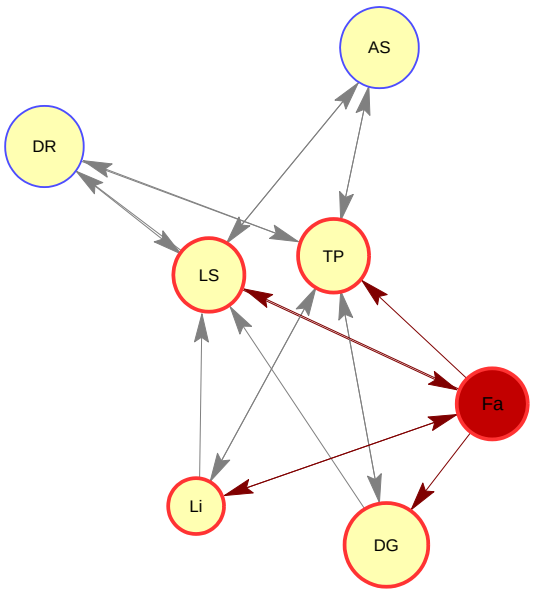
DR → LS : Land closer to the roads is more expensive, making typical plots of land smaller.

DG → TP : More time required to travel and transporting goods to and from the farm means that less time is available for work on cultivation.

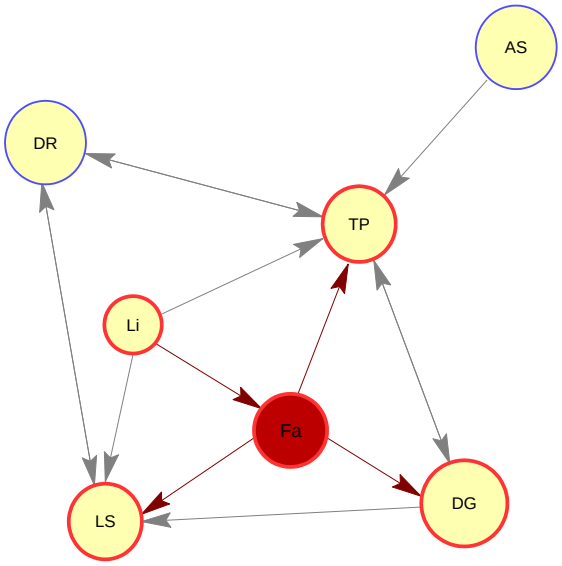
AS → TP : Agricultural shock (e.g. flooding, drought, market fluctuation) directly affects production.

PA → AS : Pest attack is part of agricultural shock by definition.

HS → La : The more people in the household, the more manpower is available for food cultivation.



(a) Graph learnt with uninformative prior



(b) *A posteriori* graph

Figure 2: Inferred and *a posteriori* causal graphs depicting causal relationships in the famine data set. The highlighted nodes signify nodes directly connected to the target variable *famine/food insecurity*. Features are represented as **HS**-Household Size, **LS**-Land Size, **La**-Labour, **DR**-Distance to main Road, **DG**-Distance to Garden, **TP**-Total Production, **AS**-Agricultural Shock, **PA**-Pest Attack, **Li**-Livestock, and **Fa**-Famine.

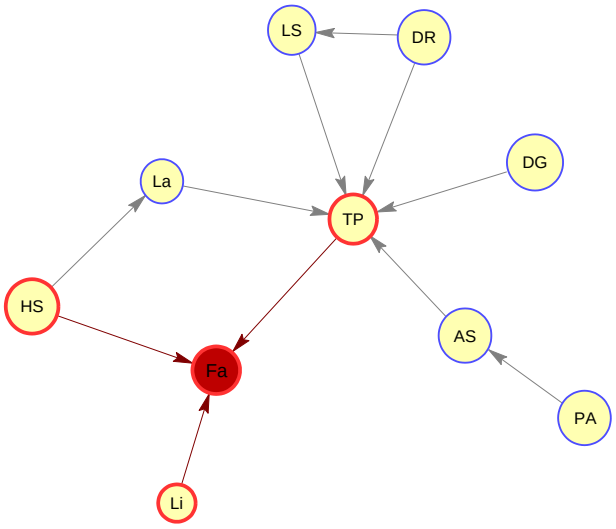


Figure 1: Intuitive *a priori* graph of causal relationships in the famine dataset. Features are represented as **HS**-Household Size, **LS**-Land Size, **La**-Labour, **DR**-Distance to main Road, **DG**-Distance to Garden, **TP**-Total Production, **AS**-Agricultural Shock, **PA**-Pest Attack, **Li**-Livestock, and **Fa**-Famine. The rest were assumed to be independent.

Graph Learnt with Uninformative Prior

The discovered graph was learned from the data using the committee method of causal discovery, without incorporating any prior domain knowledge. It is shown in Figure 2(a). The nodes $\{LS, DG, TP, Li\}$ represent the direct causes and effects of our target variable *famine*$\{Fa\}$. It is interesting to note that the discovered graph has as its direct causes *landsize* and *livestock* as the major determinants of whether a household is likely to face famine in a given year or not, which are quite intuitively plausible causes.

Edge	PC	EPC	MWST	GES	LiNG	K2
LS→Fa	1	0	0	1	0	0
Fa→LS	0	1	0	1	0	1
Fa→DG	0	1	1	0	0	1
Fa→TP	0	1	0	1	0	1
Li→Fa	1	0	1	0	0	0
Fa→Li	0	1	0	0	0	1

Table 2: Table showing relative voting strengths of the different committee algorithms for causative edges related to famine. The numbers represent whether a particular algorithm voted for or against.

Table 2 shows how each algorithm voted. It represents how confident the committee was about each of the edges related to famine/food security. It is interesting to note that two algorithms EPC and K2 agree exactly on the same set of causes while LiNGAM that assumes non-Gaussian distributions does not find any causes. A voting threshold of 2 was used to obtain the uninformed graph depicted in Figure 2 (a).

A Posteriori Graph

The *a posteriori* graph was obtained by including the *a priori* graph in to the committee and doing the voting again. The effect of this inclusion is to strengthen the vote on the prior causative links in the graph. A voting threshold of 3 was used for the *a posteriori* graph, and we double the weight of votes from the prior knowledge model (the weight of votes determines our confidence in the structure learning as opposed to our assessment of causes from prior knowledge). The graph is depicted in Figure 2 (b). Intuitive explanations for some of these causal relationships is as follows:

Li → LS : The more livestock a household has the more likely they are to look for a larger piece of land, or conversely, the more land a household has, the more likely they are to have livestock.

Fa → TP : This is somewhat counterintuitive, as we expect the reverse that low total production of crops leads to a state of famine. However this causal relationship does have a plausible interpretation. Total production is the amount of crops produced excluding those which are sold. During a shortage, families often cut down to one meal a day in order to conserve their resources. If they produce perishable crops then they are likely to sell the rest in order to build up a financial buffer.

Fa → LS,DG : In these cases we would also expect the most probable relationships would be the reverse. It is plausible however that some members of the committee may lay greater emphasis on the effects of famine than the causes for instance K2 is a causal structure algorithm that theoretically will orient edges differently based on the ordering of the features. It is hence likely that if the target $\{Fa\}$ is analysed first, the algorithm will give a stronger weight to the effectual relationship of consecutive features with the target.

Another interesting result is that we obtain several intuitive causes of variable $\{TP\}$, the total production. The fact that the two derived graphs depict some intuitive causative characteristics is interesting because it provides some support that the methods used are able to derive plausible causes from observational data. The bidirectional nature of some relationships arises because for a natural dataset like the one we used, the features intrinsically and intuitively tend to have bidirectional causal influence depending on which one manifests first. A bidirectional link can also indicate the presence of a hidden cause, which influences both of the observed variables.

Note that it is difficult to find relationships such as *PA* → *AS* statistically. As there is only one cause it is difficult to test with conditional independence.

The methods we used other than EPC were based on implementations from four packages ; BNT-SLT[1], BNT[2], LiNGAM[3] and Causal Explorer[4]. All these experiments were implemented using Matlab on an Intel Core 2 Duo CPU 1.79 GHz laptop.

Classification of famine risk

We further tested the accuracy of our true graph and derived graphs by splitting the data into training and testing data, training several models on a training set and measuring their prediction accuracy on the test set. Four datasets were used; the full dataset and three datasets derived from the direct causes and effects of the target based on the *a priori* graph, the inferred graph and the *a posteriori* graph.

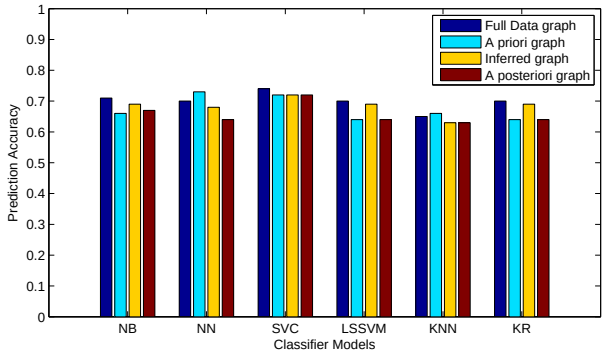


Figure 3: AUC of various standard classifiers with; a complete featureset (Full Data graph) and reduced sets of causally related features with the target (true graph, discovered graph and blended graph). Algorithms used include NB:Naive Bayes, NN:Neural Network, SVC:Support Vector Classifier with a linear kernel, LSSVM:Least Squares Support Vector Classifier, KNN:K-Nearest Neighbour and KR:Kernel Ridge regression classifier

Results from the AUC curve show that a reduced set of parents and causes provides comparable accuracy on prediction. Not only is this advantageous in the predictive sense, but it also has advantages in reducing amount of data collected due to a reduced feature set. A reduced set of features, which we believe are direct causes and effects of the target variable, also make a more robust basis for prediction when there may be interventions on some of the variables (e.g due to humanitarian relief or implementation of new agricultural policies).

Conclusion

This paper presents preliminary results in the causal analysis of famine in Uganda. The causal models we derive can be used to inform policy making with regards to household food security in a country like Uganda where food security or famine is a threat to many households, and indicate the

[1] *http://bnt.insa-rouen.fr/*

[2] *http://www.ai.mit.edu/~murphyk/Software/BNT/bnt.html*

[3] *http://www.cs.helsinki.fi/group/neuroinf/lingam/*

[4] *http://discover.mc.vanderbilt.edu/discover/public/causal_explorer/*

extent to which housefold food shortages can be predicted based on socioeconomic factors.

Discovering causal relationships from natural datasets is quite a challenging task with current techniques. We use a committee of differently motivated structure learning algorithms, in order that spurious results from an individual algorithm may be balanced by rival algorithms. However, there are a limited number of underlying concepts (e.g. measures of conditional independence and model fitness). As is often the case in similar causal studies with observational data, the results are only strong enough to form a basis for further investigation. That is, our results are not enough to make strong conclusions about causal relationships in this domain but do provide confirmation or disconfirmation for particular cause/effect hypotheses.

We have also shown that prediction accuracy is maintained using the variables that we find to be direct causes or effects. This has implications for the number of variables which may need to be collected in future studies. Where we find bidirectional causes we postulate that there may be confounding variables and more data needs to be collected.

Concerning prediction performance we reason that inferences based on variables which have a direct causal link with the target variables are more robust under manipulations of the variables. A density modelling approach to prediction using the entire dataset may not be reliable if the data distribution is shifted due to some external intervention, for example due to a relief effort. Predictions based on true direct causes of a target variable can be expected to have the same reliability whether or not those causes have been manipulated.

Future work will include comparing local causal graphs from the various regions and also between famine prone districts and relatively non-famine prone districts. This paper presents a global view of 57 districts, but some districts may have peculiar behavior given other uncollected data, teasing out these districts and their peculiarities will be the focus of this extension.

It would also be useful to formulate likelihood terms for the members in the causal committee, and produce a more sophisticated structure prior so that we can use Bayesian methods to infer the posterior model.

Acknowledgments

We thank the anonymous reviewers, whose comments helped to improve this paper. This work was funded in part by the Dutch NUFFIC NPT project. We would also like to acknowledge the Uganda Bureau of Statistics (UBOS) for providing us with the famine data.

References

Aliferis, C. F.; Tsamardinos, I.; and Statnikov, A. 2003. HITON, A Novel Markov Blanket Algorithm for Optimal Variable Selection. In *Proc. of the 2003 American Medical Informatics Association (AMIA) Annual Symposium*, 21–25.

Bojer, T.; Hammer, B.; Schunk, D.; and von Toschanowitz, T. 2001. Relevance determination in learning vector quantization. In M, V., ed., *European Symposium on Artificial Neural Networks*. d-facto publications. 271–276.

Chow, C. K., and Liu, C. N. 1968. Approximating discrete probability distribution with dependence trees. *IEEE Transactions on Information Theory* 14(3):462–467.

Cooper, G. F., and Herskovits, E. H. 1992. The induction of probabilistic networks from data. *Machine Learning* 9(4):309–347.

Elidan, G.; Lotner, N.; Friedman, N.; and Koller, D. 2000. Discovering hidden variables: A structure-based Approach. *Neural Information Processing Systems* 13:479–485.

Hoyer, P.; Shimizu, S.; and Kerminen, A. 2006b. Estimation of linear, non-gaussian causal models in presence of confounding latent variables. In *Proc. of the third European Workshop on Probabilistic Graphical Models (PGM2006)*.

Munteanu, P., and Bendou, M. 2002. The EQ framework for learning equivalence classes of Bayesian networks. In *First IEEE International Conference on Data Mining (IEEE ICDM)*.

Mwebaze, E., and Quinn, J. 2009. Fast Commitee-Based Structure Learning. In *NIPS 2008 Workshop on Causality. Forthcoming*.

Nedevschi, S.; Sandhu, J. S.; Pal, J.; Fonseca, R.; and Toyama, K. 2006. Bayesian networks: an exploratory tool for understanding ict adoption. In *Information and Communication Technologies and Development, 2006. ICTD '06. International Conference on*, 277–284.

Okori, W.; Obua, J.; and Baryamureeba, V. 2009. Famine Disaster Causes and Management Based on Local Community's perception in Northern Uganda. *Research Journal of Social Sciences* 4:21–32.

Pearl, J. 2000. *Causality : Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Robinson, R. W. 1977. Counting unlabeled acyclic digraphs. In Little, C., ed., *Combinatorial Mathematics V*, volume 622 of *Lecture Notes in Mathematics*. Berlin: Springer.

Salih, S. 1994. Food Security in East and Southern Africa. *Nordic Journal of African Studies* 13(1):3–27.

Shimizu, S.; Hoyer, P. O.; Hyvarinen, A.; and Kerminen, A. 2006. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Machine Learning Research* 7:2003–2030.

Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction and Search*, volume 81. Berlin: Springer Verlag.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction and Search*. Cambridge: Cambridge University Press.