Europ. J. Agronomy 75 (2016) 80–88

Contents lists available at ScienceDirect

European Journal of Agronomy

journal homepage: www.elsevier.com/locate/eja

Bayesian network modeling of early growth stages explains yam interplant yield variability and allows for agronomic improvements in West Africa☆

Denis Cornet[a,b,*], Jorge Sierra[c], Régis Tournebize[c], Benoît Gabrielle[d], Fraser I. Lewis[e]

[a] CIRAD, UMR AGAP, F-34398 Montpellier, France
[b] IITA-Benin, BP 08-0932 Cotonou, Benin
[c] INRA, UR1321 ASTRO, Agrosystèmes Tropicaux, Petit-Bourg F-97170, Guadeloupe, France
[d] UMR ECOSYS, INRA, AgroParisTech, Université Paris-Saclay, 78850 Thiverval-Grignon, France
[e] Section of Epidemiology, VetSuisse Faculty, University of Zürich, Winterhurerstrasse 270, Zürich CH8057, Switzerland

ARTICLE INFO

Article history:
Received 28 October 2015
Received in revised form 20 January 2016
Accepted 21 January 2016
Available online 1 February 2016

Keywords:
Additive Bayesian network modeling
Cataphyll
Early growth
Vegetatively propagated crops
Yam (*Dioscorea* spp.)
Yield variability

ABSTRACT

Yams (*Dioscorea* spp.) are important species, especially for resource-poor farmers of West Africa, where crop yields are affected by early plant size hierarchy linked with uneven emergence. Although the causes of this phenomenon are not fully known, yams, like other vegetatively propagated crops, have heavy planting material that is liable to induce such interplant variability. In addition, planting practices may mitigate this phenomenon via the selection of the seed-tuber size or state. To gain further insight into yam interplant variability, this study identified and quantified, for the first time, the direct and indirect dependency between planting practices, early growth variables and yield components of *Dioscorea rotundata* and *Dioscorea alata*, the two main food yam species. The experimental dataset came from six field trials carried out in Benin at two locations between 2007 and 2009. Additive Bayesian network modeling was used for structure discovery—its directed acyclic graph offers an ideal background for discussing complex systems when theoretical knowledge is lacking, e.g., for yams. Here we found that the emergence date was the only direct cause of plant yield variability common to both species. For *D. rotundata*, we observed a direct contribution of the cataphyll number to the plant tuber weight. These combined results suggest the existence of some uncontrolled latent variables (i.e., seed-tuber physiological age and reserves). For *D. alata*, the model did not reveal any effect of seed-tuber size, despite a strong effect noted for *D. rotundata*. We suggest that the transposition of traditional native *D. rotundata* planting practices may have led to oversized *D. alata* seed-tubers, resulting in wastage of planting material. This study demonstrated that traditional West African cropping systems have a serious drawback concerning the uncontrolled wide range of physiological ages and reserves in seed-tuber lots, which affect the plant size hierarchy and ultimately the marketable yield.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Yams (*Dioscorea* spp.) belong to a C3 monocotyledonous genus grown for food, pharmaceutical products and ornamental purposes (Ayensu, 1972; Cornet et al., 2007). Food yams are cropped for their underground tubers, which represent a key source of carbohydrates in many regions worldwide (Kennedy, 2003; Asiedu and Sartie, 2010). Projections have shown that yam consumption will increase rapidly in West Africa, leading to a higher production rate for yam compared to cassava (Scott et al., 2000). *Dioscorea rotundata* and *Dioscorea alata* are the two top ranked species in terms of economic importance. Although *D. rotundata* is the most cultivated species in West Africa, where 90% of the world production is located, *D. alata* is the most ubiquitous yam species and is grown from Japan to West Africa, and throughout Central America (Orkwor et al., 1998). Despite its importance regarding food security and household income, particularly for resource-poor farmers throughout the world, few studies have focused on yam physiology and cropping systems (Marcos et al., 2011; Cornet et al., 2014). For example, over the past 45 years the Web of Science® has referenced

☆ This paper is in memoriam of Bertrand Ney (1956–2013).

* Corresponding author at: CIRAD, UMR AGAP, INRA, UR1321 ASTRO, Petit-Bourg F-97170, Guadeloupe, France. Fax: +33 590 841663.
E-mail addresses: denis.cornet@cirad.fr, denis.cornet@antilles.inra.fr (D. Cornet).

http://dx.doi.org/10.1016/j.eja.2016.01.009
1161-0301/© 2016 Elsevier B.V. All rights reserved.

nearly 500,000 publications for the single maize species (*Zea mays*), but there have been 35 times fewer publications for the whole yam *Dioscorea* genus.

In a recent study, Cornet et al. (2014) highlighted the strong plant size hierarchy in West African yam fields and the resulting potential adverse effects on total and marketable yield. Competition between neighboring plants is negligible in West African yam cropping systems. This absence of competition allowed the authors to study the crop yield variability at the individual plant scale. Cornet et al. (2014) pointed out the role of individual emergence date to explain part of this interplant variability and concluded that other complex interacting processes involved in early growth stages might play an important role. Unlike potatoes (*Solanum tuberosum*), yams do not benefit from a certified tuber seed production system to enable fast uniform sprouting. Many factors may affect the timing and vigor of yam emergence. In the absence of quality seed tuber production, some of these factors are stochastic (i.e., physiological age and nutrient content of seed-tubers) while others can be managed or controlled: seed-tuber size, seed-tuber state (presprouted or not) and planting date (Ferguson, 1973; Orkwor et al., 1998).

Because of the lack of expert knowledge, the number of variables involved, and the complexity of the interactions, Bayesian network modeling has been used for structure discovery and parameter learning (Heckerman et al., 1995; Korb and Nicholson, 2004). Bayesian network analysis is a form of graphical modeling focused on structure discovery: determining an optimal statistical model, i.e., graphical structure, directly from observed data. Whilst relatively uncommon in plant development studies, Bayesian network analysis is now being applied to an increasing extent in areas of biology, medicine, ecology or epidemiology (Porth et al., 2013; Ward, 2013). In recent years, Bayesian network modeling has been successfully applied in plant disease epidemiology studies (Kim et al., 2014; Zhu et al., 2013) or to estimate agriculture's environmental risks (Nash et al., 2013). Lewis and McCormick (2012) showed that while multivariable regression seeks to identify covariates associated with some output variables (e.g., plant yield), Bayesian network analysis goes much further in also empirically separating these into those directly and indirectly dependent upon the output variable. Bayesian network modeling has the potential to reveal far more about key features of biological complex systems than existing commonly used approaches (Lewis and McCormick, 2012). Moreover, in Bayesian network modeling, no attempt is made to reduce the dimensionality (e.g., in exploratory principal components analysis) which makes biological interpretation of results easier. Its probabilistic formalism provides a natural means to deal with the stochastic nature of biological systems and measurements (Needham et al., 2007).

The objective of this study was to assess how planting practices and early growth variables affect plant yield formation in the two major yam species. More specifically, it aimed at: (i) discovering and quantifying the dependency structure among practices at planting (i.e., planting date, seed-tuber state and weight), early growth variables (i.e., emergence date, stem and cataphyll number) and plant yield components (i.e., tuber number and weight), (ii) comparing these dependency structures for the two major food yam species, and (iii) discussing the implications of these findings both for farmers and scientists.

2. Materials and methods

2.1. Experimental sites

The dataset used in this study came from six field trials carried out between 2007 and 2009 at two locations: AfricaRice—Cotonou Station (Benin, 6°25N, 2°19E, 23 m asl) and Glazoue (Benin, 7°56N, 2°15E, 200 m asl). We used the two cultivars belonging to the main yam species, i.e., *D. alata* 'Florido' and *D. rotundata* 'Morokorou'. Morokorou is a traditional early-maturing variety originating from north Benin, which produces 1–3 cylindrical tubers. Florido was introduced into West Africa from Puerto Rico in the early 1970s and produces two to five round tubers (Doumbia et al., 2004). The field experiments were both located in a forest-savannah transition zone. The climate is sub-equatorial with a bimodal rainfall pattern, with rain falling mainly from March to July and September to October. Administrative maps of Benin with the two locations, their soil characteristics and weather data are available in the Supplementary material, Figs. A1, A2 and Table A1. The pedo-climatic conditions of the experimental sites are mainly representative of yam growing area of Western Africa (Orkwor et al., 1998).

All trials followed traditional planting systems used in West Africa: entire seed tubers were planted in mounds, without staking, at a density of 0.7 plants m^{-2}. The planting dates were February 20th, 2007, April 25th, 2008, and March 25th, 2009. Final harvests were conducted at full crop senescence, between December and February. Each year seed-tubers were bought from one farmer, and were from the same field under the same climatic and cropping conditions. Fertilizer was applied at a rate of 60 kg N ha^{-1}, 30 kg P ha^{-1} and 140 kg K ha^{-1} one month after emergence. Additional N was applied 2 months after emergence as urea at a rate of 60 kg N ha^{-1}. Although the experiments in Glazoue were not irrigated, in Cotonou the crop was irrigated to field capacity at planting. Afterwards it was irrigated according to a water balance, and further supplementary irrigation (totaling between 80 to 110 mm depending on the cropping season) was applied to replace estimated evapotranspiration using overhead sprinklers (Marcos et al., 2009). The plants did not show any visual sign of water or nutritional stress. Weed control was done by hand roughly on a 2-week basis. Experiments used in this study are representative and cover a wide range of traditional practices for yam in West Africa (i.e., planting dates, planting material, soil preparation...). All experiments had a completely randomized design with yam species being the only treatment, with two levels (*D. rotundata* and *D. alata*), and using four replications of 25 plants per treatment.

2.2. Explanatory variables

Three categories of explanatory variables were defined: variables reflecting seed-tuber management practices at planting, variables describing the plant early growth stage, and yield components (Table 1). All variables except the seed-tuber state are continuous variables and a graphical presentation of their dispersion is available in Supplementary material (Fig. A3). The seed-tuber state was added because seed-tuber germination and stem elongation occur during storage at dormancy break (Orkwor et al., 1998). Since the emergence date is independent of the soil moisture (Onwueme, 1976), rainfall was not included in the model.

In this study, the early growth stage corresponds to the period between planting and the appearance of the first true leaf. Indeed in some yam species leaves at the first few nodes may be reduced to form modified shield-shaped cataphylls (Fig. 1). These cataphylls are thick, lack a distinct leaf lamina and are limited in aerial spread (Onwueme, 1978). As in true seedlings yam vines develop without cataphylls (Okezie et al., 1986), we postulated that the number of nodes carrying cataphylls could be related to reserve availability in seed-tubers, and might therefore indicate the early nutritional status of yam plantlets. This variable is not available because it is not possible to analyze a particular seed-tuber and then plant it. Our specific objective in the current analysis was to investigate whether the number of nodes carrying cataphylls could provide further information to explain the yield components (i.e., not directly

Table 1
Explanatory variables used to build the Bayesian model explaining plant yield.

Category	Variable	Description
Practices	Planting date	Planting date (Julian day)
	Seed-tuber state	State of the seed-tuber prior to planting (binary variable): presprouted (1) or not (0).
	Seed-tuber weight	Weight of the planted seed-tuber (g).
Early growth	Latency	Period between planting date and emergence date (number of days).
	Emergence date	Date of emergence (Julian day). Individual emergence was recorded when a tuber sprout emerged from the soil.
	Stem number	Number of main stem per mound.
	Cataphyll number	Number of nodes carrying cataphylls per mound.
	Cataphyll number per stem	Number of nodes carrying cataphylls divided by the number of main stems per mound.
Yield components	Tuber number	Tuber number per mound
	Tuber weight	Mean tuber fresh weight per mound (g plant^{-1}).

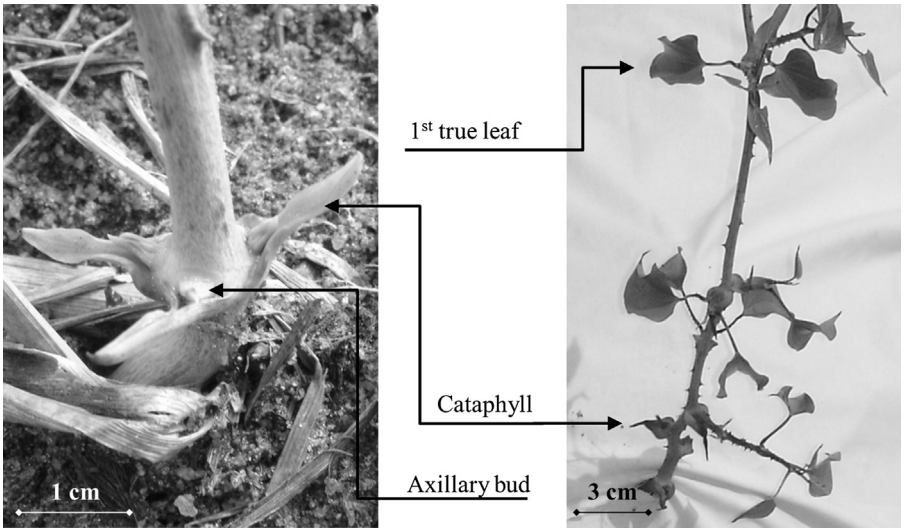

Fig. 1. Morphology of yam (D. rotundata) at the early growth stage (from the planting date until appearance of the first true leaf).

explained by planting date or seed-tuber weight). On the other hand, we assessed the variability of reserves in seed-tuber lots with a subsample of 25 seed-tubers of each species, randomly selected from the planting material of 2009 to measure dry matter, N, P and K content.

Like any model, statistical crop models are a simplified representation of reality and contain inevitable errors among which measurement errors may leads to biased models (Lobell, 2013). In this study the emergence date variable was most prone to measurement error because its value was aggregate every three days. In order to roughly estimate the measurement error linked to emergence date aggregation we utilize the simulation extrapolation approach recommended by Lobell (2013), using the SIMEX library (Cook and Stefanski, 1994; He et al., 2012). Results showed that the corrections of the effect estimates from a naïve generalized linear model were rather small in absolute value for both species (Supplementary material, Figs. A4 and A5). Moreover measurement error issues are minimized by the use of a large panel datasets (i.e., two locations and three growing seasons; Lobell and Ortiz-Monasterio, 2007). However, it is recommended not using this model outside the range of experimental variation (Sheehy et al., 2006).

2.3. Additive Bayesian network

The influence of emergence characteristics over plant yield is typical in multivariable regression modeling (Klemke and Moll, 1990; Rebetzke et al., 2007; Fayaud et al., 2014). By extending this approach to an analogous multivariate regression model in which all variables are simultaneously considered as potentially mutually statistically dependent, it is possible to gain substantially enhanced insight into the plant yield formation system under study (Lewis and Ward, 2013). When the analytical task is to identify statistical dependencies with one or more response variables, the additive Bayesian network structure discovery approach is well suited. The Bayesian network modeling approaches presented here are similar to those used by Lewis and McCormick (2012). Bayesian networks are graphical models comprising a set of conditionally independent generalized linear models combined in such a way as to be probabilistically coherent (i.e., no cycles in the graph), while maximizing the fit to observed data. The models aimed at formally describing interrelationships between explanatory variables and plant yield components.

All modeling was carried out in R, using the abn library (R Development Core Team, 2011; Lewis et al., 2011). Some dependence relationships between variables were banned from the structure discovery analysis to maintain a logical timeframe (e.g., planting date cannot be dependent on the emergence date, while the converse could be). Prior to model fitting, each variable was standardized to a mean of zero and standard deviation of one to account for the scale difference between features. This transformation had no effect on the identification of dependencies between variables (Neal, 1993).

2.3.1. Globally optimal model selection

The main purpose of Bayesian network structure discovery is to estimate the joint dependency structure of the random variables in the available data. For example, if X, Y and Z are three random variables, then a directed acyclic graph with serial connection between

nodes *X*, *Y*, *Z* (i.e., arc from *X* to *Y*, and from *Y* to *Z*) implies that $P(X,Y,Z) = P(X)P(Y|X)P(Z|Y)$. In data analyses, once the joint probability structure is known then we have complete information, i.e., given this we can then directly estimate any desired parameter or variable effect. An exact structure discovery approach was used to identify a globally optimal directed acyclic graph (Koivisto and Sood, 2004).

In order to determine the globally optimal model for each yam species (i.e., a model with the best goodness of fit to the observed data), it is assumed that all structures are equally supported in the absence of any data (an uninformative prior on structures). The log marginal likelihood (Mackay, 1992), typically referred to as the network score, was used to compare all models. Uninformative parameter priors were used throughout, specifically Gaussian distributions with means of zero and variance of 1000 for the marginal mean effects in each individual multivariable regression. By using uninformative priors, the structure discovery process effectively began from a point equivalent to no prior knowledge. Specifying and justifying informative parameter priors is impractical for every combination of variables across every model under comparison (Firestone et al., 2013). In these models, the goodness of fit (network score) and model parameters were estimated numerically rather than analytically using Laplace approximations at each node (Tierney and Kadane, 1986).

The globally optimal model (i.e., the model with the maximum goodness of fit over all possible directed acyclic graph structures before bootstrapping) had a total of 10 nodes and 17 arcs for *D. alata* and 12 nodes and 18 arcs for *D. rotundata* (Supplementary material, Fig. A6).

2.3.2. Adjustment for over-fitting

Once the globally optimal model has been identified then the next task is to check this model for over-fitting. As is the case with any model selection metric in multi-model selection, the marginal likelihood may identify structural features which, if the study was repeated many times, would likely only be recovered in a tiny fraction of instances (Babyak, 2004). To correct for such over-fitting, a parametric bootstrapping approach can be used in Bayesian network modeling (Friedman et al., 1999). This is conceptually straightforward, although computationally intensive, as for each simulated (bootstrap) dataset we need to repeat the exact same model search as that conducted with the original data. We took our chosen model—identified by applying the exact structure discovery search to the study data—and then simulated datasets from this, the same size as the original observed data, and checked how often the different structural features were recovered. These simulations were generated using open source JAGS software (Plummer, 2003) and the rjags library in R. We further removed all dependencies (arcs in the directed acyclic graph) which have insufficient statistical support to be considered robust, i.e., which were not recovered in at least a majority (50%) of the bootstrap results (Poon et al., 2007).

The result of the analyses was a statistically robust additive Bayesian network (henceforth called the final best directed acyclic graph) with the parameters in this model being exactly the same as in a standard multivariable logistic regression, except we then had many more of these, i.e., a set for each variable (Firestone et al., 2013). The parameters had the usual interpretation as posterior marginal mean effects for each covariate. The mean effects of the various variables in our study were estimated, along with their posterior 95% confidence intervals.

Ten thousand bootstrap datasets were generated and fitted using an identical exact model search. Pruning all arcs from our two globally optimal directed acyclic graphs, which were not recovered in at least 50% of the directed acyclic graphs based on the bootstrapping, resulted in the removal of three arcs for *D. alata* (Supplementary material Figs. A6 and 7).

The final best directed acyclic graphs for *D. alata* and *D. rotundata* comprised 14 and 18 arcs, respectively, with log marginal likelihoods of −4259 and −3631 (Supplementary material, Fig. A7).

2.3.3. Graphical representation

In structure discovery, the objective is to identify the factorization which best represents the study data, i.e., a Bayesian network represented visually by a directed acyclic graph. The directed acyclic graph comprises a set of nodes connected by directed links (arcs). Each node denotes a random variable and arcs define a given factorization of the joint probability distribution of all the random variables. The usual notation involves squares for discrete nodes and circles for continuous nodes. In a graphical model, all variables in the same component (collection of connected arcs-ignoring direction) are jointly statistically dependent. This means that knowing the value of one variable in this component can potentially generate information about likely values of any other variable in this component. If a variable has no arcs, either emanating from it or terminating at it, then it is statistically independent. Arcs in a Bayesian network model only denote statistical dependency, unless otherwise stated. Causal dependency can only be asserted using obvious real interpretation (like logical timeframe binding variables) or expert knowledge.

In the directed acyclic graphs, white nodes belong to the category of variables linked to practices at planting, black nodes belong to variables of plant yield components and gray nodes belong to variables linked to the plant's early growth. Solid arcs indicate a negative relationship between two variables while dashed arcs denote a positive relationship. Each arc is labeled with the standardized marginal posterior median and the frequency at which each arc was recovered in bootstrapping in brackets. The standardized median is the effect size indicator and gives indications on the direction (positive or negative) and the strength of the dependency between the two variables linked by the arc. The frequency at which each arc was recovered in the bootstrapping gives us the level of directional support between two variables (the maximum value possible being 10,000: 100% support).

3. Results

The complete dataset consisted of 439 observations for *D. alata* and 388 for *D. rotundata*. As the cataphyll occurrence was low in *D. alata* species (less than 2%), it was not possible to estimate marginal posterior densities for Cataphyll number and Cataphyll number per stem. Indeed these estimations presuppose that the data contains sufficient information to accurately estimate the joint probability distribution of the variables in the data. Consequently, these last two variables were dropped from the directed acyclic graph structure discovery for *D. alata*.

All the selected variables used in both models were justified using log marginal likelihood and bootstrapping. They showed at least one dependency (Figs. 2 and 3). The statistical confidence attributable to each dependency was given by its 95% confident interval (CI); that is, intervals containing the true value of the effect of one variable upon another with a probability of 95% (Fig. 4). All of the effect parameters had a 95% CI which did not pass through the origin (Fig. 4); these would therefore typically be considered as having a strong degree of statistical support. Globally, confidence intervals of *D. alata* were narrower allowing us to be more confident in the identified relationships than for *D. rotundata*.

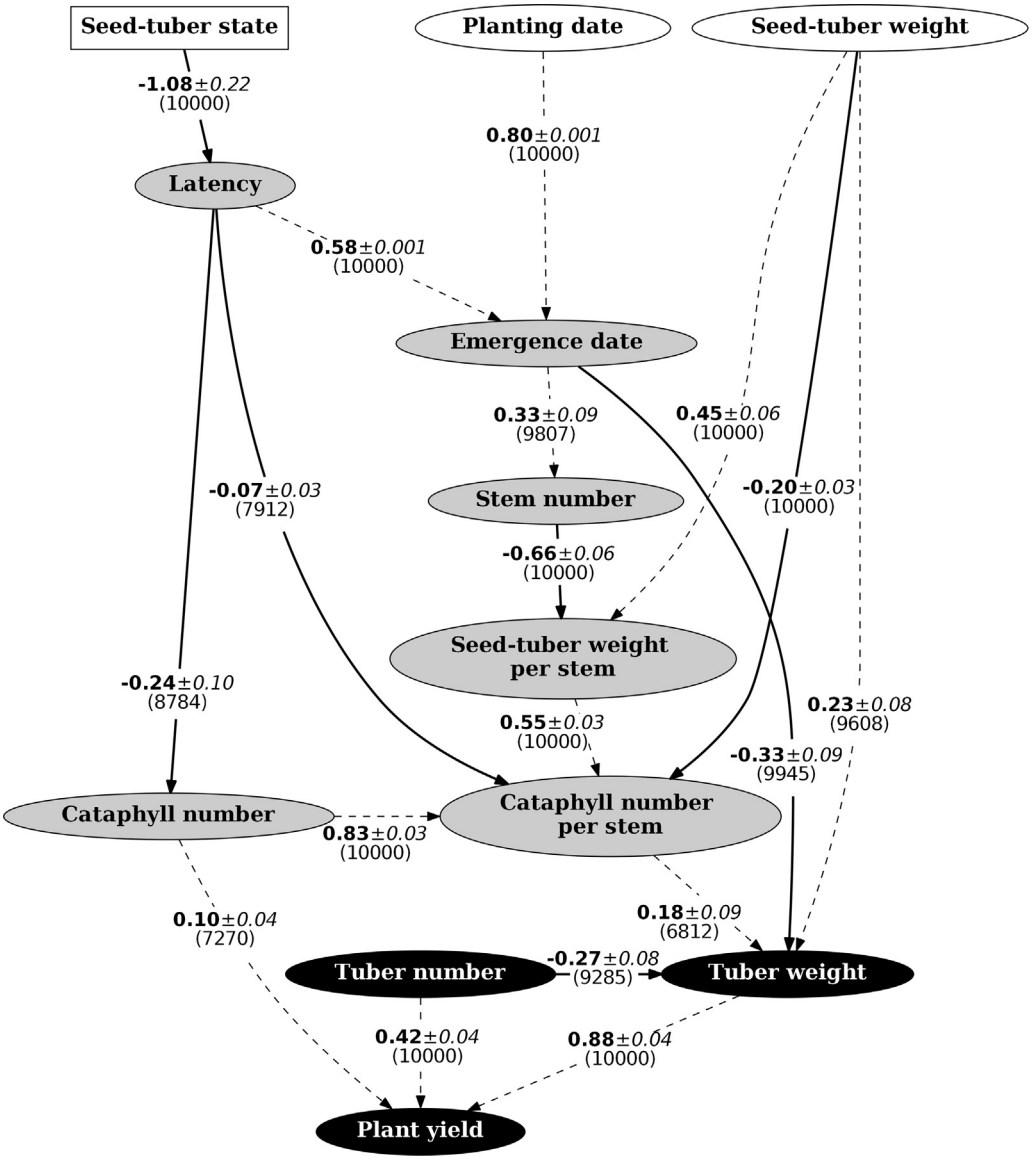

Fig. 2. Directed acyclic graph of the final best multivariate regression model for D. rotundata early growth variables on plant yield and yield components using an exact search additive Bayesian model. Solid arcs indicate a negative relationship between two variables, while dashed arcs denote a positive relationship. Arcs are labeled with the standardized median marginal posterior density and the frequency at which each arc was recovered in bootstrapping in brackets.

3.1. Emergence and early growth

The results of the analysis of the dry matter and nutrient content confirmed that the dry matter and nutrient content were highly variable within the seed-tuber lot (Fig. 5). The *D. rotundata* seed-tuber nutrient content was always higher but also more variable than that of *D. alata*.

For both species, the chain of events leading to emergence showed the following logical pattern: presprouted tubers had shorter latency which, with early planting, allowed early emergence.

For *D. rotundata* the cataphyll number decreased with the longer latency phase (Fig. 2).

For *D. alata*, the tuber state was dependent on the planting date (Fig. 3). Indeed, despite high 95% probability intervals, late planting showed more presprouted seed-tubers (Fig. 4). Thereafter, the stem number depended upon the emergence date for *D. rotundata* and upon the planting date for *D. alata*. The stem number and seed-tuber weight determined the seed-tuber weight per stem in both species. While the seed-tuber weight per stem in *D. alata* did not influence any other variable, for *D. rotundata* it determined the cataphyll number per stem, together with seed-tuber weight, latency and cataphyll number.

3.2. Plant yield components

The multivariable models provided statistical evidence that, for both species, the tuber number and emergence date were directly dependent on the final tuber weight (Figs. 2 and 3). The plant tuber weight decreased with increasing tuber number and earlier emergence date. In Fig. 2, the multivariate model provides evidence that the seed-tuber weight and the cataphyll number per stem were also directly ruled by the tuber weight, and therefore it was also obviously in the Markov blanket of the plant tuber weight of *D. rotundata*. Similarly, *D. alata* also had an additional variable to the common set of ascendants, i.e., the seed-tuber state (Fig. 3).

For *D. rotundata*, the tuber number did not show any ascendant, while for *D. alata*, it was directly and indirectly dependent

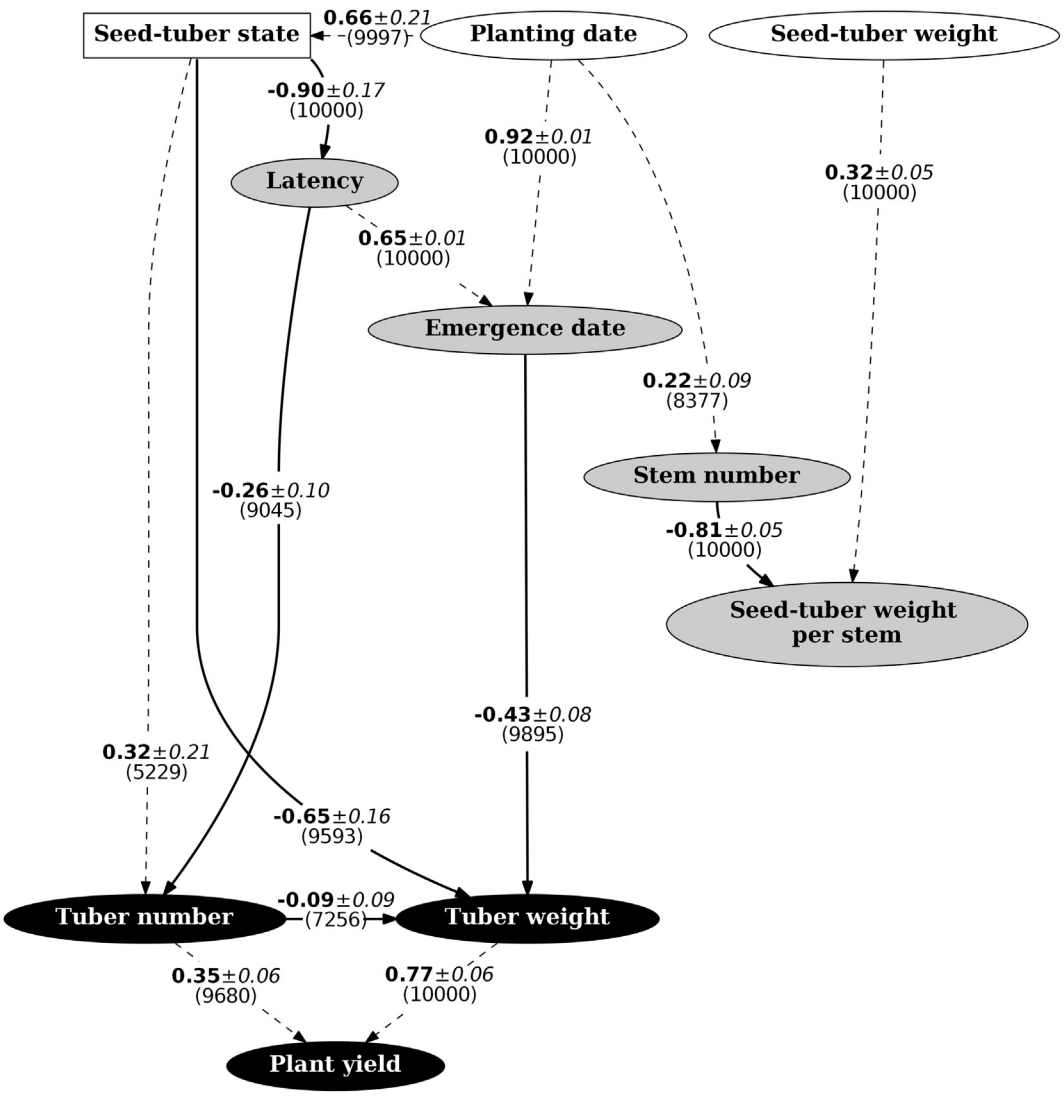

Fig. 3. Directed acyclic graph of the final best multivariate regression model for D. alata early growth variables on plant yield and yield components using an exact search additive Bayesian model. Solid arcs indicate a negative relationship between two variables, while dashed arcs denote a positive relationship. Arcs are labeled with the standardized median marginal posterior density and the frequency at which each arc was recovered in bootstrapping in brackets.

on the seed-tuber state through latency. The final plant yield of both species depended on the tuber number and tuber weight. For *D. rotundata* the cataphyll number also had a direct effect on the plant yield.

4. Discussion

Given the intrinsic property of directionality in Bayesian networks (Friedman and Koller, 2003), the directed acyclic graphs presented a general framework allowing for: (i) some new insight into the biological framework of emergence patterns leading to yam yield formation, (ii) the identification of opportunities for improvement and action priorities that will lead to improved yam cropping practices in West Africa, and (iii) consideration of the agronomic importance of yam seed-tuber quality.

4.1. Biological framework

Directed acyclic graphs offer a way of studying the influence of cropping practices and early plant growth on plant yield formation. This kind of multivariate analysis is uncommon in plant development (Nolivos et al., 2011), yet it enables us to tackle complex systems with a lack of expert knowledge, which is typically the case with yams. Both directed acyclic graphs confirmed the importance of planting practices and early growth on yield formation. This strong influence contrasts with what is known about other tuber crops such as potatoes. In fact the seed-tuber size range is also much wider for yams than for potato: 200–1000 g for yam in traditional West African cropping systems (depending on the available planting material), while it is more homogenous and much smaller (around 50 g) for potato seed-tubers in intensified cropping systems (Orkwor et al., 1998; Van Ittersum, 1992).

The emergence date is directly dependent on the planting date and latency. Given our effect size indicator (i.e., standardized median posterior), these dependencies were strong and positive. The later the planting date and the longer the latency, the later the emergence date. But for *D. rotundata* the directed acyclic graph allowed us to understand and quantify a more complex situation. Indeed, the planting date also had an indirect effect on the emergence date through the seed-tuber state. Later planting increased the number of presprouted tubers and thus decreased the latency. Thereafter, this shorter latency resulted in an earlier emergence date. Thus, until the breaking of seed-tuber dormancy, the planting date had a single direct and positive effect on the emergence

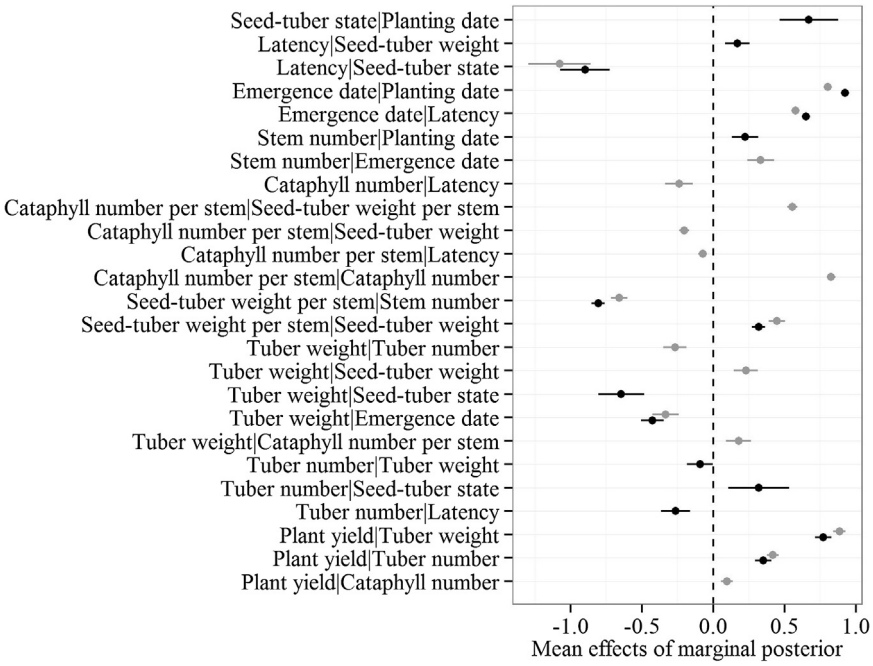

Fig. 4. Forest plot of the median effect (dot) of the marginal posterior distribution for D. rotundata (grey) and D. alata (black). Quantile-based 95% probability intervals are given using horizontal lines.

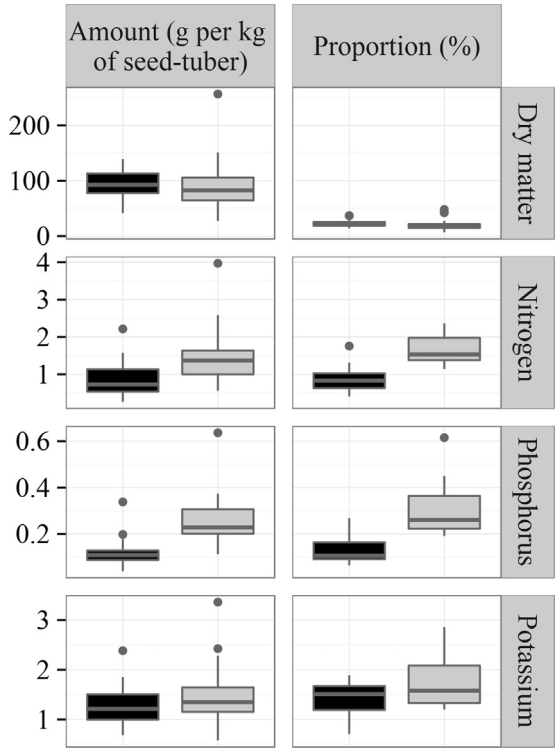

Fig. 5. Reserves in seed-tubers of D. alata (black) and D. rotundata (grey). The filled box corresponds to the inter-quartile range (IQR). The upper and lower whisker extends from the box to the highest and lowest value that is within 1.5 * IQR. Data beyond the end of the whiskers are outliers and plotted as points.

date. Afterwards, the planting date may have opposing direct and indirect effects that may lead to counter-intuitive results. With late rainy season onset, delayed planting often leads to the tuber sprouting and root and stem elongation in farmers' storage units (Orkwor, 1998).

For the descendent variables, the directed acyclic graph showed us that the emergence date was the only direct cause of plant tuber weight variability common to both species. This is in accordance with previous studies (Cornet et al., 2014; Marcos et al., 2011). But the directed acyclic graph highlights the direct effect of emergence rather than planting date. This means that, for a given planting date, the unevenly emerging stand still influenced the plant tuber weight. Given its influence on yield and yield variability in the field, this highlights a serious weakness in traditional cropping systems in West Africa, namely the uncontrolled and wide physiological age range in seed-tuber lots.

For *D. rotundata*, the emergence date also exhibited some complex indirect effects on plant tuber weight: the later the emergence, the higher the number of main stems, the lower the seed-tuber reserves per stem, the fewer the cataphyll number per stem and finally the lower the plant tuber weight. The direct contribution of the cataphyll number to the plant tuber weight and yield is a novel result that questions the emergence functional morphology in *D. rotundata*. As for cassava (*Manihot esculenta*), the emergence functional morphology is the result of selection and can be better understood by looking at the wild relatives (Pujol et al., 2005). The wild relatives of *D. rotundata* (i.e., *D. prahensilis*) originated from forested zones of West and Central Africa where, once initiated, the shoot must grow through a poorly illuminated understory (Di Giusto et al., 2001). Cataphylls thus provide a source of buds without placing a high respiration load on the seedling, which would result if a lamina developed (Wright et al., 2000). The number of nodes carrying cataphylls is thus dependent on the seed-tuber reserve, i.e., the seed-tuber size and also the seed-tuber carbohydrate and mineral nutrient contents. In agreement with this, the results indicated that seed-tuber reserves varied within a seed-tuber lot, even between equally sized seed-tubers (Fig. 5). In the model, the cataphyll number could thus represent the nutritional status of the plantlet related to the seed-tuber reserves unexplained by seed-tuber weight. As the plants develop, the cataphylls could be an indicator of the quantity and the time of seed-tuber reserve utilization by the plant. Applications of this relationship are not only of interest in functional morphology but there could also be use-

ful applications as an indicator for phenotyping early growth vigor or as an indicator of the end point of early plant growth (Hanley et al., 2004). Moreover, this result suggests that the cataphyll number per stem and seed-tuber weight should be included along for further investigation into their potential biological significance for controlling plant growth and final tuber weight.

4.2. Agronomic opportunities for improvement and action priorities

The seed-tuber weight influenced the plant tuber weight for *D. rotundata* but not for *D. alata*, which is not in agreement with the results reported by Njoku et al. (1984). Ferguson (1973) showed an asymptotic response of *D. alata* to seed-tuber size from 80 up to 250 g, while Kayode (1984) showed that seed-tubers larger than 400 g could be used to obtain a maximum *D. rotundata* tuber yield. In fact *D. alata* is a native of south-east Asia and was introduced later into West Africa (Doumbia, 2004). It seems that traditional practices transferred from *D. rotundata* may have led to oversized *D. alata* seed-tubers. For instance, in the Caribbean, *D. alata* is planted using much smaller seed-tubers (i.e., 100 g; Marcos et al., 2011). In order to avoid wastage, it could be necessary to look for the optimum seed-tuber size for *D. alata* in West Africa.

As to the influence of the seed-tuber state on the plant yield component, it seems that delayed planting (i.e., after seed-tuber sprouting) might drastically decrease the marketable yield. Indeed, presprouted seed-tubers of *D. alata* caused an increase in tuber number and a decrease in tuber weight. Yam prices on West African markets are based on tuber size, with bigger tubers attracting a higher price (Orkwor et al., 1998). On the other hand, it could be beneficial for farmers to manage their seed-tuber stocks on the basis of the seed-tuber state. Farmers could use unsprouted seed-tubers for ware yam production. Keeping the presprouted ones for seed-tuber production could enable farmers to produce more and smaller tubers for the next year's planting.

4.3. Importance of seed-tuber quality for yam production

The importance of better control of the seed tuber quality comes out of this discussion. This is obvious both for the explanatory variables like seed-tuber weight and the uncontrolled latent variables such as the physiological age and seed-tuber reserves. The need to develop seed-yam systems in West Africa is currently well understood and taken into account by donors and international development agencies (WECARD, 2011; DFID, 2014; IITA, 2014). Although most projects focus on clean (healthy) seed yam production, directed acyclic graph representations of the consequences of using uncontrolled planting material clearly supports the use of clean but also good quality yam seed-tubers (i.e., calibrated, uniform in terms of physiological age and nutrient content).

Cornet et al. (2014) claimed that until we have a more effective production system of quality seed-tubers, cohorts could be a suitable experimental unit for analyzing processes in field yam populations. A cohort can be defined as a group of individuals at the same phenological stage (Deaton and Winebrake, 2000). Cornet et al. (2014) advised building cohorts based on the emergence date. We recommend, based on our results, refining the yam cohort concept by grouping plants based on their belonging to similar categories of emergence date and cataphyll number.

5. Conclusions

The use of a Bayesian network makes it possible to represent complex systems for non-experts in a way that facilitates automated analysis. Directed acyclic graphs offer a way of studying the dependency between cultural practices, early plant growth variables and plant yield components. All of the selected variables used in both models were statistically justified and could be considered as having a strong degree of statistical support. Directed acyclic graphs have a general framework to enhance insight into the yam biological framework, and identify opportunities for cropping system improvement.

For *D. rotundata*, the direct contribution of the cataphyll number to the plant tuber weight and yield is a novel result that questions the emergence functional morphology. Cataphyll number could be an indicator of the quantity of seed-tuber reserves and the timing of their utilization by the plant. Applications of this relationship are not only of interest in functional morphology but could also find useful applications in other disciplines (e.g., phenotyping).

For *D. alata*, the model did not show any effect of seed-tuber size. It is suggested that traditional practices transposed from native *D. rotundata* may have led to oversized *D. alata* seed-tubers and that reducing the seed-tuber size may lead to less wastage. We demonstrate the influence of the *D. alata* seed-tuber state on yield components, directly and indirectly through latency. This relationship explains why delayed planting (i.e., after seed-tuber sprouting) might drastically decrease the marketable yield. It also provides basis of understanding needed to improve the farmers' seed-tuber stock management.

Both directed acyclic graphs confirmed the influence of planting practices and early growth on yam (*Dioscorea* spp.) yield formation in West Africa. Given the influence of early growth on plant yield variability, this study demonstrated a serious weakness in traditional cropping systems (i.e., the uncontrolled and wide range of physiological ages and reserves in seed-tuber lots), and highlighted the importance of better control of seed tuber quality.

Acknowledgments

We thank M. Sodanhoun, C. Adiba, D. Damissi and J. Lawson for field assistance, Dr R. Bonhomme for helpful comments on the manuscript, and A. Scaife and D. Manley for the English editing.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.eja.2016.01.009.

References

Asiedu, R., Sartie, A., 2010. Crops that feed the World 1 Yams. Yams for income and food security. Food Secur. 2, 305–315, http://dx.doi.org/10.1007/s12571-010-0085-0.

Ayensu, E.S., 1972. Anatomy of Monocotyledon. Vi Dioscoreales. Oxford University Press, London, UK.

Babyak, M.A., 2004. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom. Med. 66 (3), 411–421, http://dx.doi.org/10.1097/01.psy.0000127692.23278.a9.

Cook, J., Stefanski, L.A., 1994. A simulation extrapolation method for parametric measurement error models. J. Am. Stat. Assoc. 89, 464–467.

Cornet, D., Sierra, J., Bonhomme, R., 2007. Characterization of the photosynthetic pathway of some tropical food yams (*Dioscorea* spp.) using leaf natural ^{13}C abundance. Photosynthetica 45, 303–305, http://dx.doi.org/10.1007/s11099-007-0050-0.

Cornet, D., Sierra, J., Tournebize, R., Ney, B., 2014. Yams (*Dioscorea* spp.) plant size hierarchy and yield variability: emergence time is critical. Eur. J. Agric. 55, 100–107, http://dx.doi.org/10.1016/j.eja.2014.02.002.

DFID, 2014. Giving seed-yams the credit they deserve. Project CPP25. http://r4d.dfid.gov.uk/PDF/Outputs/ResearchIntoUse/CPP25.pdf (last accessed 21.03.14).

Di Giusto, B., Anstett, M.C., Dounias, E., Doyle, B.M., 2001. Variation in the effectiveness of biotic defence: the case of an opportunistic ant-plant protection mutualism. Oecologia 129, 367–375, http://dx.doi.org/10.1007/s004420100734.

Doumbia, S., Tshiunza, M., Tollens, E., Stessens, J., 2004. Rapid spread of the Florido yam variety (*Dioscorea alata*) in Ivory Coast. Introduced for the wrong reasons and still a success. Outlook Agric. 33 (1), 49–54, http://dx.doi.org/10.5367/000000004322877773.

Fayaud, B., Coste, F., Corre-Hellou, G., Gardarin, A., Dürr, C., 2014. Modelling early growth under different sowing conditions: a tool to predict variations in intercrop early stages. Eur. J. Agric. 52, 180–190, http://dx.doi.org/10.1016/j.eja.2013.09.009.

Ferguson, T.U., 1973. The effect of sett characteristics and spacing on growth, development and yield of yams (*Dioscorea* spp.). In: PhD Thesis. University of the West Indies, St. Augustine, West Indies.

Firestone, S.M., Lewis, F.I., Schemann, K., Ward, M.P., Toribio, J.A.L.M.L., Dhand, N.K., 2013. Understanding the associations between on-farm biosecurity practice and equine influenza infection during the 2007 outbreak in Australia. Prev. Vet. Med. 110, 28–36, http://dx.doi.org/10.1016/j.prevetmed.2013.02.003.

Friedman, N., Goldszmidt, M., Wyner, A., 1999. Data analysis with Bayesian networks: a bootstrap approach. In: Kaufmann, M. (Ed.), Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. San Francisco, USA, pp. 196–205.

Friedman, N., Koller, D., 2003. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. Mach. Learn. 50 (1–2), 95–125, http://dx.doi.org/10.1023/A:1020249912095.

Hanley, M.E., Fenner, M., Whibley, H., Darvill, B., 2004. Early plant growth: identifying the end point of the seedling phase. New Phytol. 163, 61–66, http://dx.doi.org/10.1111/j.1469-8137.2004.01094.x.

He, W., Xiong, J., Yi, G.Y., 2012. SIMEX R package for accelerated failure time models with covariate measurement error. J. Stat. Softw. 46, 1–14.

Heckerman, D., Geiger, D., Chickering, D.M., 1995. Learning Bayesian networks—the combination of knowledge and statistical. Data Mach. Learn. 20, 197–243, http://dx.doi.org/10.1023/A:1022623210503.

IITA, 2014. Project YIISWA: Yam improvement for income and food security in West Africa, http://www.iita.org/web/yiifswa/home (last accessed 21.03.14).

Kayode, G.O., 1984. Effects of sett size and spacing on tuber yield of white Guinea yam (*Dioscorea rotundata*) in the rainforest and savanna zones of Nigeria. Exp. Agric. 20, 53–57, http://dx.doi.org/10.1017/S0014479700017580.

Kennedy, D., 2003. Agriculture and the developing world. Science 302, 357, http://dx.doi.org/10.1126/science.302.5644.357.

Kim, Y., Yoo, S., Gu, Y., Lim, J., Han, D., Baik, S., 2014. Crop pests prediction method using regression and machine learning technology: survey. IERI Procedia 6, 52–56.

Klemke, T., Moll, A., 1990. Model for simulation of potato growth from planting to emergence. Agric. Syst. 32, 295–304, http://dx.doi.org/10.1016/0308-521X(90)90096-9.

Koivisto, M., Sood, K., 2004. Exact Bayesian structure discovery in Bayesian networks. J. Mach. Learn. Res. 5, 549–573, http://dx.doi.org/10.1145/203330.203334.

Korb, K.B., Nicholson, A.E., 2004. Bayesian Artificial Intelligence. Chapman and Hall/CRC, Boca Raton, US.

Lewis, F.I., Brulisauer, F., Gunn, G.J., 2011. Structure discovery in Bayesian networks: an analytical tool for analysing complex animal health data. Prev. Vet. Med. 100 (2), 109–115, http://dx.doi.org/10.1016/j.prevetmed.2011.02.003.

Lewis, F.I., McCormick, B.J.J., 2012. Revealing the complexity of health determinants in resource-poor settings. Am. J. Epidemiol. 176 (11), 1051–1059, http://dx.doi.org/10.1093/aje/kws183.

Lewis, F.I., Ward, M.P., 2013. Improving epidemiologic data analyses through multivariate regression modelling. Emerg. Themes Epidemiol. 10, 4–14, http://dx.doi.org/10.1186/1742-7622-10-4.

Lobell, D.B., Ortiz-Monasterio, J.I., 2007. Impacts of day versus night temperatures on spring wheat yields: a comparison of empirical and CERES model predictions in three locations. Agron. J. 99 (2), 469–477.

Lobell, D.B., 2013. Errors in climate datasets and their effects on statistical crop models. Agric. For. Meteorol. 170, 58–66.

Mackay, D.J.C., 1992. Bayesian interpolation. Neural Comput. 4 (3), 415–447, http://dx.doi.org/10.1007/978-94-017-2219-3_3.

Marcos, J., Lacointe, A., Tournebize, R., Bonhomme, R., Sierra, J., 2009. Water yam (*Dioscorea alata* L.) development as affected by photoperiod and temperature: experiment and modeling. Field Crops Res. 111 (3), 262–268, http://dx.doi.org/10.1016/j.fcr.2009.01.002.

Marcos, J., Cornet, D., Bussière, F., Sierra, J., 2011. Water yam (*Dioscorea alata* L.) growth and yield as affected by the planting date: experiment and modeling. Eur. J. Agric. 34, 247–256, http://dx.doi.org/10.1016/j.eja.2011.02.002.

Nash, D., Riffkin, P., Harris, R., Blackburn, A., Nicholson, C., McDonald, M., 2013. Modelling gross margins and potential N exports from cropland in south-eastern Australia. Eur. J. Agric. 47, 23–32.

Neal, R.R., 1993. Bayesian learning via stochastic dynamics. In: Giles, C.L., Hanson, S.J., Cowan, J.D. (Eds.), Advances in Neural Information Processing Systems, 5. Morgan Kaufmann, San Mateo, USA, pp. 475–482.

Needham, C.J., Bradford, J.R., Bulpitt, A.J., Westhead, D.R., 2007. A primer on learning in Bayesian networks for computational biology. PLoS Comput. Biol. 3 (8), 1409–1416, http://dx.doi.org/10.1371/journal.pcbi.0030129.

Njoku, J.E., Nwoke, F.I.O., Okonkwo, S.N.C., 1984. Pattern of growth and development in *Dioscorea rotundata* Poir. Trop. Agric. 61 (1), 17–19.

Nolivos, I., Van Biesen, L., Swennen, R.L., 2011. Modelling an intensive banana cropping system in Ecuador using a Bayesian Network. Acta Hortic. 919, 89–98.

Okezie, C.E.A., Nwoke, F.I.O., Okonkwo, S.N.C., 1986. Field studies on the growth pattern of *Dioscorea rotundata* Poir propagated by seed. Trop. Agric. 63 (1), 22–24.

Onwueme, I.C., 1976. Performance of yam (*Dioscorea* spp.) planted without water. J. Agric. Sci. Cambridge 87, 413–415.

Onwueme, I.C., 1978. Set weight effects on time of tuber formation and on tuber yield characteristics in water yam (*Dioscorea alata* L.). J. Agric. Sci. 91, 317–319, http://dx.doi.org/10.1017/S0021859600046402.

Orkwor, G.C., Asiedu, R., Ekanayake, I.J., 1998. Food Yams: Advances in Research. IITA and NRCRI, Ibadan, Nigeria.

Plummer, M., 2003. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing, March 20–22, Vienna, Austria.

Poon, A.F.Y., Lewis, F.I., Pond, S.L.K., Frost, S.D.W., 2007. Evolutionary interactions between N-linked glycosylation sites in the HIV-1 envelope. PLoS Comput. Biol. 3, 110–119, http://dx.doi.org/10.1371/journal.pcbi.0030011.

Porth, I., Klapste, J., Skyba, O., Friedmann, M.C., Hannemann, J., Ehlting, J., El-Kassaby, Y.A., Mansfield, S.D., Douglas, C.J., 2013. Network analysis reveals the relationship among wood properties, gene expression levels and genotypes of natural *Populus trichocarpa* accessions. New Phytol. 200 (3), 727–742, http://dx.doi.org/10.1111/nph.12419.

Pujol, B., Mühlen, G., Garwood, N., Horoszowski, Y., Douzery, E.J.P., McKey, D., 2005. Evolution under domestication: contrasting functional morphology of seedlings in domesticated cassava and its closest wild relatives. New Phytol. 166, 305–318, http://dx.doi.org/10.1111/j.1469-8137.2004.01295.x.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (accessed 21.3.14.) http://www.R-project.org.

Rebetzke, G.J., Richards, R.A., Fettell, N.A., Long, M., Condon, A.G., Botwright, T.L., 2007. Genotypic increases in coleoptile length improves stand establishment, vigour and grain yield of deep-sown wheat. Field Crops Res. 11, 10–23, http://dx.doi.org/10.1016/j.fcr.2006.05.001.

Scott, G.J., Rosegrant, M.W., Ringler, C., 2000. Global projections for root and tuber crops to the year 2020. Food Policy 25, 561–597, http://dx.doi.org/10.1016/S0306-9192(99) 00087-1.

Sheehy, J.E., Mitchell, P.L., Ferrer, A.B., 2006. Decline in rice grain yields with temperature: models and correlations can give different estimates. Field Crops Res. 98 (2–3), 151–156.

Tierney, L., Kadane, J.B., 1986. Accurate approximations for posterior moments and marginal densities. J. Am. Stat. Assoc. 81 (393), 82–86, http://dx.doi.org/10.1080/01621459.

Van Ittersum, M.K., 1992. Variation in the duration of tuber dormancy within a seed potato lot. Potato Res. 35, 261–269, http://dx.doi.org/10.1007/BF02357706.

Ward, M.P., 2013. Bayesian graphical modelling: applications in veterinary epidemiology. Prev. Vet. Med. 110, 1–3, http://dx.doi.org/10.1016/j.prevetmed.2013.02.007.

WECARD, 2011. Promotion of improved yam minisett technology to improve productivity and reduce excessive use of food yam for planting in West Africa. http://www.coraf.org/database/projet/programmedetail.php?detail=SC/02/CP/USAID/2009-11/ (last accessed 21.03.14.).

Wright, I.J., Clifford, H.T., Kidson, R., Reed, M.L., Rice, B.L., Westoby, M., 2000. A survey of seed and seedling characters in 1744 Australian dicotyledon species: cross-species trait correlations and correlated trait-shifts within evolutionary lineages. Biol. J. Linn. Soc. 69, 521–547, http://dx.doi.org/10.1111/j.1095-8312.2000.tb01222.x.

Zhu, Y.G., Liu, D.Y., Chen, G.F., Jia, H.Y., Yu, H.L., 2013. Mathematical modeling for active and dynamic diagnosis of crop diseases based on Bayesian networks and incremental learning. Math. Comput. Model 58, 514–523.