Using Bayesian Networks to Predict the Success of Wheat Yields in the Free State

Holly Alice Judge
Tariro Nathan Banganayi

March 2023

1 Introduction

Bayesian networks are probabilistic models that represent a set of random variables and their conditional dependencies using a directed acyclic graph. Bayesian networks are powerful because they allow us to map probabilistic distributions that are conditioned on the relationships between observed and unobserved random variables. They specifically allow us to make inferences on the probabilistic distributions of unobserved random variables using information that we can feed to the model as a prior.

Bayesian networks can be used in the agricultural industry to otpimise the production of crops. The Free State province of South Africa is the country's largest producer of wheat. As well as being a a staple food in diets for households across the country, wheat is a massive source of income for the entire agricultural industry. Farming wheat is an incredibly resource intensive and volatile activity because of the large network of dependencies that affect the success of a wheat yield. These dependencies can come from environmental factors that farmers have little control over such as the weather and climate, as well as factors that are within the control of farmers such as fertilisation and genetic modification.

Climate change and drought have created the necessity for hardier wheat crops that are able to withstand harsher, more stressful conditions. Genetic modification is an increasingly common solution to the hardships faced by farmers as it allows for crops to be modified to be resistant to certain stresses. Genetic modification in commercial agriculture is expensive and highly regulated and should therefore only be utilised with due cause. The objective of this network is to determine the extent to which the introduction of the SeCspA gene to the wheat crop will increase the yield of wheat under stressful conditions.

The SeCspA gene modifies Cold Shock Proteins in the wheat crop to allow for resistance to water-scarce and cold conditions. It is costly to introduce and therefore can only be justified under the right conditions. The network will model the probability and utility of the genetic modification under different observations of random variables associated with the conditions of wheat production in the Free State to see which conditions would justify the introduction of the gene.

The network will be used by commercial wheat farmers in the Free State as the probability distributions are specifically sourced from data in the Free State. With minor adjustments however, the network can be applied to any region or crop to predict the probability of a successful yield. The network can also be used by regulatory bodies to decide whether to approve a genetically modified seed. The potential use cases of this network are to model the utility of genetic modification in the face of drought and extreme temperature as well as to allow farmers to better understand the relationships between the variables that affect their yields at the end of a season.

2 Problem analysis

Soil quality and drought resistance are among the predominant characteristics that affect the success of a yield in the South African wheat industry. These characteristics themselves are dependant on other random variables that can be observed or unobserved.

There are 18 nutrients essential for proper crop development [1]. The absence of any one of these nutrients has the potential to decrease crop yield by negatively affecting the crop's growth [1]. Soil is a major source of these nutrients to crops [1]. Therefore, soil quality has a direct and profound affect on the crop yield. Soil moisture is one of the main factors influencing soil nutrients [2]. Therefore, soil quality is dependent on soil moisture. Soil moisture is the water stored in the soil and is affected by precipitation and temperature [2]. Therefore, in our Bayesian network soil moisture is dependent on precipitation and temperature. Common knowledge suggests that temperature is dependent on precipitation and sunlight.

In the South African context, a crop's ability to survive in drought-like conditions has a direct impact on crop yield. The mineral potassium (K) reduces the deleterious effects of drought stress on plants by the most significant factor [3], hence drought resistance is heavily dependent on potassium levels in the soil. Cold shock proteins (CSPs) enhance acclimatisation of bacteria to severe environmental circumstances, such as droughts [4]. The *Escherichia coil* CSP genes CspA and CspB have been modified to plant-preferred codon sequences and named SeCspA and SeCspB [4]. Scientific investigations have shown SeCspA transgenic wheat lines possess significant and stable improvements in drought tolerance when compared to control plants not containing the gene [4]. Because SeCspA's confer drought tolerance, drought resistance is dependent on the crops being genetically modified to include this protein.

This model assumes independence between the crops' resistance to drought and soil quality. Factors external to this model that could impact the outcomes include natural disasters and pests.

2.1 Data sets

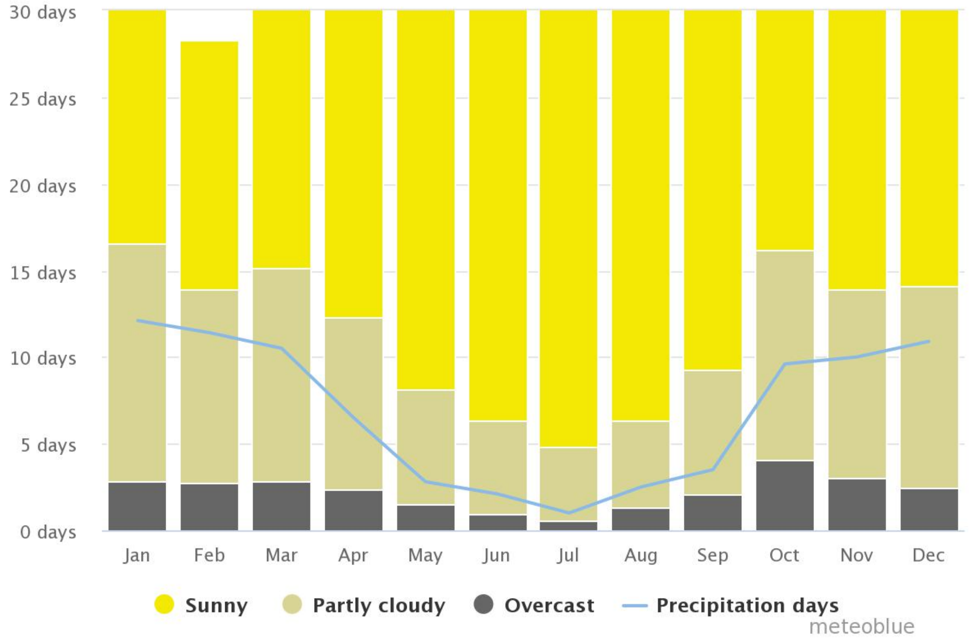


Figure 1: Average sunlight and precipitation in the Free State from 1993 to 2023 [5]

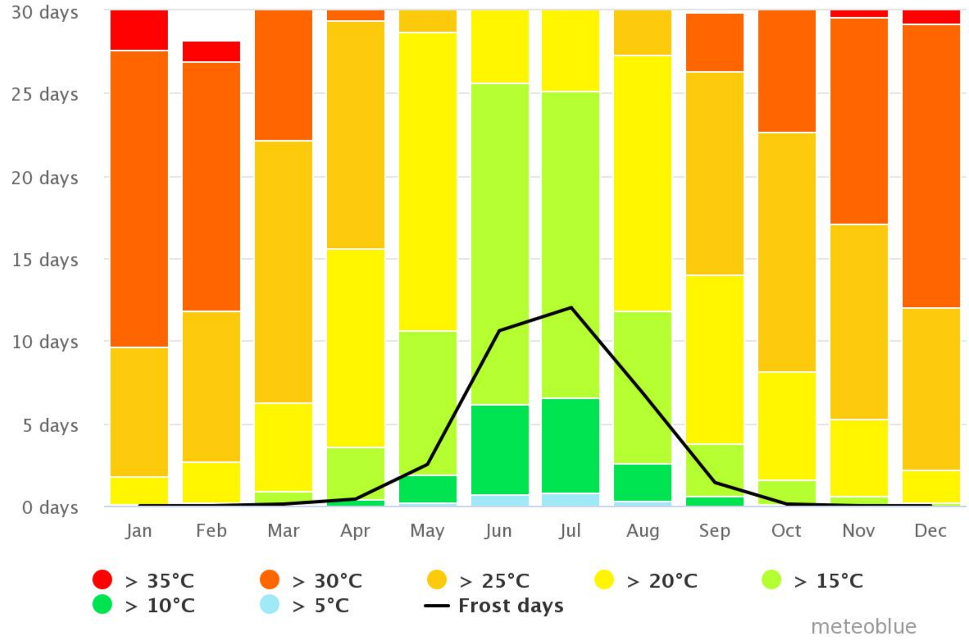


Figure 2: Average temperature in the Free State from 1993 to 2023 [5]

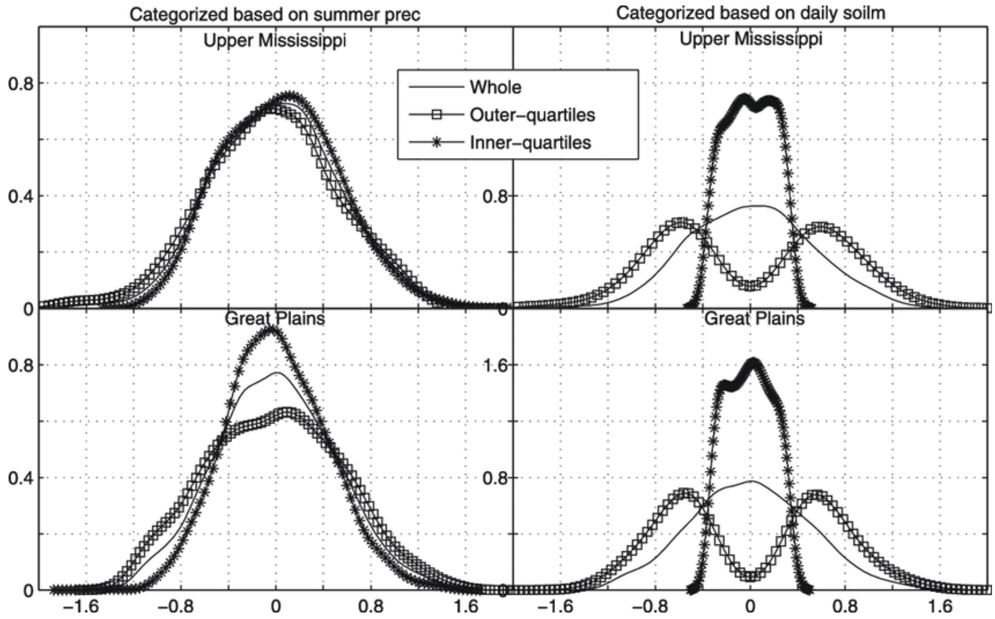
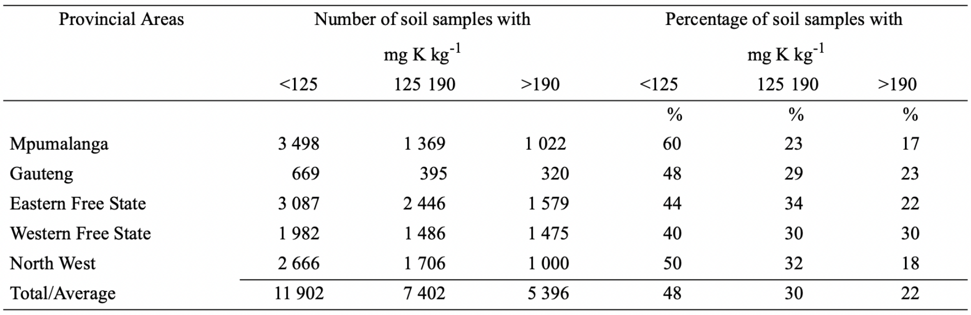


Figure 3: Probability distribution function of daily soil moisture over the two regions under two categorisations (both include outer and inner quartiles): one based on summer total precipitation and the other based on daily soil moisture. [6]



Provincial Areas	Number of soil samples with mg K kg^{-1}			Percentage of soil samples with mg K kg^{-1}		
	<125	125 190	>190	<125	125 190	>190
				%	%	%
Mpumalanga	3 498	1 369	1 022	60	23	17
Gauteng	669	395	320	48	29	23
Eastern Free State	3 087	2 446	1 579	44	34	22
Western Free State	1 982	1 486	1 475	40	30	30
North West	2 666	1 706	1 000	50	32	18
Total/Average	11 902	7 402	5 396	48	30	22

Figure 4: The number and percentage of top soil samples received during three seasons (2003-2005) from farmers in the maize growing areas of five provinces containing less than 125, between 125 and 190, and more than 190 mg K kg^{-1}[3]

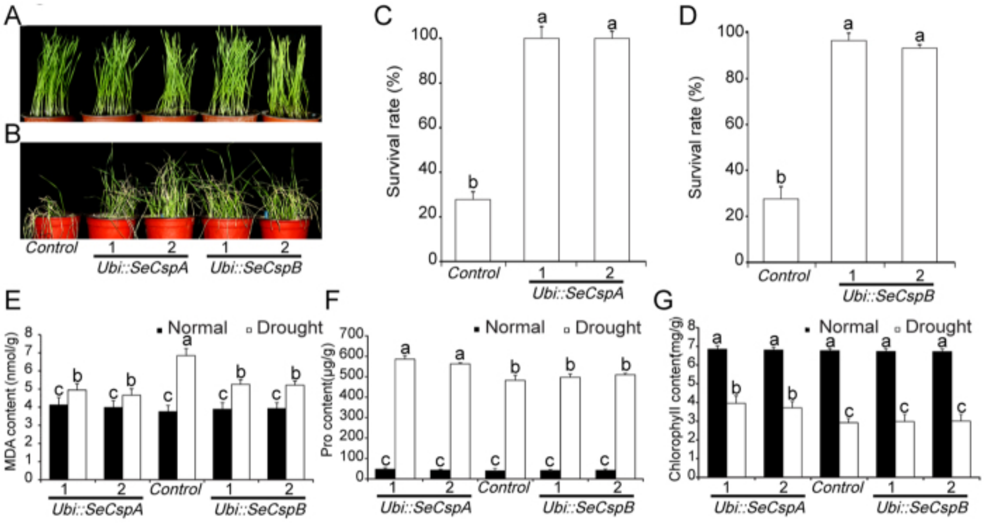


Figure 5: Phenotypes of transgenic wheat lines and parental wheat plants (control) under drought stress. (A) Phenotypes of transgenic wheat lines grown under normal conditions for three weeks. (B) Phenotypes of the transgenic wheat lines after rehydration for one week. (C) and (D) Survival rates of the transgenic wheat lines. (E) MDA content. (F) Free proline content. (G) Chlorophyll content. Measurements were made after 7-day-old plants were treated with drought conditions for one week. Vertical bars bearing different letters in (C,D,E,F), and (G) indicate significant differences between transgenic and control wheat plants [4]

3 Decision Network Model

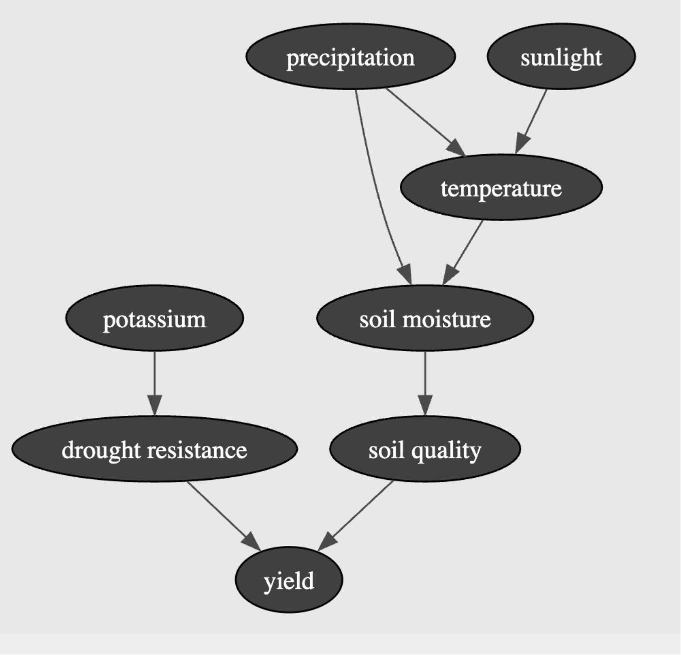


Figure 6: Bayesian Network

3.1 Structure

A successful yield is dependent on drought resistance and soil quality, with a high probability of good soil quality and a high probability of drought resistance resulting in a high probability of a successful yield.
Soil quality is dependent on soil moisture, where increased soil moisture results in higher soil quality. Soil moisture is dependent on precipitation and temperature. A high probability of precipitation increases the probability of moist soil. A low probability of high temperatures increases the probability of moist soil. Temperature is dependent on sunlight and precipitation. A high probability of sunlight increases the probability of high temperatures, whereas a high probability of precipitation decreases the probability of high temperatures.
The crops' resistance to drought is dependent on the potassium levels in the soil, where a high probability of potassium increases the crops resistance to drought. The crops resistance to drought is also dependent on whether the crop contains the SeCspA gene, with a high probability of the crop containing the gene correlating to a high probability of drought resistance.

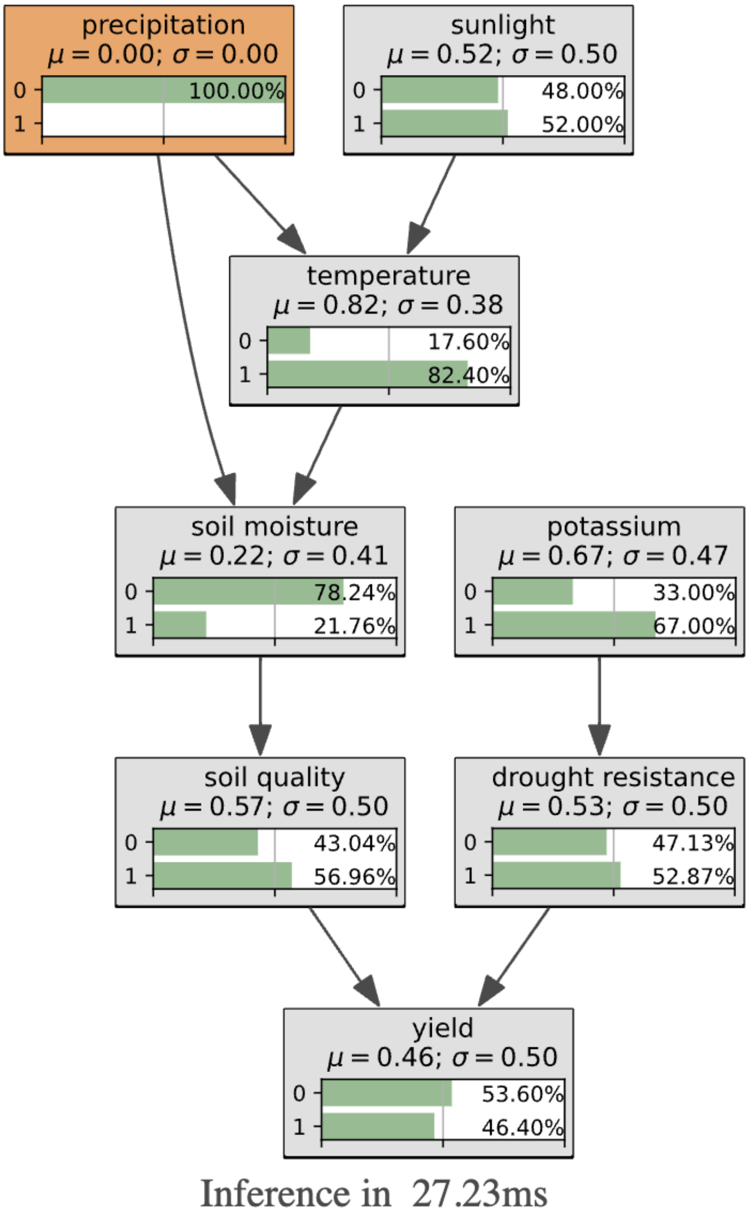
3.2 Weights

Our weightings were informed by common knowledge, the data sets, and the following statistics derived from expert knowledge.

Potassium: South African farmers only apply two thirds of the potassium that is annually removed in the grain [3].
Precipitation: Using Figure 1, it can be derived that during the wheat growing season, October to December, there is a 33% probability of precipitation.
Sunlight: Using Figure 1, it can be derived that during the wheat growing season, October to December, there is a 52% probability of sunlight.
Temperature: Using Figure 2, it can be derived that during the wheat growing season, October to December, there is a 98% probability it will be hot (above 20 degrees Celsius).
SeCspA: The germination rates of the seeds with SeCspA were more that 62.3% under polyethylene glycol-simulated drought stress compared to the control seeds which had a growth rate of 37.5% [4]. These results demonstrate that the growth rate of seeds with SeCspA is 40% higher than the growth rate of seeds without SeCspa under drought conditions.

4 Model Testing and Evaluation

4.1 Inferences



precipitation
$\mu = 0.00; \sigma = 0.00$
0
1
100.00%
sunlight
$\mu = 0.52; \sigma = 0.50$
0
1
48.00%
52.00%
temperature
$\mu = 0.82; \sigma = 0.38$
0
1
17.60%
82.40%
soil moisture
$\mu = 0.22; \sigma = 0.41$
0
1
78.24%
21.76%
potassium
$\mu = 0.67; \sigma = 0.47$
0
1
33.00%
67.00%
soil quality
$\mu = 0.57; \sigma = 0.50$
0
1
43.04%
56.96%
drought resistance
$\mu = 0.53; \sigma = 0.50$
0
1
47.13%
52.87%
yield
$\mu = 0.46; \sigma = 0.50$
0
1
53.60%
46.40%
Inference in 27.23ms

Figure 7: Inference when low precipitation is observed

The power of this Bayesian network lies in its ability to perform inference based on evidence and observation of random variables. We can prove that the network is capable of making useful inferences by passing observation parameters into the network at certain key points in the directed acyclic graph. We will predict the likelihood of a successful yield under two observed conditions. The first observation (Figure 7) is one in which we observe low precipitation. This is indicative of drought conditions. The posterior for the random variable 'yield' in this scenario is: 0.46. This means there is a 46% chance of a successful yield given low precipitation. Next we will observe low precipitation and high drought-resistance (Figure 8) in the crop. The posterior for the random variable 'yield' in this scenario is: 0.65. This means there is a 65% chance of a successful yield given low precipitation and high drought-resistance.

One can make inferences further up the DAG as well in order to find the probability distributions of a different unobserved random variable. In the next example we will make inferences about the soil quality under certain observations. When low precipitation is observed (Figure 7) the posterior for the random variable 'soil quality' is 0.57. That means there is a 57% chance of good soil quality when there is low precipitation. If we observe low precipitation and low temperatures (Figure 9) the posterior for the random variable 'soil quality' is reported as

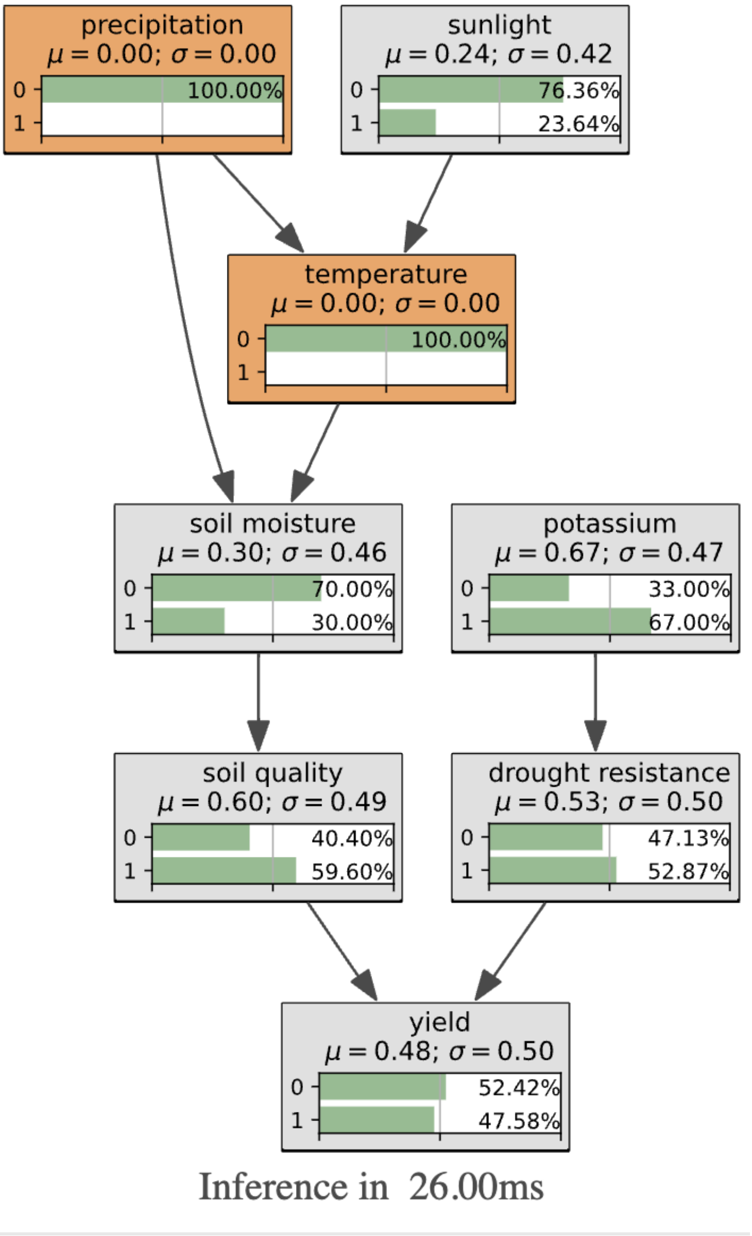

precipitation
$\mu = 0.00; \sigma = 0.00$
0
1
100.00%
sunlight
$\mu = 0.24; \sigma = 0.42$
0
1
76.36%
23.64%
temperature
$\mu = 0.00; \sigma = 0.00$
0
1
100.00%
soil moisture
$\mu = 0.30; \sigma = 0.46$
0
1
70.00%
30.00%
potassium
$\mu = 0.67; \sigma = 0.47$
0
1
33.00%
67.00%
soil quality
$\mu = 0.60; \sigma = 0.49$
0
1
40.40%
59.60%
drought resistance
$\mu = 0.53; \sigma = 0.50$
0
1
47.13%
52.87%
yield
$\mu = 0.48; \sigma = 0.50$
0
1
52.42%
47.58%
Inference in 26.00ms

Figure 8: Inference when low precipitation and low temperature are observed

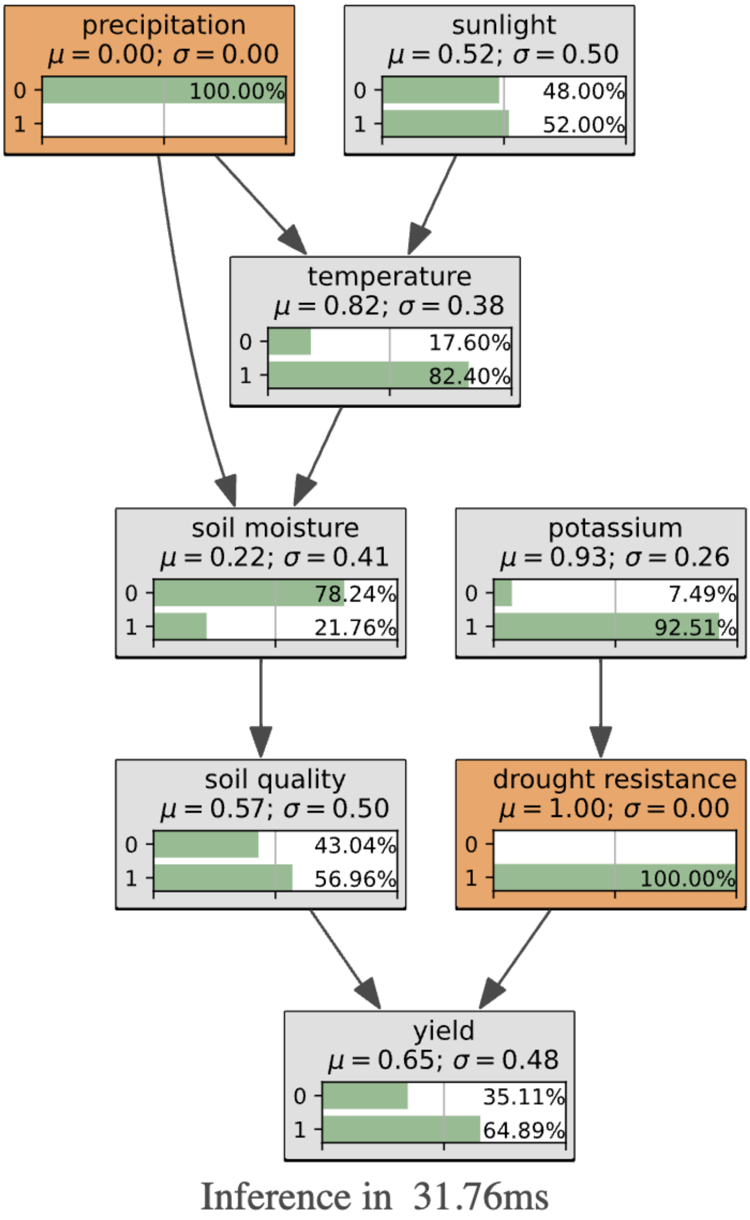

precipitation
$\mu = 0.00; \sigma = 0.00$
0
1
100.00%
sunlight
$\mu = 0.52; \sigma = 0.50$
0
1
48.00%
52.00%
temperature
$\mu = 0.82; \sigma = 0.38$
0
1
17.60%
82.40%
soil moisture
$\mu = 0.22; \sigma = 0.41$
0
1
78.24%
21.76%
potassium
$\mu = 0.93; \sigma = 0.26$
0
1
7.49%
92.51%
soil quality
$\mu = 0.57; \sigma = 0.50$
0
1
43.04%
56.96%
drought resistance
$\mu = 1.00; \sigma = 0.00$
0
1
100.00%
yield
$\mu = 0.65; \sigma = 0.48$
0
1
35.11%
64.89%
Inference in 31.76ms

Figure 9: Inference when low precipitation and high drought-resistance are observed

0.6. That means there is a 60% chance of good soil quality when there is low precipitation and low temperatures.

These hypothetical observations validate that the Bayesian network is able to make inferences about events represented by unobserved random variables based on evidence or observed data.

4.2 Decision Networks

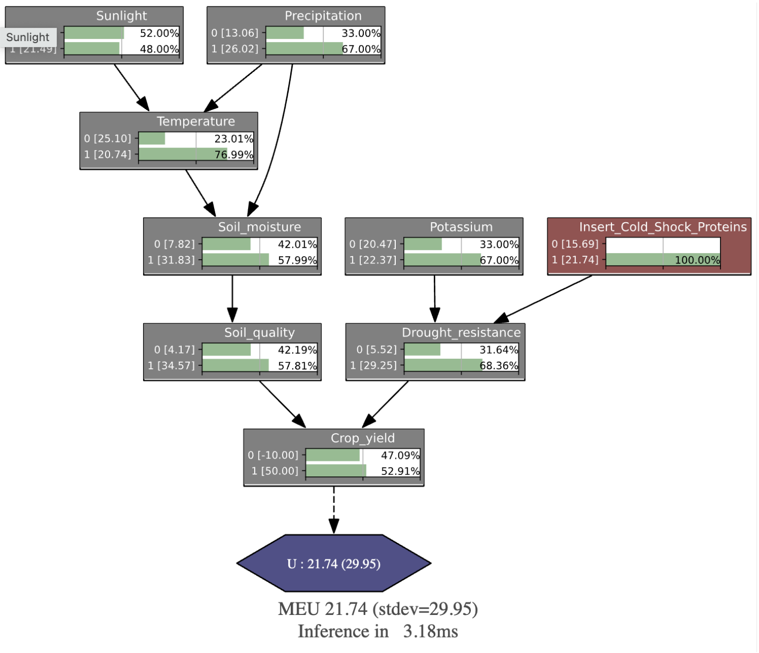


Figure 10: Influence Diagram

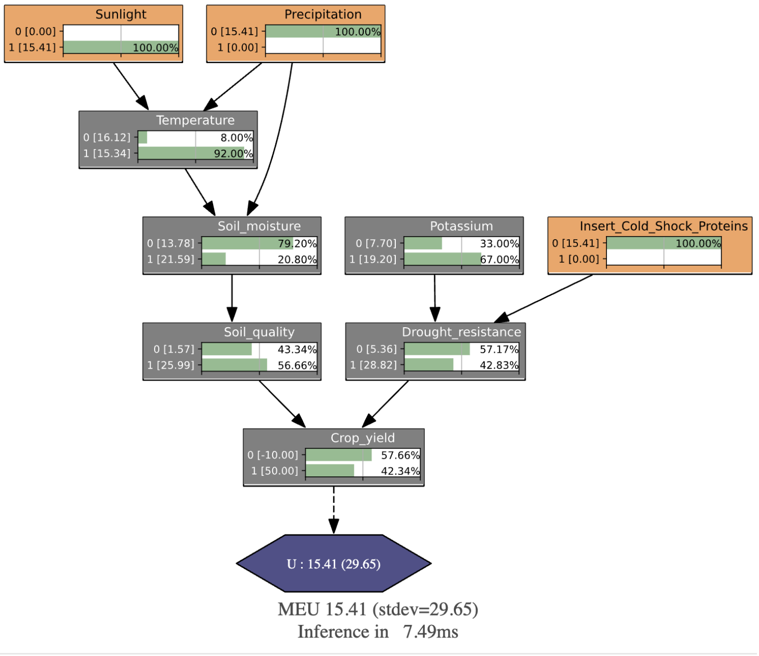


Figure 11: Decision Network with low observed precipitation, high observed sunlight, and no genetic modification

Decision Networks amalgamate Bayesian networks and decision theory with the objective of quantifying the effects of making certain decisions given that the events that influence those decisions can be modelled as random variables. In this example, the decision taken will be whether to bolster the wheat crop's drought resistance through the introduction of the SeCspA gene. It would be incorrect to assume that genetic modification is always the correct option as genetic modification of crops is a risky endeavour that can have negative

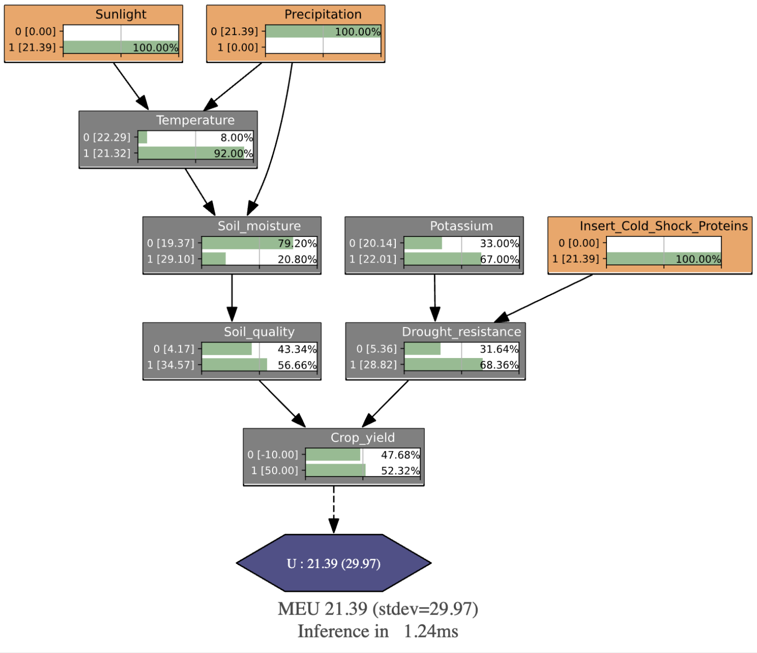


Figure 12: Decision Network with low observed precipitation, high observed sunlight, and added genetic modification

consequences for surrounding biodiversity and plant life. Therefore, decisions about genetic modification should only be made when there is a reasonable gain in utility to outweigh the risks. In our decision network we encoded the utility of a good yield as 50 and the utility of a bad yield as -10.

We declared the decision network with the observations of low precipitation and high sunlight to obtain a utility score of 15.41 when no genetic modification was used(Figure 11). This utility score rose to 21.39 when genetic modification was introduced(Figure 12). This represents a 38.8% increase in utility gained by making the decision to introduce genetic modification to the wheat crop.

4.3 Use Cases

This section will explore three potential use cases of this network and its ability to make inferences and validate decisions.

4.3.1 Use Case 1

The Bayesian network can help a wheat farmer decide whether to use seeds genetically modified with SeCpsA. The model will provide a recommendation based on the soil conditions and weather forecast for the season.

4.3.2 Use Case 2

The Bayesian network can be used by a regulatory agency that is deciding whether to approve the genetically modified seed for commercial use. The model can be used to show the potential benefits and risks of the seed, particularly with regards to it's effect on drought resistance and overall yield.

4.3.3 Use Case 3

The Bayesian network can be used for decision support by farmers in the Free State by allowing them to validate the effects that their actions will have on the predicted yield of wheat that season. The model can easily be adapted so that these actions can go beyond the introduction of genetically modified wheat and can include decisions about irrigation (based on sunlight, precipitation and temperature) and changing of fertilisers (based on soil quality).

5 Conclusion

Bayesian networks are powerful tools that allow us to make inferences about the probability distributions of unobserved variables based on random variables that we can observe. They can be extended to form Decision Networks that allow us to measure the utility of taking certain actions based on the probability distributions of events upon which those actions are conditioned. In this example, a Bayesian network can be used to predict the likelihood having a successful wheat yield in the Free State under stressful drought conditions where we observe low precipitation and high sunlight. Decision Networks can be used to measure the gain in utility of introducing genetic modification to the wheat crop in order to increase its ability to withstand the harsh conditions brought on by drought. Bayesian Networks proved to be appropriate in this use case because they were very sensitive to different observations at different locations on the directed acyclic graph. These networks provide an opportunity for statistical modelling in the sphere of agriculture to be used in order to increase food security and economic growth in the sector.

References

[1] M. ME, M. F. Nciizah AD, and W. IIC, “Tillage, crop rotation and crop residue management effects on nutrient availability in a sweet sorghum-based cropping system in marginal soils of south africa”, Agronomy **10**, `https://doi.org/10.3390/agronomy10060776` (2020).

[2] Q. A. Panhwar, A. Ali, U. A. Naher, and M. Y. Memon, “Fertilizer management strategies for enhancing nutrient use efficiency and sustainable wheat production”, Woodhead Publishing Series in Food Science, Technology and Nutrition, Organic Farming **34**, 17–39 (2019).

[3] J. V. Biljon, D. Fouche, and A. Botha, “Threshold values and sufficiency levels for potassium in maize producing sandy soils of south africa”, South African Journal of Plant and Soil **25**, 65–70 (2008).

[4] T.-F. Yu, Z.-S. Xu, J.-K. Guo, Y.-X. Wang, B. Abernathy, J.-D. Fu, X. Chen, Y.-B. Zhou, M. Chen, X.-G. Ye, and Y.-Z. Ma, “Improved drought tolerance in wheat plants overexpressing a synthetic bacterial cold shock protein gene secspa”, Sci Rep **7**, `https://doi.org/10.1038/srep44050` (2017).

[5] *Meteoblue*, (2023) `https://www.meteoblue.com/en/weather/historyclimate/climatemodelled/bloemfontein_south-africa_1018725`.

[6] M. Rui and W. Guiling, “Impact of sea surface temperature and soil moisture on summer precipitation in the united states based on observational data”, Journal of Hydrometeorology **12**, 1086–1099 (2011).