Utrecht University
Computing Science

Master thesis

Reuse of Bayesian Networks

A Case-study in Classical and African Swine Fever

Petra Geels

supervised by
Prof. dr. ir. L.C. van der Gaag
Dr. S. Renooij

External advisor:
Prof. dr. ir. G van Schaik
Faculty of Veterinary Medicine

March 19, 2018

Abstract

Developing a Bayesian Network has a high workload, also for domain experts, when not enough data is available to learn the model. We aim to reduce this workload by reusing an existing Bayesian Network when developing a new network. We study this by developing an initial model for African Swine Fever (ASF) by reusing the already existing Classical Swine Fever (CSF) model. African Swine Fever is a highly contagious disease, which is currently present in Poland and the Czech Republic. The risk of contamination in the Netherlands is substantial, and especially because no vaccine is available, a quick diagnosis is essential. Therefore, we developed a Bayesian Network to support early detection of the disease without having to wait for lab results.

The existing model for CSF consists of five phases, each representing a part of the body affected. These phases are used as a base, on which to build the reused model. The initial structure of the ASF model is determined, using only literature, very limited expert interviews and data of inoculation studies. When learning the parameters of the model, the probabilities of the CSF model where reused where possible. The remaining conditional probability tables are determined by using a variant of the EM algorithm.

The resulting network displays how a good initial model can be made in significant less time compared to developing a new one.

Contents

1 Introduction

Bayesian Networks are a powerful tool for combining expert knowledge with knowledge from data into a decision model. A Bayesian Network is a probabilistic data structure depicted as a directed acyclic graph. The graph consists of variables represented by the nodes and their interdependence represented by the arcs. For each node there is a probability table, representing the prior probabilities of the variable. Constructing such a Bayesian Network model demands a lot of effort and time, from both developers and domain experts. When an initial structure of the model is made, conditional probability tables (CPTs) have to be determined, the model has to be tested and results should be discussed with domain experts. This will be repeated until a proper model is found, causing an intensive iterative process for both developers and domain experts. Bayesian Networks are frequently developed to support medical and veterinary doctors with diagnoses, combining doctor's knowledge with prior experiences. Until now, for each application, a whole new model is developed. However, diseases can show a few or even a lot of the same clinical symptoms. This suggests parts of a network might be reusable for different diseases. Hence, we suggest reuse of software can be of help for faster development of Bayesian Networks. Besides reducing workload and thus cost, reuse of software also improves the reliability of the software (Kang and Frakes [16]) by accumulated checks (Lim [17]) and more time for details. We aim that reuse in the case of a Bayesian Network will give a better first model, and therewith the iterative development process will be substantially shortened.

We will explore the above concept for two swine fevers, Classical and African Swine Fever. In animal healthcare, an also relatively new research field is syndromic surveillance of diseases. Syndromic surveillance, or disease surveillance, amounts to monitoring the health of animals by on-going collection, validation and interpretation of data (Veldhuis [28]). The intent is to early detect and control diseases or disease outbreaks before diagnoses are confirmed (J. Henning [15]). In the Netherlands, disease monitoring is done by GD Animal Health (GD). The national pig health monitor consists of several surveillance components. Pig veterinarians can phone to a help desk for expert advise on pig health problems that they encounter. Furthermore, the veterinarians have to enter data of their finding during monthly obligatory farm visits in an Online Monitor programme. We will construct a decision model for ASF as support of syndromic surveillance by valuable interpretation of this monthly health data. This model will give a probability of a pig having ASF. With this probability, the veterinarian can decide whether to warn the authorities, who will take the appropriate actions.

Such a model is already been made for Classical Swine Fever (CSF) and found successful (van der Gaag et al. [27]). The CSF and ASF viruses show nearly the same clinical picture, so we believe a lot of the CSF model can be reused. The CSF model is built up from five distinguishable parts, each describing a part of the pig's body affected. We attempt to use the five parts as a base when reusing the model. In general, preservation of such a structure in a Bayesian Network can make the model reusable for other diseases of animals or even human diseases.

African Swine Fever is a highly contagious disease with a mortality rate up to 100% (for Animal Health [10]). The first clinical signs are aspecific and pigs can even die without showing signs. Currently the disease is not present in the Netherlands but it has already been detected in Poland and recently OIE World organisation for Animal Health announced a wild boar infected with ASF was found in the Czech Republic [23]. ASF can be spread via direct contact between infected and healthy animals, via fomites (cloths, vehicles, pork, etc.) and via ticks carrying the virus [9]. Since many people from Poland are visiting our country, there is an increased risk of infecting wild boars, by bringing infected meat. Today's methods for elimination of the disease rely on rapid diagnosis, movement restrictions, hygiene protocols, quarantine and

culling (de Carvalho Ferreira [9], for Animal Health [10]). As the pig industry is substantial in the Netherlands, occurrence of the disease will have an enormous (economical) impact. ASF is logically a notifiable disease and rapid diagnosis with e.g. syndromic surveillance is essential.

As ASF is spreading in Europe already, quick development of an appropriate model is essential. Therefore, in this thesis we explore if we can shorten the development time of a Bayesian Network by reusing an existing network, applied for Classical and African Swine Fever.

To accomplish this, we first define the similarity of the diseases using literature and a few expert interviews. With the five parts of the CSF model as a base, we will determine the network's structure with the obtained knowledge. We will refine the structure of the model by analysis of data from inoculation studies. When the structure is determined, we define which probabilities of the CSF model are reusable and learn remaining CPTs with an Expectation-Maximisation (EM) based algorithm.

The remainder of this paper will be organised as follows. In section 2 we describe Classical Swine Fever by the five phases that are distinguished for CSF. We will illustrate the surveillance model, based on these phases. In section 3, we give an overview of today's knowledge about (clinical signs of) African Swine Fever. In section 4, we define a first rough model for ASF, by first determining general reusability of the CSF model. Thereafter, we establish what is reusable for ASF specific and the order of the phases. In section 5, we process data of inoculation studies, to conclude the structure of the ASF model in section 6. In section 7, we will explain the algorithm used for parameter learning in section 7 and how we applied it in our case. The results and discussion are listed in section 8.

2 The Bayesian Network for Classical Swine Fever

We start off with a description of the existing Classical Swine Fever (CSF) model by first explaining the course of a CSF infection and thereafter explaining the Bayesian Network developed for CSF by van der Gaag et al. [27].

The CSF network is designed by van der Gaag, in collaboration with an experimental CSF expert and a senior epidemiologist from the Central Veterinary Institute of the Netherlands van der Gaag et al. [27]. In-depth interviews were held with Dutch swine veterinarians, of which some were operating during a CSF epidemic in the Netherlands. First a network of 42 variables was generated, which was evaluated briefly. This evaluation revealed the network gave many false CSF warnings. The reviews however displayed that a veterinarian easily can tell most of the cases are not CSF. While diagnosing, veterinarians used information about the combination of the clinical signs that were not covered by the model. When a pig showed certain signs without other signs first, they could rule out CSF for this case. With this knowledge the course of CSF consisting of five phases, as described below, was determined. Besides including this expert knowledge in the model, other variables that did not contribute anything were removed, resulting in a final model of 32 variables [27] which is shown in figure 1.

2.1 Classical Swine Fever

A Classical Swine Fever infection typically progresses through five phases (van der Gaag et al. [27], van der Gaag [26]). The first reaction of the pig's body after entrance of the virus will be an immune reaction. This reaction will show as high fever and general malaise, manifested by signs such as a loss of appetite and lethargy. Following the immune reaction, the intestinal tract is affected, with an inflammation of the mucous membranes as a result. The pigs will have abnormal faeces, mostly aspecific diarrhoea after initial dry faeces, as a consequence of the high fever. In the third phase, the virus also affects the respiratory tract: coughing, nasal discharge, conjunctivitis and breathing difficulties are observed with the diseased pigs in this stage. In the fourth phase of the infection, the virus enters the blood stream. Leaking blood vessels will cause cyanosis and pin-point skin haemorrhages, mostly seen at the ears and in the groins respectively. In the last phase of CSF, the central nervous system is assaulted, whereby ataxia is developed. Due to accumulating failure of body systems, the pig will die ([27], [26]).

2.2 The CSF network

The final network for CSF is based on the five phases as illustrated above. The phases are modelled by hidden variables, i.e. a variable that can not be observed, along with the evidence they give for CSF[27]. Each phase is described by a single hidden variable. The hidden variable is not an observable sign, but it holds information about combinations of the observable variables. For each phase i the hidden variable ϕ_i is called CSF Phase i. Since we know that phase i will most likely develop into phase $i+1$ (for $i = 1, ..., 4$), the phase variables are connected in chronological order. Therefore, the CPTs capture the probability that phase i occurs given the presence or absence of phase $i-1$, $P(\phi_i|\phi_{i-1})$ and $P(\phi_i|\neg\phi_{i-1})$ respectively. As the disease processes chronologically through the phases, the latter is put to zero. The other probabilities are assessed by an expert, since predisposing factors have to be taken into account.

In the lower part of the model (together with *Trembling* and *Stillborn piglets* in the upper part) are 14 observable variables, which represent all the clinical signs. Each sign is connected directly or indirectly to the corresponding phase they appear in. For some phases, one or two underlying causes in the body are also added as hidden variable. For example for phase 1,

`Body Temperature` and `Malaise` are added as cause of for example `Fever` and `Lethargy`, but are not observable themselves.

In addition to these variables, also seven so called "explaining away" variables (Wellman and Henrion [29]) are added to the network; these variables reside in the upper part of the model. An explaining away variable is another possible cause of an observed sign, which can explain why the sign is observed without the disease being present. As an example in the network, when a pig has diarrhoea but not likely CSF, the diarrhoea might be caused by a change of food, so `Feed` is added as explaining away variable to the model.

The performance of the CSF model is determined in terms of the *sensitivity* and *specificity*. The sensitivity of the model is the percentage for which the model actually finds that a pig has CSF out of all cases that the pigs have CSF. The specificity is the percentage of all not diseased pigs, for which the model indeed returns a negative diagnosis for CSF. We will clarify how to determine this with a Bayesian Network in section 8.1. For establishing the sensitivity, experimental data is used from three different countries. The sensitivity is determined for each day, by filling in the clinical signs shown by a pig so far. Adding up the number of pigs found positive for CSF thus far, gives the cumulative sensitivity up to that day. For specificity, data of pigs without CSF was collected by veterinarians in the Netherlands. The enhanced network described above was found to have a specificity up to 99% [27], and a cumulative sensitivity of 30% [25].

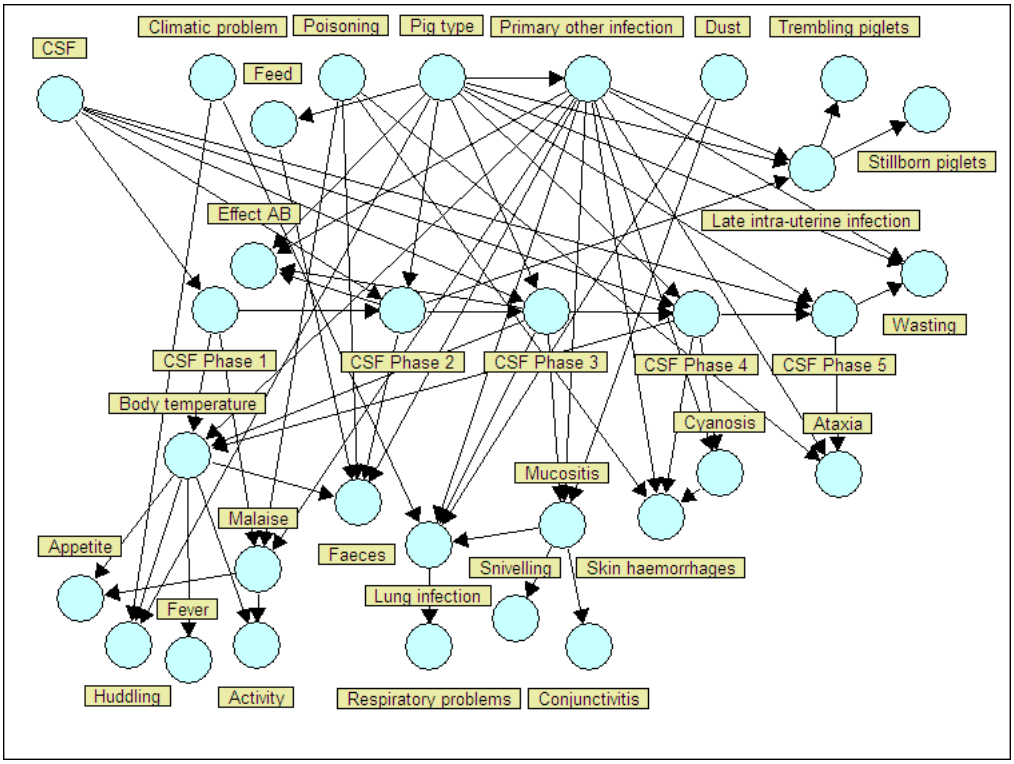

Figure 1: Bayesian Network for early detection of CSF

3 African Swine Fever

In this section, to get a proper picture of the disease, we will explain African Swine Fever in detail.

African Swine Fever is caused by the African Swine Fever virus (ASFV), which is a DNA virus [10]. ASFV can be transmitted in various ways: via direct contact between pigs, via fomites (cloths, vehicles, pork, etc.) contaminated by (excretions of) infected pigs, and via ticks as carrier of ASF [9]. The virus enters the body via the tonsils or a mucous membrane in the throat, from there it enters the blood stream via lymph nodes near the mouth and therewith, the virus spreads through viremia (i.e. the virus is present in the blood) [11]. During the inoculation studies of Guinat et al. [12], both inoculated or infected via direct or indirect contact pigs are found viraemic 1-3 days before the virus was detected in nasal or oral excretion, and just before/at the same time as the onset of clinical signs. This substantiates the virus enters the bloodstream very quickly, independent of the infection route.

A variety of different strains of the virus are known, each with a different virulence, i.e. a different capacity of infecting the host [5], and an incubation period from 4 to 19 days [9, 10]. In recent studies, an average incubation period of 4.4 and 6.15 was found for European virus strains ([12, 20] respectively). Highly virulent strains will cause peracute or acute disease, which have both a mortality rate up to 100% [10]. But also a sub-acute and chronic form, caused by less virulent strains, exist [11, 9]. These are not reported in Africa but they are in Europe and the Caribbean [2]. Below we will discuss the disease by summarising the clinical signs of each form of ASF.

3.1 ASF described by clinical signs

With the peracute form of ASF pigs may die without showing any clinical signs. Only a high fever (41-42°C), increased respiratory rate and hyperaemia (redness) of the skin can be seen in infected pigs when consistently observed. When these signs are observed, pigs will die within 1-3 days [9].

With the acute form, high fever is often seen as the first sign [2, 13]. The fever is manifested by signs as loss of appetite or anorexia, depression and huddling together [10, 2]; these are the most consistent signs of ASF [20, 12]. After the initial phase, other signs are commonly seen [2, 9]. One is redness of the skin, especially at the abdomen and extremities of the pig's body [10, 2]. It is described as flushing of the skin [2, 11], erythema [24] or hyperemia (i.e. increase of blood flow [7]) [20, 12]. Besides flushing, also cyanosis, a blue-ish purple discoloration [8], of the skin is frequently observed [9, 10]. The pigs also show respiratory problems: an increased respiratory rate is typical [10, 9]. Beside, conjunctivitis, an inflammation of the outer membrane of the eye [11], and sometimes thick whitish [2] discharges of the eyes [10] and nose [9] may be seen. The gait of the pigs is affected too, they show incoordination [9, 2, 10] with sometimes hind legs appearing weak [2]. The last frequently occurring clinical signs are vomiting, abortion [11] and bleeding from the rectum [20]. Less regular signs are, constipation or diarrhoea (sometimes bloody); bloody foam at the nostrils [2]; bleeding from nose and rectum [2] and abdominal pain [2]. In the recent study of Olesen et al. [20], also convulsions are reported. The last stage of the disease is characterised by coma and death [9]. 1-2 Days before death [10], anorexia, listlessness [10] or lethargy [11], cyanosis and incoordination can be seen [10] as the last stage of the disease [11]. Death will occur within 6-13 days [10].

For the sub-acute form, the most important signs are a fluctuating fever, depression and loss of appetite [2, 9]. Furthermore, swelling of the joints, dyspnoea (shortness of breath)[11] and heart failure will occur [2, 9]. Some pigs will show some less severe signs also seen with the acute form, and some pigs give an alert impression [9]. With this form they can be sick for

5-30 days and will die, possibly due to heart failure [2], in 15-45 days [10]. The mortality rate is lower than for the (per)acute form, and is extremely variable from 30 to 70% [10].

Pigs with the chronic form of ASF show some unspecific [11] and extremely variable signs [9] and can be sick for many months [9, 10]. The most commonly seen signs are loss of weight [11] or emaciation [9], stunting of growth, respiratory signs, skin ulcers, fever peaks [9, 11] and arthritis [9, 11] or swollen joints [2]. For the chronic disease, the mortality rate is low [10], less than 30% [11].

Variables CSF	**Variables ASF**
CSF	**ASF**
No	*No*
Yes	*Yes*
Climatic problem	**Climatic problem**
No	*No*
Yes	*Yes*
Dust	**Dust**
Normal	*Normal*
Abnormal	*Abnormal*
Feed	**Feed**
Unbalanced	Unbalanced
Balanced	Balanced
Pig type	**Pig type**
Suckling piglet	
Weaned piglet	*Suckling piglet*
Finishing pig	*Weaned piglet*
Sow	*Finishing pig*
Boar	
Poisoning	**Poisoning**
No	*No*
Yes	*Yes*
Primary other infection	**Primary other infection**
None	*None*
Respiratory	*Respiratory*
Respiratory+intestinal	*Respiratory+intestinal*
Snivelling	**Snivelling**
No	*No*
Yes	*Yes*
Stillborn piglets	-
No	
Yes	
n.a.	
Trembling piglets	-
No	
Yes	
n.a.	
Wasting	**Wasting**
No	*No*
Yes	*Yes*
Effect AB	-
No	
Yes	

Table 1: All variables with values of the upper part of the network, for the CSF and ASF model

4 The initial ASF model

Classical and African Swine Fever are often stated to be indistinguishable by clinical signs only[10, 2], which is promising for reusing the CSF model to construct a BN for ASF. To study the reusability of the CSF model, we will compare the two diseases by clinical signs. The five phases described for CSF in section 2, each represent a part of the pig's body affected. Clinical signs are connected to specific phases, as these signs show when that particular part of the body is affected by the virus. As many sings are corresponding for both diseases, we will maintain this global structure in the model. We will only rename the phases to the corresponding parts of the body affected; so *CSF Phase 1* will become *Immune reaction* etc. Per part, we will determine which signs should be reused, deleted or added in the case of ASF. Furthermore, we try to find whether an ordering in which the parts are affected, as with CSF, exists for ASF. For all reused variables we can copy or easily adapt the CPTs, so not too many new probabilities have to be estimated (by a domain expert).

Using this approach, the general reusability of the model will increase, since for each virus, we can check whether parts of the body are affected, with what kind of signs this is expressed, and if there is a pattern in the course of the disease.

In this section, we will only use literature and a few interviews to determine an initial model for ASF. First, we will show some results for the general reusability of the CSF model. After that, we will decide which variables will be reused, added or deleted, on the basis of the body parts.

4.1 General reusability

Van der Gaag (personal communication, June 2017) declared some parts of the CSF model were not as well designed as others. First of all, the part of the network about the late intra-uterine infection of sows has been considerably less validated than the other parts, and therefore cannot be reused just like that. Also, the validation of the effect on boars is poor. Lastly, the effect of antibiotics (AB) on a pig is ambiguous to use as explaining away variable. When the pig has low resistance caused by CSF infection, bacteria will also attack the pig more easily and so an antibiotic treatment can seem to have effect, even if the pig is (also) affected by the virus. Therefore, the late intra-uterine part and the variable `effect AB` will be left out of the initial ASF model and we will develop the model only for suckling piglets, weaned piglets and finishing pigs.

Besides the above, a mistake was found in the CSF model. The vertex representing the presence of dust in a pen, is now connected with lung infection. However, dust in a pen will not cause a lung infection by itself, but can cause coughing or related respiratory problems. So dust should be connected to respiratory problems instead of the lung infection vertex, which is what we will apply for the ASF model.

4.2 Reuse of the five phases

4.2.1 Immune reaction

The first part of the CSF model we will discuss is the immune reaction. Variables that belong to this phase are CSF phase 1, Body temperature, Appetite, Huddling, Fever, Activity and Malaise. All values of the variables can be found in table 2. High fever is frequently named to be the first sign of a pig with ASF. Loss of appetite, depression and huddling together are also signs for ASF (see section 3). Apart from these, lethargy is often described as depression, but from a veterinary perspective lethargy and depression are the same (van Schaik, personal

Phase	Variables CSF	Variables ASF
Immune	**Body temperature**	**Body temperature**
	Normal	*Normal*
	Increased	*Increased*
	Appetite	**Appetite**
	Normal	*Normal*
	Decreased	*Decreased*
	Huddling	**Huddling**
	No	*No*
	Yes	*Yes*
	Fever	**Fever**
	No	*No*
	Yes	*Yes*
	Activity	**Activity**
	Normal	*Normal*
	Lethargic	*Lethargic*
	Malaise	**Malaise**
	No	*No*
	Yes	*Yes*
	-	**Erythema**
		No
		Yes

Table 2: All variables with values belonging to the immune reaction, for the CSF and ASF model

communication, June 2017). Besides this, there is one new variable that possibly has to be added here. For ASF, flushing of the skin is regularly described. This could be part of the immune reaction and thus be different from cyanosis, which is bleeding under the skin and caused by the virus attacking the circulatory system. In conclusion, all the variables in the immune part appear to be reusable and we add a variable for flushing of the skin, named as `Erythema`, with values *No* and *Yes*. The model describing the immune reaction for ASF is shown in figure 2 and the corresponding values in table 2.

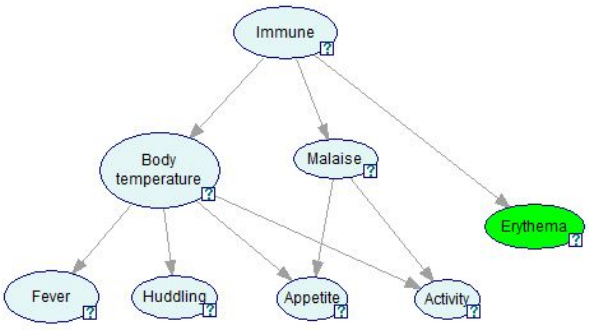

Figure 2: Model of the immune reaction for ASF.

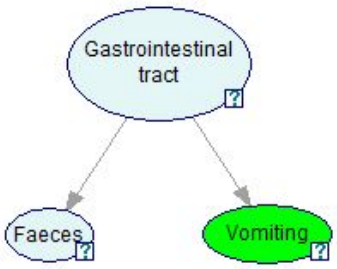

Figure 3: Model of the gastrointestinal tract for ASF.

Phase	**Variable CSF**	**Variable ASF**
Gastrointestinal tract	**Faeces**	**Faeces**
	Normal *Dry* *Aspecific diarrhoea* *Marked diarrhoea*	*Normal* *Dry* *Aspecific diarrhoea* *Marked diarrhoea* *Bloody diarrhoea*
	Late intra-uterine inf.	**-**
	No *Yes* *n.a.*	
	-	**Vomiting**
		No *Yes*

Table 3: All variables with values belonging to the gastrointestinal tract, for the CSF and ASF model

4.2.2 Gastrointestinal tract

Secondly, we look at Phase two of CSF, the virus' attack of the gastrointestinal tract. The model contains only `Faeces` as a variable for this part. As stated in section 3, with an ASF infection, vomiting, sometimes (bloody) diarrhoea and abdominal pain is observed. Hence, vomiting will be included in the ASF model, with values *No* and *Yes*. Since diarrhoea only appears sometimes, we will keep it in the model but with different probabilities. Also, the values of faeces will change a little. Bloody diarrhoea will be added since it is mentioned for ASF where it is not for CSF. Aspecific diarrhoea is diarrhoea which clearly belongs to another disease than CSF and ASF, so will be reused. Abdominal pain is not observable by veterinarians (van Schaik, personal communication, June 2017), so will not be added. This part of the ASF model is shown in figure 3 and values in table 3.

4.2.3 Respiratory tract

The third phase of CSF is the affection of the respiratory tract. Variables included are respiratory problems, snivelling, mucositis and lung infection. The clinical signs for the respiratory tract seem quite similar from literature for both swine fevers. Only the discharges of nose and eye might be bloody with ASF. As these signs are less regular, we will determine with the inoculation study data, if this should be added as value. de Carvalho Ferreira (personal

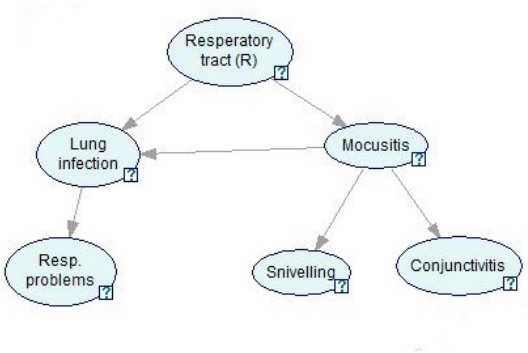

Figure 4: Model of the respiratory tract for ASF.

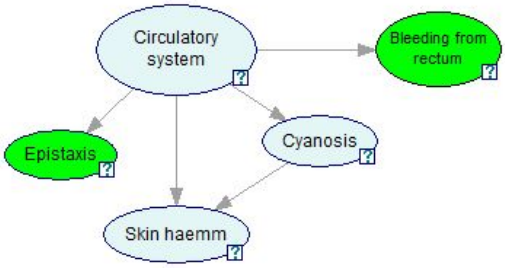

Figure 5: Model of the circulatory system for ASF.

communication, July 26, 2017) stated that the respiratory problems where not that special or different for both diseases. So for now, we assume the respiratory problems can be reused as-is. This part of the model and values per variable can be found in figure 4 and table 4 respectively.

4.2.4 Circulatory system

Next, we look at the signs of the virus' attack at the circulatory system. In the CSF model, this part consists of the variables skin haemorrhages and cyanosis. Both signs are seen with ASF as well, and hence should be reused. Beside these two, bleeding from nose and rectum are signs for ASF and should be appended, named *Epistaxis* and *Bleeding from rectum* respectively, both with values *No* and *Yes*. This part of the initial model can be seen in figure 5 with the variables explained in table 5.

4.2.5 Nervous system

In the final stage of CSF, the nervous system is affected. The variables included for CSF are ataxia (a collective noun for bad coordination of muscle movement) and wasting. In the CSF model, ataxia has the values *No*, *Aspecific* and *CSF-specific* (which is sitting like a dog). For ASF the signs stated are: incoordination and sometimes hind legs appearing weak. This suggests we should change the value names of the values of this variable to *No*, *Incoordination* and *Dog sitting*. As the pig will die eventually, wasting should be included for ASF too. This part of the model and the corresponding variables are shown in figure 6 and table 6.

Phase	Variable CSF	Variable ASF
Respiratory tract	**Mucositis**	**Mucositis**
	No	*No*
	Yes	*Yes*
	Lung infection	**Lung infection**
	No	*No*
	Yes	*Yes*
	Respiratory problems	**Respiratory problems**
	No	*No*
	Yes	*Yes*
	Snivelling	**Snivelling**
	No	*No*
	Yes	*Yes*
	Conjunctivitis	**Conjunctivitis**
	No	*No*
	Yes	*Yes*
	Effect AB	-
	No	
	Yes	

Table 4: All variables with values belonging to the respiratory tract, for the CSF and ASF model

Phase	Variable CSF	Variable ASF
Circulatory system	**Cyanosis**	**Cyanosis**
	No	*No*
	Yes	*Yes*
	Skin haemorrhages	**Skin haemorrhages**
	No	*No*
	Yes	*Yes*

Table 5: All variables with values belonging to the circulatory system, for the CSF and ASF model

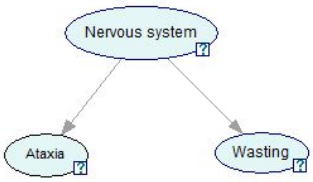

Figure 6: Model of the nervous system for ASF.

Phase	Variable CSF	Variable ASF
Nervous system	**Ataxia**	**Ataxia**
	No	*No*
	aspecific	*Incoordination*
	csf-specific	*Dogsitting*
	Wasting	**Wasting**
	No	*No*
	Yes	*Yes*

Table 6: All variables with values belonging to the nervous system, for the CSF and ASF model

4.3 Sequence of the phases in ASF from literature and interviews

Now that we have defined the parts of the initial ASF model, we have to determine the order in which they appear. All the variables of the upper part of the CSF network that are not discussed above, will be reused in the ASF network as well, together with their probabilities, because they are connected with one or more variables we want to reuse in the middle and/or lower part of the network.

Since both diseases are caused by a virus, it is likely that an ASF infection starts with an immune reaction of the body too. Also, as said above, high fever is frequently mentioned as the first sign, which substantiates this assumption. De Carvalho Ferreira (personal communication, July 26, 2017) who has done a PhD study of ASF, declared that the signs of damage of the circulation system, nervous system, gastrointestinal and respiratory tract all seem to appear at once. So we can not assume the phases will be sorted as with CSF. In section 3.1, anorexia, listlessness, cyanosis and incoordination are named to be seen in the last stage of the disease. Nevertheless, these signs can be present during the whole infection, so we can not say they appear later than others. Accordingly, further research is needed to determine how the phases should be connected. In the next section we will complete the structure of the ASF model by using data of inoculation studies, which are experimental setups where pigs are observed very closely after inoculating (some of) the pigs with ASF, to complete the structure of the ASF model.

Samples	Inoculated pigs Latent period*	Within-pen contact pigs Time to onset of infectiousness†	Between-pen contact pigs
Blood ‡	4.8 (± 1.3)	10.3 (± 1.6)	13.9 (± 3.0)
Blood §	3.6 (± 1.0)	10.4 (± 1.4)	13.1 (± 3.0)
Oral swab ‡	5.4 (± 1.3)	8.5 (± 1.5)	9.2 (± 1.5)
Nasal swab ‡	5.4 (± 1.4)	7.6 (± 2.6)	11.3 (± 0.5)
Rectal swab ‡	4.9 (± 1.4)	9.3 (± 2.9)	11.0 (± 1.6)
	Incubation period*	Time to onset of clinical signs†	
Clinical score >3	4.4 (± 1.0)	9.9 (± 1.6)	12.7 (± 2.0)

*Average number of days post inoculation (± standard deviation).
†Average number of days post exposure (± standard deviation).
‡Results by quantitative real-time polymerase chain reaction.
§Results by virus titration.

Figure 7: Results of the average duration of latent and incubation period for the inoculated pigs and time to onset of infectiousness and clinical signs for the contact pigs

5 Inoculation studies

To gain more insight in the ordering of the signs appearing, we will look at inoculation studies. As mentioned above, inoculation studies are an experimental setting with a group of pigs, where some of the pigs are inoculated with the disease. All the pigs are observed daily and the clinical signs they show are listed. With this daily data of clinical signs, we can check whether some parts of the body seem to be affected before others. To achieve this, we will make plots where we count each day how many pigs are showing certain signs in each part of the body.

Below, we will look at two different inoculation studies. From the first one, from Guinat et al. [12], we have the clinical data per day. Of the second inoculation study of Olesen et al. [20], we do not have the data, but the results are described in great detail in their paper.

5.1 Inoculation study of Guinat et al. [12]

Guinat et al. [12] recently performed an inoculation study with the Georgia strain of ASF. Stated that this strain is not significantly changing [12], this is a pretty good study to use as basis to further define the structure of the network. The goal of this study was to get more detailed information about the clinical signs, viremia and virus excretion of the ASF virus, via different infection routes. For this, they inoculated a fraction of 40 pigs with the virus strain in a controlled environment. The pigs were divided in four rooms. In each room, only some of the pigs were inoculated with ASF. The other pigs are within-pen or between-pen contact pigs, susceptible to the ASF virus. In room B and C, four pigs were separated by a 80 cm high partition, being the between-pen contact pigs, the remaining pigs were within-pen contact pigs.

Each pig was examined daily to obtain the rectal temperature, clinical signs and blood, oral, nasal and rectal fluid samples. When gathering the clinical signs a form is used with ten clinical signs listed, each with a score of severity. The form can be found in appendix A. Fever is determined as a temperature higher than $40°C$ for two consecutive days. From the fluid samples the location of the virus in the body at a given day post inoculation is determined, which is shown in figure 7. For welfare reasons, the pigs were euthanized as soon as they showed a rectal temperature of $40°C$ or higher for three consecutive days, or showing three different clinical signs.

To gain more insight in the moment of onset of the clinical signs corresponding to the five phases, we used this data to make plots that show which signs appear on which day post

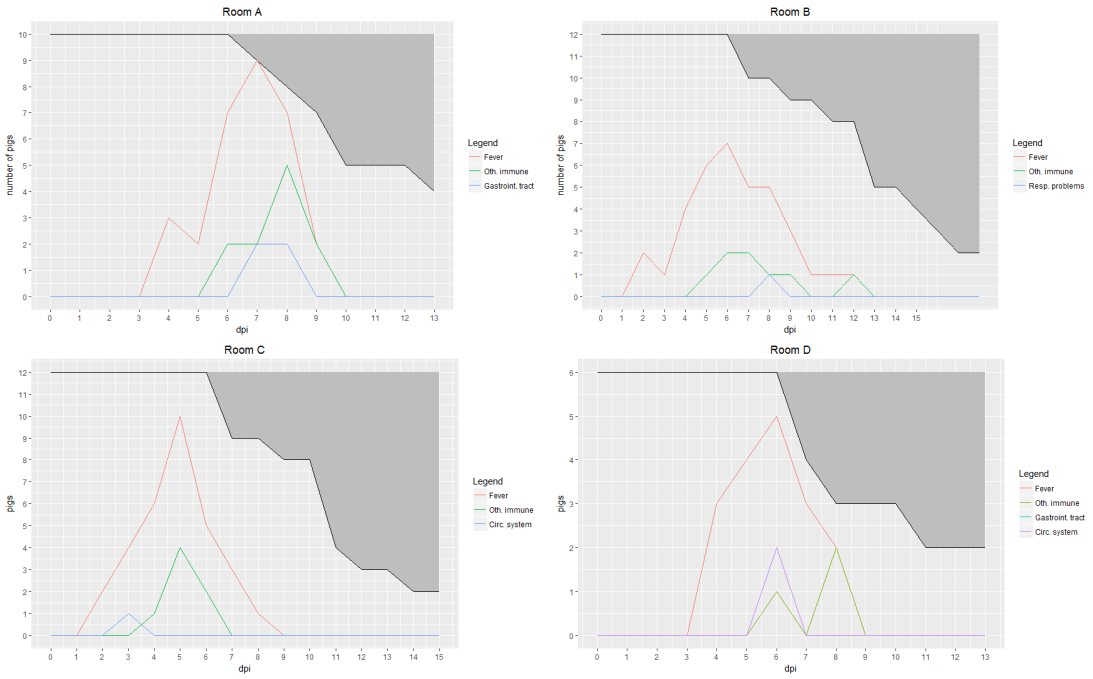

Figure 8: Clinical signs observed in room A-D.

infection (dpi). Note here that the definition of dpi by Guinat et al. [12], actually means days post inoculation of the inoculated pigs. The contact pigs were of course not infected immediately. Since we are interested in the onset of clinical signs after infection, we had to shift the data. As Guinat et al. [12] found that clinical signs appeared at day 4.4 ± 1.0 for inoculated pigs, and 9.9 ± 1.6 for withing-pen contact pigs (see figure 7), we assume the pigs were infected with ASF 5.5 days after inoculation. So we shift the data for the within-pen contact pigs so that dpi 0 is at day 4. We rounded down, as rounding up resulted in the within-pigs showing signs and dying earlier than the inoculated pigs, which is not plausible. For between-pen contact pigs we did the same, we defined dpi 0 on day 7.

For each day, we counted how many pigs have signs corresponding to a phase. Defining these counts, we are indifferent how high the signs are scored (as in the assessment form is shown). So for example, when a pig shows lethargy with score one or higher, we count this pig for showing an immune reaction. The plots are shown in figure 8 for each rooms separately and figure 9 shows the result for all rooms together. Note that we look at fever apart from the other immune reactions, to verify the statement in section 3 that fever is often seen as the first sign for the acute form of ASF. To gain the right conclusion for our model, we had to slightly change the definition of fever. As we make a model for single day observations, we have to define fever as a temperature at one day. Since equipment and age of the pigs can influence the definition of fever, we will define fever as one day above $40°C$ conform with the temperature specified by Guinat et al. [12].

As we can see in the plots explained above, fever is always the first sign seen. Other immune reactions appear later than fever, and sometimes earlier and sometimes later than the other phases. Of the remaining phases circulatory system sometimes appear earlier but mostly together with the gastrointestinal tract, respiratory tract and nervous system. The last four phases are hardly observed, which can be due to the fact that the pigs were euthanized before natural death.

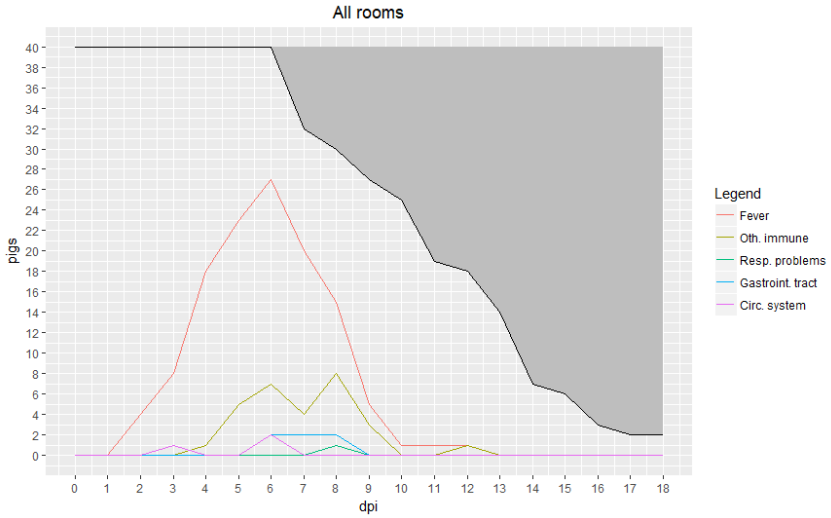

Figure 9: Clinical signs observed in all rooms together.

5.2 Inoculation study of Olesen et al. [20]

Even more recently Olesen et al. [20] preformed a inoculation study with an virus isolate from Poland. We do not have data of this inoculation study, but the results are fairly detailed. As this is the closest country to the Netherlands where ASF is reported, it seems useful to check whether there are some big differences with the study in the previous section. In this study there were again four pens used, now creating four groups: 1) inoculated pigs, 2) within-pen contact pigs, 3) between-pen contact pigs 4) air-contact pigs. Olesen et al. [20] also obtained the viremia, clinical signs and transmission parameters (e.g. time until onset of infectiousness or onset of clinical signs) of the virus for within-pen, between-pen but also for air-contact pigs. The pigs of group 1 were intranasally inoculated on post infection day 0 (PID as by their definition). As above, the contacts pigs were not directly infected with the ASF virus, so we shifted the data for these pigs by the same argumentation as described above, using the transmission parameters stated in figure 11. So PID means indeed post infection day from now on. Clinical signs were again collected with a form, which can be found in appendix B.

We summarised the results given in paragraph 3.2.1 of Olesen et al. [20] to get a same plot as in the previous section. We translated the text as much as possible into numbers, obtaining the number of pigs showing clinical signs of a certain phase, per day. The results are shown in figure 10. Here, fever and other immune reactions appear together as first signs seen, followed by circulatory problems and attack on the nervous system. After that problems with the gastrointestinal tract seem to appear. Respiratory problems were not reported.

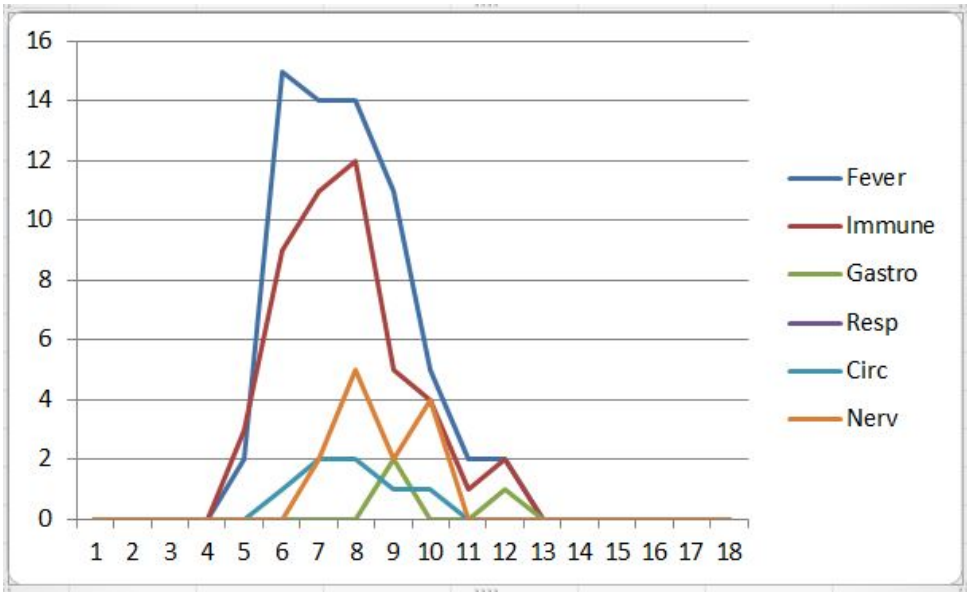

Figure 10: Clinical signs observed in all pens together by Olesen et al. [20]

	Inoculated pigs (1*a*)	Within-pen contact pigs (2*a*)	Between-pen contact pigs (3*a*)	Air-contact pigs (4*a*)
	Incubation period (days)	Time until onset of clinical signs (days) §		
	n = 4	*n* = 4	*n* = 4	*n* = 4
CS > 3	6.0 ± 2.7	8.3 ± 1.5	10.0 ± 0.8	12.8 ± 1.0

Samples	Latent period (days) ξ	Time until onset of infectiousness (days) §		
Sera +	5.5 ± 2.7	9.0 ± 0.0[a]	9.5 ± 1.0	12.0 ± 1.7[a]
Sera #	5.5 ± 2.7	9.0 ± 0.0[a]	9.5 ± 1.0	12.0 ± 1.7[a]
EDTA #	4.5 ± 3.1	7.7 ± 1.2[a]	8.5 ± 1.0	9.7 ± 1.2[a]
Nasal swab #	5.3 ± 2.5	8.3 ± 1.5	11.0 ± 0.0	11.8 ± 1.5
Oral swab #	7.3 ± 3.2[b]	9.8 ± 2.5	11[c]	12.0 ± 1.7[c]
Rectal swab #	7.0 ± 3.6[d]	8.3 ± 1.2[d]	10.5 ± 1.0	12.0 ± 1.7[d]

ξ Average number of days post inoculation (± standard deviation), § average number of days post exposure (± standard deviation), # results by qPCR, + results by virus isolation.

[a] Pigs 7 (group 2*a*) and 16 (group 4*a*) did not become viremic, and these pigs are not included in the calculations.

[b] Viral DNA was not detected in oral swabs obtained from pig 1 (group 1*a*), and the pig is not included in the calculation.

[c] Cq values below 35 were not consistently detected in oral swabs from pigs 10, 11, 12 (group 3*a*) and 14 (group 4*a*), and these pigs are not included in the calculations. A standard deviation cannot be calculated for group 3*a*.

[d] Viral DNA was not detected in rectal swabs from pigs 1 (group 1*a*), 7 (group 2*a*) and 16 (group 4*a*), and these pigs are not included in the calculations.

	Inoculated pigs (1*b*)	Within-pen contact pigs (2*b*)	Between-pen contact pigs (3*b*)	Air-contact pigs (4*b*)
	Incubation period (days)	Time until onset of clinical signs (days) §		
	n = 4	*n* = 4	*n* = 4	*n* = 6
CS > 3	6.3 ± 1.3	12.3 ± 1.0	13.3 ± 1.5[a]	15.0 ± 1.7[a]

Samples	Latent period (days) ξ	Time until onset of infectiousness (days) §		
Sera +	4.8 ± 1.5	12.0 ± 2.3	13.8 ± 2.1	15.4 ± 2.6[b]
Sera #	5.3 ± 1.9	12.0 ± 2.3	13.8 ± 2.1	15.4 ± 2.6[b]
EDTA #	4.5 ± 1.7	10.3 ± 1.0	12.8 ± 2.8	14.8 ± 3.0[b]
Nasal swab #	5.3 ± 1.3	12.5 ± 1.3	13.3 ± 2.1[c]	15.0 ± 1.7[c]
Oral swab #	6.7 ± 1.5[d]	13.3 ± 1.0	15.0 ± 1.2	14.7 ± 0.6[d]
Rectal swab #	5.8 ± 1.5	10.3 ± 0.5	14.8 ± 1.0	14.3 ± 0.6[e]

ξ Average number of days post inoculation (± standard deviation), § average number of days post exposure (± standard deviation), # results by qPCR, + results by virus isolation.

[a] Pigs 32 (group 3*b*), 38, 39 and 40 (group 4*b*) did not reach a clinical score above 3, and these pigs are not included in the calculation.

[b] Pig 38 (group 4*b*) did not become viremic and this pig is not included in the calculation.

[c] Cq values below 35 were not consistently detected in nasal swabs from pigs 32 (group 3*b*), 38, 39 and 40 (group 4*b*) and these pigs are not included in the calculations.

[d] Cq values below 35 were not consistently detected in oral swabs from pigs 26 (group 1*b*), 38, 39 and 40 (group 4*b*) and these pigs are not included in the calculations.

[e] Cq values below 35 were not consistently detected in rectal swabs from pigs 38, 39 and 40 (group 4*b*) and these pigs were not included in the calculation.

Figure 11: Results of transmission study from Olesen et al. [20]

6 The structure

6.1 Global structure

The result figures above suggest that fever indeed can be seen as one of the first signs of ASF. And therewith, the immune reaction should be the first in ordering of the phases in the model.

In both studies, circulatory problems seem to appear a little earlier than the remaining phases. Note that these figures are based on very few data points, so we must be careful. Beside looking at clinical signs, we can also look at the blood samples taken. In figure 7 of the first inoculation studies, we compare `Blood` ǂwith the onset of clinical signs (i.e. clinical score > 3). We see, for within-pen and between-pen contact pigs respectively, the virus is present in the blood from day 10.3($\pm$1.6) and 13.9($\pm$3.0). And the onset of the clinical signs is from 9.9($\pm$1.6) and 12.7($\pm$2.0). So the virus is present in the blood 0.4 and 1.2 day(s) after onset of clinical signs. For the second study, we compare `Sera` # with $CS > 3$ in figure 11. We see in the left table that for group 3a (between-pen contact pigs), 4a (air-contact pigs) and 2b (within-pen contact pigs) the virus is 0.5, 0.8, and 0.3 days earlier detected in blood than clinical signs start to appear, respectively. For the remaining groups the virus is in the blood less than 0.7 day after clinical onset.

Combining the above knowledge with the clinical signs data, we draw the following conclusion. As we do not see all clinical signs within one day, the above numbers substantiates that the attack of the circulatory system will appear earlier than or together with the other phases, except `Immune` what is mostly the first sign seen. This is an important contrast with CSF, were it appears at the very end. As the above reasoning is not conclusive to determine further ordering of the other four phases, we will put the last four phases together, after the immune reaction. Further testing is needed to verify this assumption.

According to the above conclusions, we will define the overall structure by connecting `Immune` to all other parts of the body and connect all these parts to `Wasting`. That way all phases not immune, appear together after the immune reaction and probably just before the pigs dies. The global structure is shown in figure 12.

6.2 Clinical signs

In section 4.2.3 we were left to determine whether to change the nodes corresponding to the respiratory problems. As only coughing is observed once in the studies ofGuinat et al. [12] and no other respiratory problems, there is no cause or enough evidence to change the snivelling node. The total structure of the ASF model is shown in figure 13.

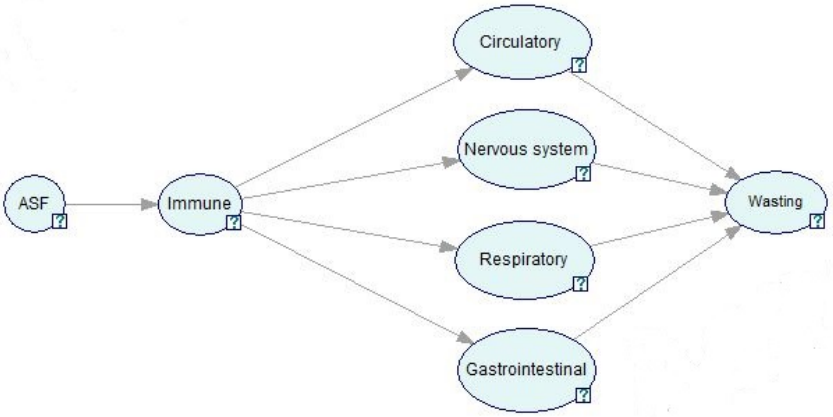

Figure 12: Global structure of the ASF network

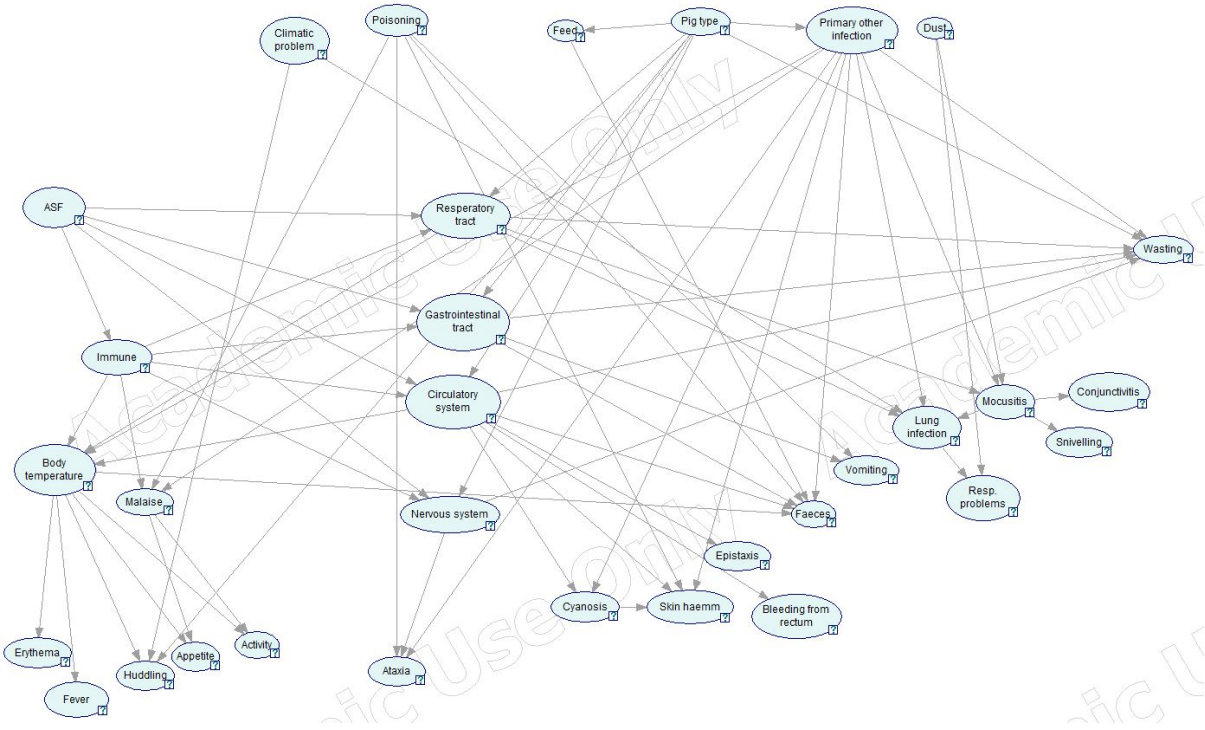

Figure 13: Total structure of the ASF network

7 Learning the parameters

Now the structure is determined. We have to obtain the second part of the BN, the conditional probability table (CPT) of each variable. We can do this by asking experts to estimate these probabilities, but we aim to keep the workload low for experts. So we will first find initial values by learning the parameters from data, whereafter we can calibrate.

As the CSF model is tested pretty well, we will reuse probabilities from the network as well for the ASF model. To learn the remaining CPTs, we will use the data of the inoculation studies of Guinat et al. [12], from section 5.1. This time, per pig we noted if a clinical sign appeared, irregardless when the sign showed in time. When a sign is not described on the assessment form (Appendix A) and not observed, it is set to Not Available (NA). This means the data has missing values, and for some variables, no data points are available. To be able to learn from this data anyway, we will use a parameter learning algorithm applicable for missing values and hidden variables.

In this section, we will first describe which probabilities of the CSF network can be reused. Then, we give a short overview of the learning algorithms used nowadays. Thereafter, we will explain two variants of the Expectation-Maximisation algorithm and we conclude by showing how we applied these algorithms in our network.

7.1 Reuse of parameters

Before learning parameters, we define which of the probabilities can be reused. Prior probabilities for the explaining away variables, such as `Dust` and `Poisoning`, can surely be reused, assuming the situation in the pig herds of the Netherlands have not changed that much. Also, the probabilities of the hidden variables excluding the five representing the body parts, can be reused. As an example, the probability that `Fever` is established, given that there is actually a raised body temperature, will be reusable, as this captures the human and/or machine error rate. The parameters we want to learn are the five hidden variables representing the body parts: `Immune` (Immune reaction), `Gastro` (Gastrointestinal tract), `Resp` (Respiratory tract), `Circ` (Circulatory system) and `Nerv` (Nervous system), and the new parameters or parameters with new values: `Erythema`, `Faeces` and `Vomiting`. Note that `Bleeding from rectum` and `Epistaxis` are both also new variables but there is no observed data available. Therefore, we can not learn these parameters and so they should be estimated by an expert.

When reusing parameters, it can appear that the number of parents of a variable in the ASF does not equal the number parents it had in the CSF model. When a variable has more parents in the ASF model, we will simply duplicate the probabilities, i.e. we use $P(X \mid Y, Z_j) = P(X \mid Y), \forall j$. In the case of less parents, we marginalise over the missing parent(s): $P(X \mid Y, Z) = \sum_i P(X \mid Y, Z, W_i)P(W_i)$ [18].

The prior probability of CSF appearing in the Netherlands is 0.0000019, for ASF, this is a little higher, 0.0019, as it is present in neighbouring countries.

7.2 Parameter learning algorithms

Learning from data is gaining popularity. Learning from data is done in machine learning, artificial intelligence and also in data mining [4] to find patterns or relations. For learning Bayesian Networks, you can learn both the structure and/or the parameters. We will only look at parameter learning as we will only learn the CPTs from data. In our case, we have to learn discrete distributions, with two or more values per variable, for both hidden and observable variables.

The general goal of learning parameters, or obtaining the CPTs, can be described as follows. We set the probabilities of parameters $\boldsymbol{w}$ to some start value. We have some data D and we want to find $\boldsymbol{w}$ which represent the data best. Finding $\boldsymbol{w}$ with missing values is done by an iterative process, taking little steps, to a global maximum.

A common way to do this, is via a Maximum likelihood estimate (MLE) method, where the likelihood of the BN is maximised. This method however has problems when there are zero observations of certain values in the data. In this method, the parameters are estimated by the number of occurrences N_{ijk} that node X_i equals value k, divided by the total number of times node i's parents are equal to the j^{th} assignment. But if the j^{th} assignment of the parents never occurs in the data, you divide by zero ([30]). In our case, the data has missing values or even hidden variables, so this problem can possibly arise. A different approach is the Maximum a posteriori (MAP), where you start with a prior (Dirichlet) distribution [30, 19], and maximise from this.

Different algorithms for both methods are known. Most well known are Expectation-Maximisation (EM) and Gradient Descent, less commonly also Markov Chain Monte Carlo based algorithms (such as Gibbs Sampling [22, 21]) are mentioned. Furthermore, a number of variations or extensions of these algorithms are explored, e.g. using auxiliary networks [18] or include expert knowledge by implementing constraints for parameters [30].

Gradient Descent is a fairly convenient method. In this algorithm, a nonlinear function describing the CPTs is determined and this function will be maximised. Each iteration, the gradient vector of partial derivative with respect to the CPTs is computed, whereafter a step is taken into this direction. When the gradient is zero, you have reached a local maximum for the CPTs [22]. Gradient descent can take a lot of iterations to find this maximum and so become computational expensive [6].

Very powerful for very complicated high dimensional probability distributions are the MCMC methods [1]. These methods are based on sampling from the CPTs [3], making a chain of possible models. These methods are not often mentioned for learning Bayesian Networks with hidden variables.

The EM algorithm is a popular [21], mathematically grounded and frequently mentioned algorithm when learning parameters including hidden variables [21, 3]. The EM algorithm iterates between computing expected values with respect to the data, and computing new probabilities with this expected values. EM has the downside that it sensitive to starting values of the CPTs and it can be computational expensive when you have a large network [21].

In this thesis, we will use a MAP version of EM as described by Neapolitan [19], to learn parameters for which all data points are available, but possibly not for all their parents. For learning the hidden variables' CPTs, we use another version of EM, using Bayes inference in the maximisation step [14].

7.3 Expectation-Maximisation algorithm by Neapolitan [19]

The general idea of EM by Neapolitan [19] is that data is extended where data points are missing. These data points should be missing at random, otherwise the missing variables are dependent on the values of the other variables. The probability that each extended row occurs is set to a starting probability. With these occurrences, a expectation with respect to the data is computed. With this expectation, a new probability is determined and this will be the new occurrence of corresponding rows. By iterating this, a (local) maximum of the prior probabilities is reached. Below we will first give some notation, after which we show the algorithm applied on an example network in section 7.3.2.

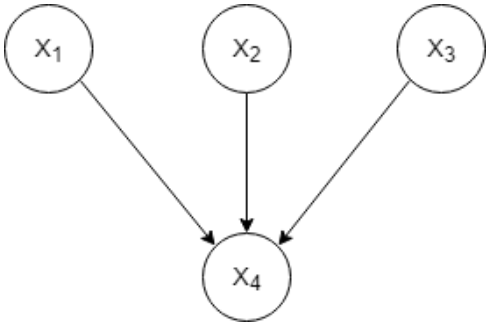

Figure 14: Example network for notation

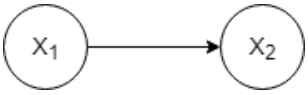

Figure 15: Bayesian Network of run-through example

7.3.1 Notation

A Bayesian Network is represented by a graph $G(V, E)$, where V are the nodes (or vertices) and E the arcs (or edges). Each node $X_i \in V$ has r_i values and c_{X_i} denotes the conjunction of its values. We define $\mathtt{PA}_i$ as the parents of X_i and $\mathtt{pa}_{ij}$ is the j^{th} instantiation of the values of the parents of X_i, whereby there is a defined ordering for all possible assignments. In total, q_i possible instantiations of X_i's parents exist. For example, we look at X_4 in the example BN in figure 14, where all variables are binary. We have $\mathtt{PA}_4 = \{X_1, X_2, X_3\}$, for which there are $q_4 = 8$ possible assignments. When assigning the ordering: $\{X_1, X_2, X_3\} = \{(1,1,1), (2,1,1), (1,2,1), (2,2,1), (1,1,2), (2,1,2), (1,2,2), (2,2,2)\}$, then $\mathtt{pa}_{43} = (1,2,1)$.

For each node, we denote its probability as

$$f_{ijk} = P(X_i = k \mid \mathtt{pa}_{ij})$$

where $k \in c_{X_i}$ stands for the current value of the node. All probabilities of the network are denoted by f, so in the example from above we have

$$f = \{f_{111}, f_{112}, f_{211}, f_{212}, f_{311}, f_{312}, f_{411}, \ldots, f_{481}, f_{412}, \ldots, f_{482}\}.$$

When learning the parameters from data, we will start with a initial probability. This probability is described by a_{ijk}'s, where

$$P(X_i = k \mid \mathtt{pa}_{ij}) = \frac{a_{ijk}}{\sum_k a_{ijk}}.$$

In words, we believe that value k appears a_{ijk} times of a total of $\sum_k a_{ijk}$ cases. For intance, in the same example network, when we have no prior knowledge, we are indifferent which value of X_1 appears more often. Then we set $a_{111} = a_{112} = 1$ and we get an initial probability f_{111} of

$$P(X_1 = 1) = \frac{a_{111}}{a_{111} + a_{112}} = \frac{1}{2}.$$

7.3.2 EM algorithm detailed

Now that we have the notation, we will explain EM as defined by Neapolitan [19] by a run through example. We will work with the Bayesian network represented in figure 15 existing of

Case	X_1	X_2
1	1	1
2	1	?
3	1	1
4	1	2
5	2	?

Case	X_1	X_2	Occurrences
1	1	1	1
2a	1	1	$\frac{1}{2}$
2b	1	2	$\frac{1}{2}$
3	1	1	1
4	1	2	1
5a	2	1	$\frac{1}{2}$
5b	2	2	$\frac{1}{2}$

Table 7: Data with missing values (left) and extended data (right) of run-through example EM [19].

two binary nodes X_1 and X_2 with an arc from X_1 to X_2. We have some data available shown in table 7(right). In this data, a few data points are missing for node X_2.

To learn with this data, we start by extending the data with all possible options of values of the node for which data is missing. In the case of our example, we duplicate the data row where X_2 is missing and fill in both possible values of X_2, as is shown in the right table of table 7. We add a column where we list the occurrences of each row. When the row does not have missing values, the occurrence is one. Where data is missing, we fill in the initial probability we believe this row will occur. We start with no knowledge at all, so we fill in the probabilities $P(X_2 = 1 \mid X_1 = 1) = \frac{1}{2}$ and $P(X_2 = 1 \mid X_1 = 2) = \frac{1}{2}$.

Now we can compute the expected value for each probability from this data as extended above. We call this expectation s_{ijk} and it is simply defined as counting the number of occurrences of node i being equal to k given the j^{th} instantiation of his parents. More formally,

$$s_{ijk} = E(X_i = k, \mathtt{pa}_{ij} | d, f) = \sum_{row \in d} P(X_i = k, \mathtt{pa}_{ij} | d, f). \tag{1}$$

where d is the data and the probability P per data row is the occurrence of this row. This is called the `expectation step` of the algorithm.

Next, we compute new probabilities f_{ijk} with this expectations, the `maximisation step`. This we do by simply adding up the expected values to our previous probabilities in the following way:

$$f_{ijk} = \frac{a_{ijk} + s_{ijk}}{\sum_{k \in c_{X_i}} a_{ijk} + \sum_{k \in c_{X_i}} s_{ijk}}.$$

In our example, we first compute the expected values for X_2 given $X_1 = 1$ ($j = 1$ when $X_1 = 1$ and 2 otherwise). We count

$$\begin{aligned} s_{211} &= \sum_{h=1}^{5} P(X_1^{(h)} = 1, X_2^{(h)} = 1 \mid d, f) \\ &= 1 + \frac{1}{2} + 1 + 0 + 0 = 2\frac{1}{2}, \end{aligned} \tag{2}$$

where (h) stands for the h^{th} row. The same way we get

$$s_{212} = 0 + \frac{1}{2} + 0 + 1 + 0 = 1\frac{1}{2}. \tag{3}$$

Now we can compute the new f_{211} of the maximazation step by:

$$\begin{aligned} f_{211} &= \frac{a_{211} + s_{211}}{\sum_{k \in c_{X_2}} (a_{21k} + s_{21k})} \\ &= \frac{2 + 2\frac{1}{2}}{2 + 2 + 2\frac{1}{2} + 1\frac{1}{2}} \\ &= \frac{7}{12}. \end{aligned} \tag{4}$$

This new probability for $f_{211} = \frac{7}{12}$ can now be filled in for row 2a in the right table of table 7. And $f_{212} = \frac{5}{12}$ as occurrence of row 2b.

The same procedure can be done for X_2 given $X_1 = 2$. After filling in all new probabilities, we can start over with computing the new expectations, then compute the probabilities again, etc. After a certain amount of iterations of the algorithm, a local maximum will be reached.

7.3.3 Defining a_{ijk}

Above, we started with simple values for a_{ijk}. But we should be careful choosing these values. Below we describe why we can not just choose 1 for each node, as this gives odd results, by explaining an example with coin tossing (Neapolitan [19]). We show that an equivalent sample size prevents this problem, and how we can still start with prior indifference as our knowledge.

Equivalent sample size Take the same example network as above in figure 15. Let both nodes represent a coin toss, with value 1 if it lands heads and value 2 if it lands tails. Assuming we have fair coins, our prior believe would be that each coin lands heads half of the time. As stated in section 7.3.1, we can take $a = b = 1$ for both nodes, getting $f_{111} = \frac{1}{1+1} = 0.5$ as our prior believe for X_1. And also:

$$\begin{aligned} P(X_2 = 1) &= P(X_2 = 1 \mid X_1 = 1)P(X_1 = 1) + P(X_2 = 1 \mid X_1 = 2)P(X_1 = 2) \\ &= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2} \end{aligned} \tag{5}$$

as our prior believe for X_2.

Now say we update the probabilities with the data from 8 tosses, shown in table 8. We get

$$\begin{aligned} P(X_2 = 1 \mid X_1 = 1) &= \frac{1 + s_{211}}{2 + s_{211} + s_{212}} \\ &= \frac{1 + 1}{2 + 1 + 2} \\ &= \frac{2}{5}. \end{aligned} \tag{6}$$

And in the same way we get

$$\begin{aligned} P(X_2 = 1 \mid X_1 = 2) &= \frac{1 + 3}{2 + 3 + 2} = \frac{4}{7}, \\ P(X_1 = 1) &= \frac{1 + 3}{2 + 3 + 5} = \frac{2}{5}, \text{ and} \\ P(X_1 = 2) &= \frac{3}{5}. \end{aligned} \tag{7}$$

With these new probabilities, the probability of $P(X_2 = 1)$ becomes:

$$\begin{aligned} P(X_2 = 1) &= P(X_2 = 1 \mid X_1 = 1)P(X_1 = 1) + P(X_2 = 1 \mid X_1 = 2)P(X_1 = 2) \\ &= \frac{2}{5} \cdot \frac{2}{5} + \frac{4}{7} \cdot \frac{3}{5} \\ &\approx 0.50286. \end{aligned} \tag{8}$$

But if we look at the data, we see the second coin landed head exactly half of the time of the experiment. So the outcome is not what we expect, as we will still believe X_2 is a fair coin after these 8 tosses.

What we did in this case, was defining a different amount of prior occurrences X_1 and X_2 and combining these different sample sizes while computing $P(X_2 = 1)$. We stated that X_1 equals 1 once out of two times. But for X_2, the initial probability is explained as the second coin lands heads once out of two times that the first coin landed head. In this case, we have defined two prior occurrences for X_1 and four for X_2 (two for each outcome of X_1). So, we do not believe $P(X_1 = 1) = 0.5$ as much as the probabilities of X_2 indicate we should.

To prevent this, we should define the same prior sample size for each node in the BN. In this example, we take for instance $a = b = 2$ for X_1 and $a = b = 1$ for X_2. This way, we also have four prior occurrences for X_1. The initial probability remains a half: $P(X_1 = 1) = \frac{2}{2+2} = \frac{1}{2}$ and the updated probability of X_1 becomes:

$$\begin{aligned} P(X_1 = 1) &= \frac{2+3}{4+3+5} \\ &= \frac{5}{12}. \end{aligned} \tag{9}$$

With this probability, we indeed get a probability of a half for X_2 after updating:

$$\begin{aligned} P(X_2 = 1) &= \frac{2}{5} \cdot \frac{5}{12} + \frac{4}{7} \cdot \frac{7}{12} \\ &= \frac{1}{2}. \end{aligned} \tag{10}$$

Formulating the above solution more formally, we should define a prior sample size N_{ij} for each node, such that the network has an equivalent sample size N [19] where:

$$N_{ij} = \sum_{k=1}^{r_i} a_{ijk} = P(pa_{ij}) \times N. \tag{11}$$

We do this by defining the a_{ijk} as

$$a_{ijk} = \frac{N}{r_i q_i}. \tag{12}$$

We will show why the network has equivalent sample size N when we define a_{ijk} as such. Recall that q_i is the number of instantiations of the parents of node i, and therefore is defined as

$$q_i = \prod_{pa \in \mathtt{PA}_i} r_{pa}. \tag{13}$$

Case	X_1	X_2
1	1	2
2	1	1
3	2	1
4	2	2
5	2	1
6	2	1
7	1	2
8	2	2

Table 8: Data on coin tossing.

With a_{ijk} as in (12), the initial probability of each node is given by:

$$\begin{aligned} P(X_i = k \mid pa_{ij}) &= \frac{a_{ijk}}{\sum_k a_{ijk}} \\ &= \frac{\frac{N}{r_i q_i}}{r_i \frac{N}{r_i q_i}} \\ &= \frac{N}{r_i q_i} \cdot \frac{q_i}{N} \\ &= \frac{1}{r_i}. \end{aligned} \tag{14}$$

With that, the probability of a certain instantiation of parents becomes the probabilities of each parent multiplied:

$$\begin{aligned} P(pa_{ij}) &= \prod_{pa \in PA_i} \frac{1}{r_{pa}} \\ &= \frac{1}{\prod_{pa \in PA_i} r_{pa}} && \text{Use formula (13)} \\ &= \frac{1}{q_i}. \end{aligned}$$

So now we know $P(pa_{ij}) = \frac{1}{q_i}$ for all i and j. Therefore,

$$\sum_{k=1}^{r_i} a_{ijk} = r_i \frac{N}{r_i q_i} = \frac{N}{q_i} = P(pa_{ij}) \times N, \tag{15}$$

and indeed (11) holds and thus the network has equivalent sample size N.

Prior indifference Bigger values of a_{ijk} represent a stronger believe in this starting value. So for prior indifference, we should take values not too large. Reasonable for describing prior indifference can thus be $a_{ijk} = \frac{1}{r_i}$ (where r are the number of values of i). Combining this with the equivalent sample size, we should take an equivalent sample size of $N = maxr$, the largest number of values of a variable appearing in the network. The rationale doing this is that for a node X_p with $maxr$ values, it is reasonable to believe we have seen each of the value once, and it is as small as possible.

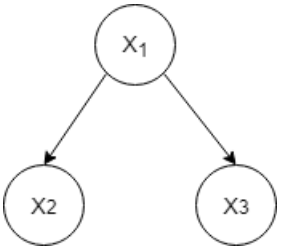

Figure 16: Example network adjusted EM

Case	X_1	X_2	X_3	$P(X_1 = 1 \mid d, f)$
1	NA	1	1	$\frac{a}{a+b}$
2	NA	1	1	$\frac{a}{a+b}$
2	NA	1	2	$\frac{a}{a+b}$
3	NA	1	1	$\frac{a}{a+b}$
4	NA	2	2	$\frac{a}{a+b}$

Table 9: Data with hidden variable

7.4 Adjusted EM for the hidden variables [14]

The EM algorithm described above works well if there only a few or none data points are missing for a certain variable. When a variable is however hidden, the above algorithm will not learn anything from the data. In this section, we will show why this happens by giving an example and thereafter explain an adjusted version of EM, which can learn with hidden variables.

Take the example BN shown in figure 16 and say we have no data points for X_1. We extend each row for X_1 and fill in the initial probability, as shown in table 9. To keep the formulas readable, we will use $a_{111} = a$ and $a_{112} = b$. We start with:

$$\begin{aligned} P(X_1 = 1) &= \frac{a}{a+b} \\ P(X_1 = 1) &= \frac{b}{a+b}. \end{aligned} \tag{16}$$

We follow the EM algorithm as above and start with computing the new expectation by counting the occurrences (`expectation step`):

$$\begin{aligned} s_{111} &= 5 \cdot \frac{a}{a+b} \\ s_{112} &= 5 \cdot \frac{b}{a+b}. \end{aligned} \tag{17}$$

Computing the new probabilities we get (`maximisation step`):

$$\begin{aligned} f_{111} &= \frac{a + 5\frac{a}{a+b}}{a+b+5\frac{a}{a+b}+5\frac{b}{a+b}} \\ f_{112} &= \frac{b + 5\frac{b}{a+b}}{a+b+5\frac{a}{a+b}+5\frac{b}{a+b}}. \end{aligned} \tag{18}$$

Now we can rewrite the upper formula as follows:

$$
\begin{aligned}
f_{111} &= \frac{a + \frac{5a}{a+b}}{a + b + \frac{5a}{a+b} + \frac{5b}{a+b}} && \text{get same denominators:} \\
&= \frac{\frac{a(a+b)}{a+b} + \frac{5a}{a+b}}{\frac{(a+b)(a+b)}{a+b} + \frac{5a}{a+b} + \frac{5b}{a+b}} && \text{merge fractions:} \\
&= \frac{\frac{a(a+b)+5a}{a+b}}{\frac{(a+b)^2+5(a+b)}{a+b}} && \text{rewrite:} \\
&= \frac{a(a+b)+5a}{\cancel{a+b}} \cdot \frac{\cancel{a+b}}{(a+b)^2+5(a+b)} && \text{simplify:} \\
&= \frac{a(a+b)+5a}{(a+b)^2+5(a+b)} && \text{rewrite:} \\
&= \frac{a(a+b+5)}{(a+b)(a+b+5)} && \text{which equals:} \\
&= \frac{a}{a+b}.
\end{aligned}
$$

As we see, the new probability equals the initial probability. So the algorithm will never learn a new probability for a hidden variable. The same holds of course for the lower formula of (18).

But when we look at the data, we see X_2 and X_3 are equal to one, three out of five times. Because of the structure of the network, this should also tell us the probability $P(X_1 = 1)$ should be about this $\frac{3}{5}^{th}$ (unless the probability of X_2 and X_3 given X_1 state otherwise). This is why we should also use information of the nodes below the hidden variable, by using Bayes inference.

And that is the idea of this adjusted algorithm. We again start with a initial probability, but now we fill this in as prior probabilities of the BN. Next, we compute the occurrence of each data row by filling in all known data points as evidence, and determine the posterior probability of the hidden variable with Bayes inference. Then, we perform the expectation and maximisation step similarly as above. Note that the expectations/occurrences are now displayed as probability that the data row occurs. We clarify this with an example.

We start with initial probabilities f for the network in figure 16:

$$
\begin{aligned}
f_{111} &= P(X_1 = 1) = 0.6 \\
f_{211} &= P(X_2 = 1|X_1 = 1) = 0.55 \\
f_{221} &= P(X_2 = 1|X_1 = 2) = 0.4 \\
f_{311} &= P(X_2 = 1|X_1 = 2) = 0.45 \\
f_{321} &= P(X_2 = 1|X_1 = 2) = 0.65.
\end{aligned} \tag{19}
$$

We use this as the prior probabilities of our network. Now for each row, we set the value of X_2 and X_3 as evidence and determine the probability $P(X_1)$ with Bayes inference. This probability is then the occurrence of the corresponding data row as shown in table 10.

As the expectation step, we compute the total occurrences as above determined. For instance, we compute

$$s_{111} = 3 * 0.588 + 0.764 + 0.639 = 3.167.$$

Case	X_1	X_2	X_3	$P(X_1 = 1 \mid d, f)$
1	NA	1	1	0.588
2	NA	1	1	0.588
2	NA	1	2	0.764
3	NA	1	1	0.588
4	NA	2	2	0.639

Table 10: Data with hidden variable

Then, in the maximisation step, we compute the new probability with

$$P(X_1 = 1) = \frac{E(X_1 = 1)}{E(X_1 = 1) + E(X_1 = 2)} = \frac{3.167}{5} = 0.6334.$$

Note that the prior probabilities now are taken into account in the inference step instead of the maximisation step. You compute these new probabilities for all nodes of the network and fill these new probabilities in as prior probabilities of the network. You can compute the occurrences $P(X_1)$ with inference again and repeat the above procedure until the (global) maximum is reached.

7.5 EM applied on our network

In the above examples, all the parameters of the network where learned from the data. As already mentioned in the beginning of this section, we want to reuse most of the probabilities and only learn `Erythema`, `Faeces`, `Vomiting`, `Immune`, `Gastro`, `Resp`, `Circ`, and `Nerv`. When a variable is *Not Available* in the data, but we reuse their probability, we will also extend the data for this variable. When determining the occurrence of each data row, all probabilities of the variables with *NA* in the data are multiplied. Only these probabilities will not change for reused variables.

`Erythema`, `Faeces`, and `Vomiting` are leaf nodes and no data is missing, so we apply the EM algorithm as described by Neapolitan [19] in section 7.3.2. Obviously, in this case we only need one iteration, counting the occurrence of each value of the node, as nothing will change in the data in the next iteration. However, we learn these variables simultaneously with the remaining, hidden variables, because some of these variables are parents of this clinical sign variables. As they are hidden, their probability will change every iteration and so the probabilities can differ per instatiation of the parents of the observed variables. Thus, also `Erythema`, `Vomiting`, and `Faeces` will be learned every iteration.

For the phase variables, we use the adjusted EM algorithm. Note that this algorithm is a MLE method and is defined for learning root nodes, so the problem of dividing by zero described in the beginning of section 7.2 will not occur. But the phase variables we want to learn in our network, are not root nodes. Therefore, we split the learning process in two parts. Using domain knowledge obtained from the inoculation studies in section 5, we know that the immune reaction appears first. So we will learn that part of the model first. This part is shown in figure 17. We learn with the data of the inoculation studies, and in this data the variables `ASF`, `Dust`, `Poisoning`, `Primary other infection`, `Feed`, and `Climatic problems` have fixed values. So learning with these parameters fixed, is the same as learning `Immune` as the root as above.

Second, we learn the probabilities of `Gastro`, `Resp`, `Circ`, and `Nerv`, and `Erythema`, `Faeces`, and `Vomiting` and set the learned probability of `Immune`. The result probability of `Immune` is $P(Immune = Yes|ASF = Yes) = 1$ and therefore we can fill in `Yes` for `Immune` in the data.

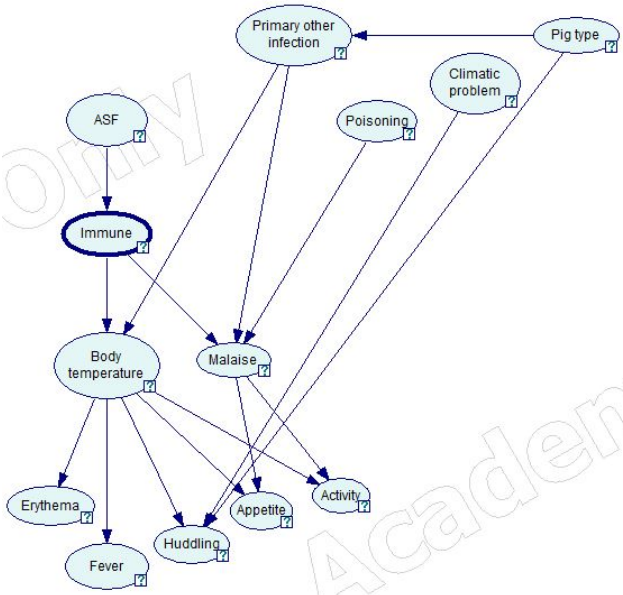

Figure 17: Only the immune reaction part of the ASF model

With `Immune` filled in, as above, all parents of the four hidden variables are now fixed, and the remaining four hidden variables become root nodes and can be learned with the adjusted EM algorithm. When learning these remaining variables, we use the whole Bayesian Network.

α	Sensitivity
0.19	10%
0.019	37.5%
0.01	45%
0.005	82.5%

Table 11: Sensitivity of the ASF model

8 Results and discussion

Now that we have determined the structure and prior probabilities of the network, we will look at the performance of the developed model. Thereafter, we will note some adjustments we had to make and we will conclude the project.

8.1 Results

To determine the performance of the network, *sensitivity* and *specificity* together are a good measurement. Recall that sensitivity establishes the percentage that the model actually evaluates a pig has ASF, out of the cases where pig are infected with ASF. Specificity, as counterpart, is the percentage of all not diseased pigs, of which the model indeed returns negative diagnosis.

The sensitivity will be defined as follows. A data row will be used as evidence, asking for the posterior probability $P(ASF = Yes)$. When this probability exceeds a certain threshold α, we claim the model returns positive for this pig having ASF. The sensitivity of the model is:

$$sensitivity = \frac{\text{ASF notices}}{\text{Total pigs in data with ASF}} \times 100 \text{ [25].}$$

Note that sensitivity alone is not saying much, but is only giving an idea. Specificity should be determined too, with data from pigs without ASF, but showing at least one sign of ASF. With the formula:

$$specificity = \frac{\text{Total pigs in data - number of ASF positives}}{\text{Total pigs in data}} \times 100 \text{ [25],}$$

it should be checked if the model is not just given ASF warnings when it appears to be another virus.

We conducted the sensitivity for a few alpha's. The prior probability for ASF was 0.0019, taking that as alpha resulted in 100% sensitivity which is not really plausible and probably will give a low specificity. So we tried some different alphas, the results are shown in table 11. Together with specificity, the most appropriate alpha should be determined. Due to lack of time, the specificity is not yet determined.

8.2 Data adjustments

Because EM explodes in run-time with every non observed variable, we made some adjustments to keep the run-time somewhat doable. On the basis of expert knowledge, we can argue some of the signs can be put to a value rather than Not Available.

First, the pigs are euthanised early because of welfare reasons, and therefore `Wasting` will probably never be seen in this experimental setting, whether it is checked or not. Besides, since we aim for an early detection model, `Wasting` is not the most important variable, as the pig is already almost dead at that point. Thus, we set `Wasting` to *No*.

Second, for `Ataxia` we decided the assign the value *No*, recall that `Ataxia` has the values *No*, *Incoordination* and *Dogsitting*. These values are not on the assessment form, but joint swelling with difficult walking is checked. As difficulty in walking is therewith checked, it is reasonable to assume `Ataxia` did not appear.

Finally, from domain knowledge, we can say there is no lung infection, as long as there are no respiratory problems according to van Schaik (personal communication, June 2017). This is substantiated by the prior probabilities of the the CSF network. The probability for no lung infection given no respiratory problems is 0.99. So in the cases no respiratory problems are observed, we will put lung infection to *No* as well, and NA otherwise. Note that now, the data points with NA are not missing at random, but as we do not learn this variable, this is not a problem.

8.3 Conclusion

The interpretation of the sensitivity without the specificity can not tell us much, as false warnings should not appear too often as well. Also, the percentage of the sensitivity is difficult to interpret, as the model is developed for a single pig. A veterinarian will always assess a herd instead of one pig. For example, three pigs with a probability higher than 30% can be enough to suspect ASF in a herd. Authorities will have to decide which combination of sensitivity and specificity must be used in practice.

Even without formal performance measures, we can evaluate the development process and the initial ASF network and answer the research question: does the reuse of an existing model shorten the development time of a Bayesian Network?

We have shown that it will. The structure of the model is fairly understandable and resembles the CSF network. Together with the good performance results of the CSF model, this means that the ASF model is a promising model. By an experienced researcher, this initial model can be developed in only months, where the design CSF model took years, of which a major part was building the initial model. Many in-depth interviews with experts were held to determine the first structure of the CSF model, where only a few are used for the ASF model. Testing this initial model will give already a good insight of the performance of the model, whereafter fine tuning already can be started.

8.4 Future research

Of course, in general, the model should be reviewed further and some (parts of) CPTs still have to be estimated by experts because they could not be learned with data. The structure on the new nodes should be firmly tested, as should the new values of variables. Testing the sensitivity should also be done per day, where it now is the worst disease picture possible per pig. But the model should perform well on every moment of the clinical picture the pig shows. Below, we will name some specialities to taken into account for further development of the ASF model.

First, the current data includes only cases for very limited values of some variables, and thus only parts of the CPTs could be learned. For example, Respiratory tract has parents Immune, ASF and Pig type. When learning, only the case (Immune = Yes, ASF = Yes, Pig type = Weaned) appears in this data. The other assignments hence can not be learned from the data and have to be completed.

Second, in the data adjustments, we did put Ataxia and Wasting to No. But these two signs together are the only children of the Nervous system node. So these assumptions can have an important impact on the CPT of this node. In further research this should be checked closely.

As last, Faeces got Circulatory system as new parent, because we added the value bloody diarrhoea which is probably caused by affection of the Circulatory system. We learned Faeces

with the data. But the fact that bloody diarrhoea is caused by the affection of the Circulatory system is not taken into account here. This should be checked with a domain expert.

Besides better testing of the performance and determining remaining probabilities, we also suggest to study if adding an extra phase would improve the model. In the first inoculation studies, we noted fever was appearing mostly before other immune signs. We would like to recommend researching if fever is a phase of ASF by itself.

Further reuse of Bayesian Networks should definitely be explored for other pig diseases and even for human diseases. This research showed that preservation of a global structure consisting of five parts of the body affected, makes a Bayesian Network easy to reuse.

Acknowledgements

I would like to thank Guinat et al. [12] for providing data of inoculation studies. Second, I would like to thank Helena Cardoso de Carvalho Ferreira for providing data and a good first idea of the African Swine Fever disease. I also want to thank Gerdien van Schaik as my external advisor from Veterinary Medicine for the pleasant and helpful meetings.

Finally, I really want to thank my supervisor Linda van der Gaag for the very pleasant and enjoyable collaboration on this project. She has given me a really good idea what conducting research is about and she has given me self-confidence during the project. I could not have imaged a better matching supervisor and project for my master thesis.

References

[1] Alex. Marcov chain monte carlo introduction. URL https://www.youtube.com/watch?v=3ZmW_7NXVvk.

[2] Daniel Beltran-Alcrudo, Marisa Arias, Carmina Gallardo, Scott Kramer, and Mary Penrith. *African swine fever: detection and diagnosis A manual for veterinarians.* 06 2017. ISBN 978-92-5-109752-6.

[3] W. Buntine. A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8(2):195–210, Apr 1996. ISSN 1041-4347. doi: 10.1109/69.494161.

[4] Vladimir Cherkassky and Filip Mulier. *Introduction*, pages 1–18. John Wiley Sons, Inc., 2006. ISBN 9780470140529. doi: 10.1002/9780470140529.ch1. URL http://dx.doi.org/10.1002/9780470140529.ch1.

[5] Wikipedia contributors. Virulence — wikipedia, the free encyclopedia, 2017. URL https://en.wikipedia.org/w/index.php?title=Virulence&oldid=797550565. [Online; accessed 23-February-2018].

[6] Wikipedia contributors. Gradient descent — wikipedia, the free encyclopedia, 2018. URL https://en.wikipedia.org/w/index.php?title=Gradient_descent&oldid=829476994. [Online; accessed 10-March-2018].

[7] Wikipedia contributors. Hyperaemia — wikipedia, the free encyclopedia, 2018. URL https://en.wikipedia.org/w/index.php?title=Hyperaemia&oldid=827722765. [Online; accessed 8-March-2018].

[8] Wikipedia contributors. Cyanosis — wikipedia, the free encyclopedia, 2018. URL https://en.wikipedia.org/w/index.php?title=Cyanosis&oldid=819502791. [Online; accessed 8-March-2018].

[9] Helena Cardoso de Carvalho Ferreira. *Towards an improbed understanding of African swine fever virus transmission.* 2013. ISBN 978-90-393-6005-7.

[10] World Organisation for Animal Health (OIE). Technical disease card: African swine fever. URL http://www.oie.int/fileadmin/Home/eng/Animal_Health_in_the_World/docs/pdf/Disease_cards/AFRICAN_SWINE_FEVER.pdf. [Online; Last updated: April 2013].

[11] Claudia Gabriel. Classical and african swine fever in domestic pigs and european wild boar. July 2012. URL http://nbn-resolving.de/urn:nbn:de:bvb:19-148216.

[12] Claire Guinat, Ana Reis, Christopher Netherton, Lynnette Goatley, Dirk Pfeiffer, and Linda Dixon. Dynamics of african swine fever virus shedding and excretion in domestic pigs infected by intramuscular inoculation and contact transmission. 45:93, 09 2014.

[13] Erin B. Howey, Vivian ODonnell, Helena C. de Carvalho Ferreira, Manuel V. Borca, and Jonathan Arzt. Pathogenesis of highly virulent african swine fever virus in domestic pigs exposed via intraoropharyngeal, intranasopharyngeal, and intramuscular inoculation, and by direct contact with infected pigs. *Virus Research*, 178(2):328 – 339, 2013. ISSN 0168-1702. doi: https://doi.org/10.1016/j.virusres.2013.09.024. URL http://www.sciencedirect.com/science/article/pii/S016817021300316X.

[14] Massachusetts institute of technology. Learning with hidden variables [pdf slides]. URL https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-825-techniques-in-artificial-intelligence-sma-5504-fall-2002/lecture-notes/Lecture18FinalPart1.pdf.

[15] Kelly J. Henning. Overview of syndromic surveillance what is syndromic surveillance? 53 Suppl:5–11, 10 2004.

[16] K. Kang and W. B. Frakes. Software reuse research: Status and future. *IEEE Transactions on Software Engineering*, 31:529–536, 07 2005. ISSN 0098-5589. doi: 10.1109/TSE.2005.85. URL doi.ieeecomputersociety.org/10.1109/TSE.2005.85.

[17] W. C. Lim. Effects of reuse on quality, productivity, and economics. *IEEE Software*, 11:23–30, 09 1994. ISSN 0740-7459. doi: 10.1109/52.311048. URL doi.ieeecomputersociety.org/10.1109/52.311048.

[18] Roger Luis, L. Enrique Sucar, and Eduardo F. Morales. Inductive transfer for learning bayesian networks. *Machine Learning*, 79(1):227–255, May 2010. ISSN 1573-0565. doi: 10.1007/s10994-009-5160-4. URL https://doi.org/10.1007/s10994-009-5160-4.

[19] Richard E. Neapolitan. *Learning Bayesian Networks*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2003. ISBN 0130125342.

[20] Ann Sofie Olesen, Louise Lohse, Anette Boklund, Tariq Halasa, Carmina Gallardo, Zygmunt Pejsak, Graham J. Belsham, Thomas Bruun Rasmussen, and Anette Btner. Transmission of african swine fever virus from infected pigs by direct contact and aerosol routes. *Veterinary Microbiology*, 211:92 – 102, 2017. ISSN 0378-1135. doi: https://doi.org/10.1016/j.vetmic.2017.10.004. URL http://www.sciencedirect.com/science/article/pii/S0378113517307757.

[21] Erik Reed and Ole J. Mengshoel. Bayesian network parameter learning using em with parameter sharing. In *Proceedings of the Eleventh UAI Conference on Bayesian Modeling Applications Workshop - Volume 1218*, BMAW'14, pages 48–59, Aachen, Germany, Germany, 2014. CEUR-WS.org. URL http://dl.acm.org/citation.cfm?id=3020299.3020305.

[22] Stuart Russell, John Binder, Daphne Koller, and Keiji Kanazawa. Local learning in probabilistic networks with hidden variables. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pages 1146–1152, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-363-8. URL http://dl.acm.org/citation.cfm?id=1643031.1643048.

[23] OIE: World Animal Health Information System. Immediate notification: First occurence of a listed disease in the country. URL https://www.oie.int/wahis_2/public/wahid.php/Reviewreport/Review?reportid=24159.

[24] J.M. Snchez-Vizcano, L. Mur, J.C. Gomez-Villamandos, and L. Carrasco. An update on the epidemiology and pathology of african swine fever. *Journal of Comparative Pathology*, 152(1):9 – 21, 2015. ISSN 0021-9975. doi: https://doi.org/10.1016/j.jcpa.2014.09.003. URL http://www.sciencedirect.com/science/article/pii/S002199751400142X.

[25] L.C. van der Gaag. A bayesian network for early detection of classical swine fever in pigs [pdf slides]. 2017.

[26] Linda C. van der Gaag. Workflow alignment of system contents: A case study in the design of a clinical decision-support system for veterinary practice. *Unfinished manuscript*. 09 2012.

[27] Linda C. van der Gaag, Janneke Bolt, Willie Loeffen, and Armin Elbers. Modelling patterns of evidence in bayesian networks: A case-study in classical swine fever. In Eyke Hüllermeier, Rudolf Kruse, and Frank Hoffmann, editors, *Computational Intelligence for Knowledge-Based Systems Design*, pages 675–684, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-14049-5.

[28] A.M.B. Veldhuis. Surveillance of emerging diseases in cattle: Application to the schmallenberg virus epidemic in the netherlands. *Utrecht University*, pages 1–9, 2016.

[29] M. P. Wellman and M. Henrion. Explaining 'explaining away'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):287–292, Mar 1993. ISSN 0162-8828. doi: 10.1109/34.204911.

[30] Yun Zhou, Norman Fenton, and Martin Neil. Bayesian network approach to multinomial parameter learning using data and expert judgments. *International Journal of Approximate Reasoning*, 55(5):1252 – 1268, 2014. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2014.02.008. URL `http://www.sciencedirect.com/science/article/pii/S0888613X14000371`.

A Clinical assessment form Guinat et al. [12]

Clinical Assessment for ASFV

Experiment: .....**Starting Date/time** .....

Responsible for assessment .....

Animal number						
Morning T°C						
Afternoon T°C						
Temperature <39 = 0 39.0 < to < 39.5= 1 39.5 ≤ to < 40 = 2 40.0 to ≤ 40.5 = 3 40.6 ≤ 41 = 4 >41 = 5						
Inappitence - Reduced eating (1) - Only picking at food (4) - Not eating (6)						
Recumbancy - Lethargic (1) - Get up only when touched (2) - Slow to get up when touched (4) - Remain recumbent when touched (6)						
Skin Haemorrhage* - Haemorrhagic areas on ears and body (1) - Generalised haemorrhage all over body (3)						
Joint Swelling - Joint swelling (1) - Severe swelling with difficulty walking (4)						
- Laboured breathing and/or coughing (1) - Severe (3)						
Ocular discharge (1) (gummed up eyes)						
- Diarrhoea (1) - Bloody Diarrhoea (4)						
Blood in Urine (4)						
Vomiting (4)						
Total	/40	/40	/40	/40	/40	/40

B Clinical assessment form Olesen et al. [20]

Temperature	0	< 39.0 °C
	1	39.0–39.5 °C
	2	39.6–40.0 °C
	3	40.1–40.5 °C
	4	40.6–41.0 °C
	5	> 41.0 °C
Alertness and recumbency	0	Alert
	1	Depressed/lethargic
	2	Only gets up when touched
	4	Gets up slowly when touched
	6	Remains recumbent when touched
Appetite	0	Normal
	1	Reduced
	4	Picking at food
	6	Does not eat
Body condition	0	Normal, full stomach
	1	Empty stomach, sunken flanks
	2	Empty stomach, sunken flanks, loss of muscle mass
	3	Emaciated
Skin	0	Normal
	1	Minimal area of the skin with observed bleeding (< 10% of the body)
	2	Moderate area of the skin with observed bleeding (10–25% of the body)
	3	Generalized skin bleeding (> 25% of the body)
Joints	0	No joint swelling
	1	Swelling
	4	Severe swelling and lameness
Respiration	0	Normal
	1	Mildly labored
	2	Labored +/− cough
	3	Severely labored
Eyes	0	Normal
	1	Small amount of exudate
	2	Moderate amount of exudate
Gastrointestinal and urinary tracts	0	No diarrhea
	1	Mild diarrhea for less than 24 h
	3	Diarrhea for more than 24 h or vomiting
	4	Bloody diarrhea or blood in urine
Neurology	0	No signs
	3	Hesitant, unsteady walk, crossing-over of legs is corrected slowly
	4	Pronounced ataxia
	6	Paralysis or convulsions