DECISION SUPPORT FOR PNEUMONIA MANAGEMENT IN PIG PRODUCTION

Michael Höhle
Department of Animal Science and Animal Health,
Royal Veterinary and Agricultural University, Denmark
E-mail: hoehle@dina.kvl.dk

ABSTRACT

This paper exploits pneumonia treatments recordings from a Danish production facility to develop a decision support system (DSS) for on-site handling of the disease. Simple visualizations of the gathered data provide an intuitive retrospective tool to identify problematic areas within the facility or illustrate implicit strategies applied by the workers. To achieve an operational level DSS a mathematical model is developed; on a daily basis each section of the production is assigned to one of the two categories: at risk or not at risk. Using logistic discrimination for this decision problem, the effect of past decisions, neighboring sections, and other covariates are included. Model parameters are estimated from a partition of the dataset and the predictive performance is measured on the remaining part of the data set. Risk of overfitting the training set by introducing additional model parameters is illustrated. The results of the DSS are presented to the users using risk maps – a user-interface in analogy with the earlier intuitive plots and independent of model choice.

1. INTRODUCTION

Modern pig production implies larger production facilities together with decreasing time spent on the individual pig. High stocking rates together with lack of early symptom detection makes health management an important issue. Especially critical are infectious diseases; capable of quickly spreading between farms an outbreak has nationwide if not even global consequences. Introducing automatic surveillance techniques and other type of on-site decision support systems in disease management is a technological possibility to compensate for the reduced amount of time used per pig, see e.g. [Madsen, 2000]. Key issue in developing a system is the ability to predict the occurrence of new cases by modeling the spread of the disease. This being a process of extensive stochastic nature makes handling uncertainty in the domain a crucial matter. Moreover does the temporal structure of the application naturally requires methods of sequential decision making under uncertainty.

The attempt of this paper is to exploit existing on-site registrations of pneumonia treatments – made by the site workers – for such a decision support model. Straight forward visualizations of the data provide valuable insights about pens and sections with increased disease frequency, effect of clearance policies against e.g. PRRS, inexpedient management routines, etc. Being intuitive to understand the visualizations enable an immediate dialogue with the daily staff; a dialogue of great importance. Summarizing plots although only allow a retrospective analysis; to achieve interacting operational level support for early disease detection and prevention, a further step toward the alarm system metaphor is needed: A mathematical model to predict the disease spread has to be developed, filtering noise caused by uncertainty and with that capturing

the systematic patterns in the data. On a daily basis, information about earlier treatments etc. are used to calculate risk of infection for pens and sections of the facility. Based on this daily risk map the staff can decide to keep an higher alert level in certain sections, perform preventive culling, etc. Crafting of these risk maps is a case of decision making: predictions should be as precise as possible to reduce costs inflicted by misclassification. For now, sequential decision making is interpreted as daily repetition of a single classifying decision. Section 4. provides a discussion on modeling a setup, where the warnings of the DSS have an impact on future treatments, thus making decisions connected over time.

2. MATERIALS AND METHODS

Data is provided by the Danish National Committee for Pig Production and originates from a Danish test-facility named *Bøgildgård*. Purpose of Bøgildgård is to evaluate boars in order to select the best for artificial insemination. Registrations cover the period from 8 January 1996 to 23 November 2000 in which a total of 21036 boars passed through the test-facility. For registration purposes each boar is equipped with an ear tag containing a unique identification number allowing registration of e.g. location, introduction and removal date, breed. At weaning a boar, selected at its birth farm for testing, is transported to Bøgildgård, where the first five weeks are spent in a climate-pen. Hereafter the boars are transfered to the test facility shown in Figure 1, where they on average spend 90 days before being removed – either by becoming an AI-boar or by being sent to the slaughterhouse.

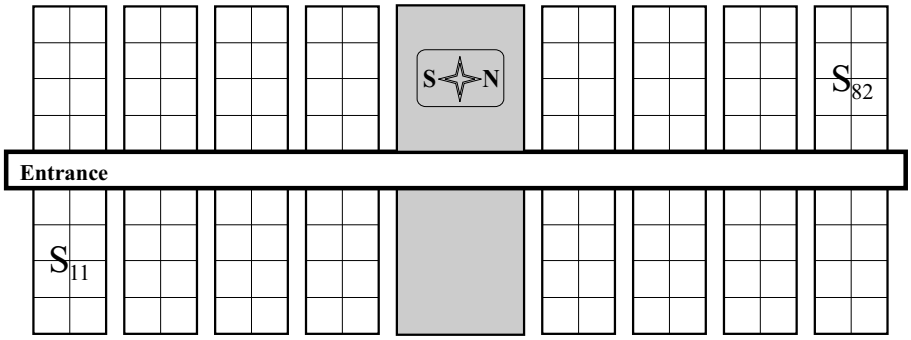


Figure 1: Map of pens and sections of the test facility at Bøgildgård. A section, e.g. S_{11}, consists of 4×2 pens. The grey shaded building contain storage space and office facilities and are therefore not used for boars.

The registered treatments cover medical intervention against a variety of diseases, e.g. diarrhea, tail bite, paralysation. To restrict attention to an infectious disease, interest is only in treatments against what at Bøgildgård is called *pneumonia*. Basically, this definition is used by the staff whenever a pig shows pneumonia like signs, i.e. panting, a dry cough, or if the pig remains lying for longer periods. On a clinical level such symptoms could be induced by PRRS, mycoplasma hyopneumonia, or some type of actinobacillus pleuropneumonia [Larsen and Bækbo, 1997].

Initial decision support consists of straight forward summary plots illustrating the temporal, spatial, and spatial-temporal aspects of the data. Simplicity of this approach makes these plots suitable as basis for a communication with the workers at Bøgildgård. Being intuitive to understand the visualizations enable an immediate communication with the staff. A forte making them convenient to illustrate results of more complicated predictive models needed when moving from retrospective analysis to operational decisions supporting the worker' daily disease

management. Core of the such an operational DSS is a spatio-temporal model for the disease spread by incorporating knowledge about the disease-state in neighboring sections, managment routines etc. Using this information, the model should forecast in which sections to expect sick pigs and alarm workers about sections with an increased risk. Formulating this spatio-temporal model in statistical terms, let s be a section in the facility and Y_t^s a binary random variable describing the disease state in section s at time t, i.e.

$$Y_t^s = \begin{cases} 1 & \text{if one or more } \textit{new} \text{ infections appear in } s \text{ on day } t, \\ 0 & \text{if } \textit{no} \text{ new infections appear in } s \text{ on day } t. \end{cases} \tag{1}$$

Arranging all Y_t^s according to their spatial position gives the farmer a picture of the current state of the disease in his facility – a so called *risk map*.

Predicting the outcome of the disease variables is done using a classifier based on *logistic discrimination* [Ripley, 1996]. Given input about relevant covariates, the classifier decides whether new cases will occur in a specific section. Within a Bayesian framework prior probabilities π_0 and π_1 are assumed for the two states of Y_t^s, respectively. Based on observed covariates, $X_t = x_t$, a statistical model is used to calculate the posterior distribution $p(Y_t^s|x_t)$. A simple classification procedure could now just pick the class with the highest posterior probability, but this would ignore that erroneously sounding an alarm for infected pigs is regarded less crucial than accidently missing to warn about true infections! Assuming that it possible to assign costs of L_{01} and L_{10} units for false-positives and false-negatives, respectively, a minimum loss Bayes classifier chooses the class

$$c(x_t) = \begin{cases} 1 & \text{if } p(1|x_t) > L_{01}/(L_{01} + L_{10}) \\ 0 & \text{otherwise} \end{cases} . \tag{2}$$

It can be shown that the decision rule (2) yields the lowest expected *total risk*, a measure combining misclassification probabilities and cost of misclassification, see [Ripley, 1996]. To obtain an operating classifier and test its performance the first step is to use a *training* data set to estimate the parameters in the parametric model of the posterior distribution. Then the classifier is applied to a *validation* data set while counting number of correct classifications (n_{00} and n_{11}) together with false-negatives, n_{01}, and false-positives, n_{10}. Classifier performance is summarized by its expected cost per case

$$p(Y_t^s = 1)(1 - Se)L_{10} + (1 - p(Y_t^s = 1))(1 - Sp)L_{01}.$$

Here $p(Y_t^s = 1)$ is the prior probability of a day with new infections, i.e. $n_{11} + n_{10}$ divided by the total number of cases, while $Se = n_{00}/(n_{00} + n_{01})$ is the *sensitivity* and $Sp = n_{11}/(n_{10} + n_{11})$ the *specificity* of the classifier.

Posterior probabilities in the logistic discrimination of Equation 2 are obtained using a Generalized Autoregressive Model (GArM) [Fahrmeir and Tutz, 1994]. By conditioning the model on past values, these can be included as additional covariates in a generalized linear model. The general form of a binary and linear GArM using the logistic link function is as follows.

$$\text{logit}(\mu_t) = f(\text{covariates}) + g(\text{past}) = x_t'\gamma + \sum_{i=1}^{l} \beta_i g_i(H_t),$$

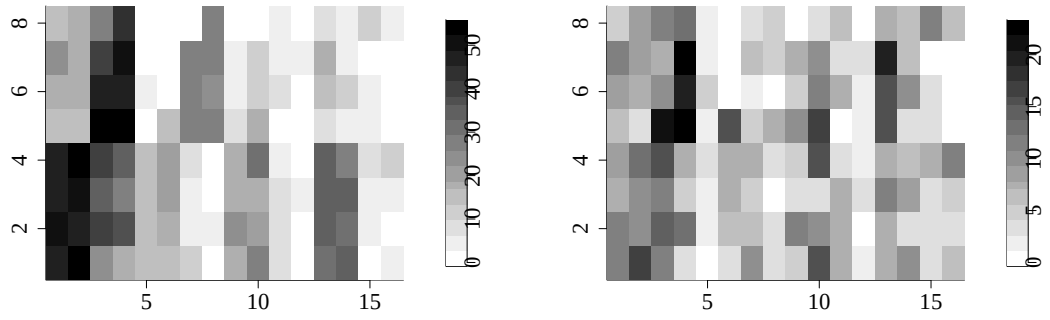
where H_t is the history up to time t, i.e. $H_t = (x_1, y_1, \ldots, x_{t-1}, y_{t-1})$. Typically, g_i is only a function of past responses y_{t-i} and possibly past linear combinations $x_{t-i}'\gamma$. Because only a

finite number, l, of past values are included, the above corresponds to a non-stationary Markov chain of order l.

Unfortunately, selecting appropriate values for misclassification costs is often not as simple as it sounds. A *receiver operating characteristic* (ROC) curve [Greiner *et al.*, 2002] depicts the sensitivity and specificity of a classifier for all possible choices of threshold values in Equation 2. Hence, an ROC curve is a useful graphical illustration of a classifiers discriminative power, independent of misclassification costs. Using the area under the curve is a way to compare classifiers in case nothing is known about misclassification costs. If costs fixed, comparing two classifiers boils down to selecting the one with lowest expected risk; in ROC space this can be translated to finding the intersection of an iso-performance line with the convex hull generated by the ROC curves of all involved classifiers, see [Provost and Fawcett, 1997]. If costs (and possibly priors) are subject to uncertainty, a Bayesian solution is to quantify the uncertainty in $r = \frac{L_{01}}{L_{01}+L_{10}}$ using a distribution function. If the setup is such that the decision maker has initially chosen a value r_0 for r, then found the optimal classifier, and now wants to investigate how robust this choice is to changes in r_0 a traditional sensitivity analysis has to be done. Typically, this boils down to investigating whether the choice of optimal classifier changes if r is varied within the interval $[r_0 - \delta, r_0 + \delta]$.

3. RESULTS

A coarse way to provide insight is to perform visualizations of the Bøgildgård data. Two types treatments against pneumonia are used: individual injections of antibiotics or addition of medication to the water supply of an entire section. Figure 2(b) shows the cumulated number of individually treated pigs for each pen at Bøgildgård, whereas (b) shows the same plot for both individual and water treatments.



(a) (b)

Figure 2: Cumulated number of treatments for each pen in the test facility; (a) shows both treatment forms, where as (b) shows individual treatments only.

The plot reveals, that pens towards the entrance – especially section S_{11} – have a higher disease risk. Discussions with the staff of Bøgildgård revealed that this might be due to an opposite tall barn together with virtually no surrounding vegetation. A setup which can cause fast wind streams to occur, stressing the ventilation system of the section.

Turning to the disease prediction system, an important issue in the design is the choice of GArM. Investigating the effect of various covariates is done by looking at histograms such as Figure 3, which reveals an influence of age of pig as well as season. Discussions with

the staff did not reveal an immediate explanation for the 40-day peak, but it might be due to declining effect of the vaccine used at introduction to Bøgildgård. To keep things simple,

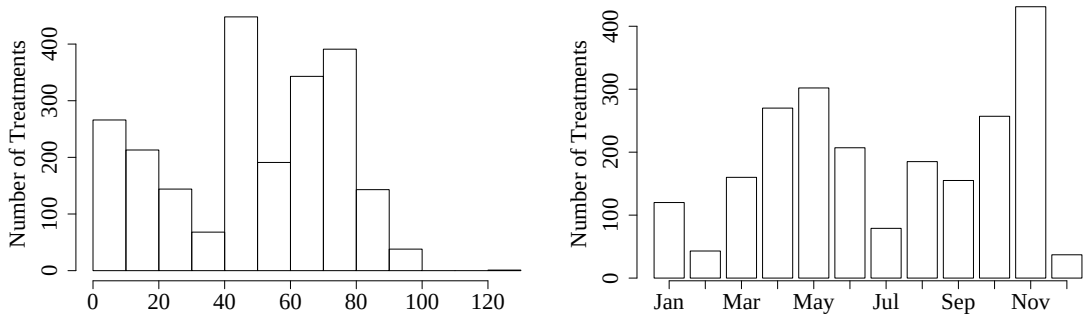


Figure 3: Histogram of treatments split according to (a) days since introduction of the boar to the test-facility and (b) month of the year.

the selected GArM model contains only main effect together with a a straightforward additive nearest neighbour impact. The mean structure for the disease occurrence in section s on day t containing x^s_{no} boars of average age x^s_{age} and t located in season x_{season} thus looks as follows.

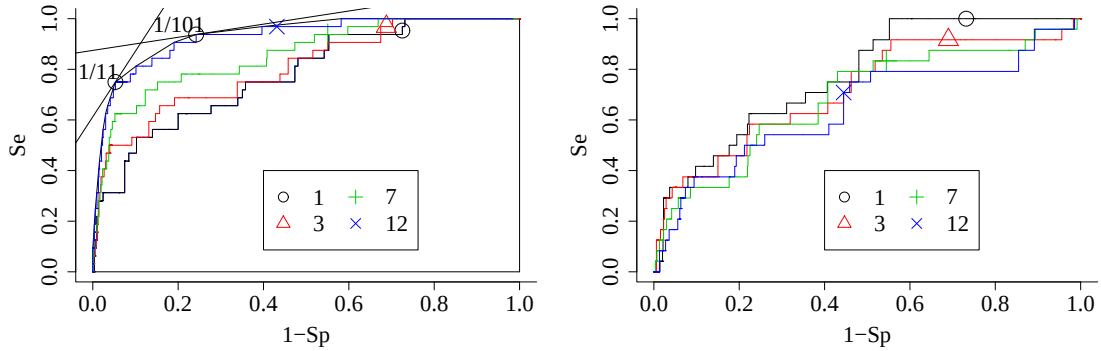
$$\text{logit}(\mu^s_t) \;=\; x^s_{\text{no}}\gamma_1 + x^s_{\text{age}}\gamma_2 + \gamma_3^{\text{season}} + \beta_0 + \sum_{s' \in N^*_4(s)} \sum_{i=1}^{l} \beta_i^{s'} y^{s'}_{t-i}, \tag{3}$$

where $N^*_4(s)$ denotes the four compass direction neighbors of s – if they exist – together with the unit itself, e.g. $N^*_4(S_{11}) = \{S_{11}\} \cup \{S_{12}, S_{21}\}$. The integer l denotes the order of the autoregressive part of the model. Choice of l is a matter of model selection — the auto-covariance function of the time series and an argument that l should be sufficiently large to cover factors such as incubation time of the disease and insertion policy can be used as guidelines. Training the classifier on half the dataset available for a specific section and then evaluating it on the remaining entries, yields a ROC curves for each model. Figure 4 shows that increased l results in nicer fits for the training data, but introduction of the large amount of extra model parameters also results in typical over-fitting revealed by deteriorating performance on the validation set. Best performance is apparently obtained by using as little memory as possible, which indicates that the model is not able to capture many systematic patterns in the data.

4. DISCUSSION

The result from the previous section show that using treatment data to develop a DSS for disease management is not an easy task. Prediction based on logistic discrimination is not able to provide a very high quality. This might both be due high irregularity in the data set and use of a black-box model depending on data only. A white-box model such as e.g. the SIR-model [Andersson and Britton, 2000] based on population dynamics integrates prior knowledge, allowing a more advantageous utilization of the data. Important is, that the visualization methods, e.g. risk maps, for communication with the daily staff is independent of the underlying model. Hence, other choices for the underlying classifier including neural networks, classification trees, etc. [Ripley, 1996], could be investigated.

An intrinsic difficulty in performing disease prediction using treatment data is that treatment only gives partial information about the current disease state; factors such as an implicit man-



(a) Training set (b) Validation set

Figure 4: ROC curves for the two data partitions obtained by various choices of GArM(l) to predicted section S_{11}. Optimal classifier for $r = 1/11$ and $r = 1/101$ are shown in (a) using the convex hull method by [Provost and Fawcett, 1997].

agement strategy together with an imprecise test procedure for disease detection add uncertainty. In theory it is possible to extend the models of this paper to cope for separation between the true – but unobserved disease state – and recorded treatments. Controlled experiments and qualified guesses are then necessary to quantify the connection between those two. Furthermore, risk maps generated by an operating DSS will lead to measures influencing the future health state; a setup corresponding to a partially observed Markov decision process Kaelbling *et al.* [1998]. Besides an almost intractable solution complexity of such a POMDP, such modeling also requires expensive investigations to quantify the effect of classification results on the future state.

References

Håkan Andersson and Tom Britton. *Stochastic epidemic models and their statistical analysis*. Number 151 in Lecture notes in statistics. Springer, 2000.

Ludwig Fahrmeir and Gerhard Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, 1994.

M. Greiner, D. Pfeiffer, and R.D. Smith. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine*, 45:23–41, 2002.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Lars Pilegaard Larsen and Poul Bækbo. *Sundhed og Sygdom hos Svin*. Landbrugsforlaget, 3rd edition, 1997.

Thomas Nejsum Madsen. A Model for Detection of Changes in the Drinking Behaviour of Young Pigs. In L.M Plà and J. Pomar, editors, *Proceedings of the International Symposium on Pig Herd Management Modeling and Information Technologies Related*, pages 61–70, 2000.

Foster Provost and Tom Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, pages 43–48. AAAI Press, 1997.

B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.