September 2003 · nr. 184

Temamøde den 25. september 2003 på Forskningscenter Flakkebjerg

Biometrianvendelse i planteforskningen

Kristian Kristensen (red.)

Biometrianvendelse i planteforskningen

Kristian Kristensen, (red.)
Sekt. for Biometri
Afdeling for Husdyravl og Genetik
Postboks 50
DK-8830 Tjele

Interne rapporter indeholder hovedsageligt forskningsresultater og forsøgsopgørelser som primært henvender sig til DJF medarbejdere og samarbejdspartnere. Rapporterne kan ligeledes fungere som bilag til temamøder. Rapporterne kan også beskrive interne forhold og retningslinier for DJF.

Rapporterne koster i løssalg 50 kr. pr. stk.
Henvendelse til:
Danmarks JordbrugsForskning
Postboks 50, 8830 Tjele
Tlf.: 8999 1010

Alle DJF's publikationer kan bestilles på nettet:
www.agrsci.dk
Tryk: Digisource.dk

Indholdsfortegnelse

Statistical modelling and deconvolution of yield meter data.....	5
<i>Frede Aakmann Tøgersen</i>	<i>5</i>
Abstract	5
Introduction	5
Parameter estimation	10
Results for Veddelev yield meter data	10
Discussion	12
Acknowledgements	13
References	13
Spiser kvæg som en grønthøster, eller er de kræsne.....	14
<i>Jens Henrik Badsberg, Lisbeth Nielsen og Hanne H. Hansen.....</i>	<i>14</i>
Baggrund og målsætning.....	14
Metoder	14
Udvidet beskrivelse af den statistiske metode	16
Resultat: Diæt og selektion	24
Konklusioner vedr. selektion	26
Litteratur.....	26
Vurdering af forsøgsudslag ved hierarkiske designs og gentagne målinger.....	27
<i>Erik Jørgensen.....</i>	<i>27</i>
Abstract	27
Introduktion.....	27
Eksempel på fejlrisikoen	28
Planteforsøg.....	30
Forsøgsdesign skal afspejles i model	31
Udvidet modellering af afhængighed mellem observationer	33
Diskussion.....	34
Modern chemometrical methods in relation to agricultural sciences – is it the light at the end of the tunnel?	35
<i>Søren Højsgaard</i>	<i>35</i>
Introduction	35
Example: Heat development when concrete hardens.....	35
Example: Prediction from IR data.....	37
PCA, PCR and PLS.....	38
Example: Children's choice of apples.....	40
Some (classical) problems where the projection methods may not be appropriate	41
Software for chemometrical analyses	41
Discussion	42
Acknowledgements	42

Modellering af QTL-effekter over tid samt deres vekselvirkning med miljøfaktorer øger statistisk styrke og biologisk forståelse.....	43
<i>Mogens Sandø Lund</i>	43
Sammendrag.....	43
Introduction.....	43
Material and Methods	44
Results.....	46
Discussion	48
References.....	49
Reducing the size of large experiments and treating missing observations.....	50
<i>Rodrigo Labouriau and Peter K. Theil</i>	50
Dansk resumé	50
Abstract	50
Introduction.....	50
Material and methods.....	55
Results.....	57
Discussion and conclusion	59
References	59
Construction of homogeneous zones for field experimentation	61
<i>Ulrich Halekoh and S.E. Olesen</i>	61
Abstract	61
Introduction.....	61
Material and Methods	62
Results.....	64
Discussion	67
References	68
Anvendelse af ufuldstændige blokforsøg.....	69
<i>Kristian Kristensen, Jakob Willas, Lise Nistrup Jørgensen og Rene Gislum</i>	69
Sammendrag.....	69
Introduktion.....	69
Enfaktorielle forsøg.....	70
Flerfaktorielle forsøg.....	73
Konklusion	77
Referencer	77

Statistical modelling and deconvolution of yield meter data

Frede Aakmann Tøgersen

Biometry Research Group, Department of Animal Breeding and Genetics, DIAS

Abstract

We consider the problem of mapping spatial variation of yield in a field using data from a yield meter on a combine harvester. The unobserved yield is assumed to be a Gaussian random function and the yield meter data is modelled as a convolution of the yield and an impulse response function. This results in a spatial covariance structure for the yield meter data that depends on the driving pattern of the combine harvester. Parameters of the impulse response function and the spatial covariance function of the yield are estimated using maximum likelihood methods. The yield is predicted using standard kriging methods.

Key words: deconvolution, Gaussian random field, precision farming, prediction, spatial statistics, yield meter.

Introduction

Precision farming is an agricultural practice where e.g. fertilizers and pesticides are allocated to the field according to locally determined requirements and requires mapping of many types of variables like yield, weed occurrence, soil properties, root zone capacity, etc. In this paper we consider the problem of modelling and mapping the spatial variation of yield in a field. Modern combine harvesters carry a differential global positioning system and a yield meter system. These two systems provide corresponding position and yield data. However, certain delay and smoothing effects in the combine harvester must be taken into account when devising a procedure for yield mapping based on the raw yield meter data.

The delay and smoothing effects in the combine harvester are due to the transport time of the grain from the cutter bar to the yield meter and mixing of grain harvested at different locations. A detailed modelling of the grain flow is not feasible and a simplifying assumption is that a yield meter observation is given by a convolution of the unobserved true yield and a so-called impulse response function, see e.g. Lark et al. 1997. Empirical estimation of the impulse response function can be performed by marking crop in a suitably small region with a dye or water and use this as an “impulse” input to the combine harvester. Subsequently, corresponding observations of the proportion of marked yield in the output of the combine harvester and of distance from the impulse region provides an approximation to the impulse response function. Non-parametric estimation of the impulse response function based on this approach is considered in Whelan and McBratney 1997 and parametric estimation in Lark et

al. 1997, Pringle, Whelan, Adams, Cook and Riethmuller 1999, and Whelan and McBratney 1999.

Attempts to deconvolve yield meter data using an estimated impulse response function and Fourier methods have been made with varying success in Lark et al. 1997, Pringle et al. 1999, and Whelan and McBratney 2000. The deconvolution problem is ill-posed and Whelan and McBratney 2000 apply presmoothing of the yield meter data in order to reduce the noise sensitivity of the solution. Regularization methods (O'Sullivan, 1986) are not considered. After deconvolution a next step would be to create a yield map by interpolating the deconvolved values.

We consider parameter estimation and deconvolution using methods from spatial statistics (Ripley, 1981, Cressie, 1993, and Stein, 1999). The yield is assumed to be a realization of a random function and a spatial model for the yield meter data then follows from a convolution of the yield with an impulse response function. Using parametric models for both the impulse response function and the covariance function of the yield allow us to compute maximum likelihood estimates of the parameters. Once the parameters of the model are computed, predictions on a dense grid of locations are easily obtained using standard kriging methods. This corresponds in fact to a deconvolution of the observed yield meter data, where regularization is imposed by the spatial covariance of the yield. The advantages of our approach are that we obtain parameter estimates without expensive experimental field trials and that the estimated spatial covariance function provides a data driven regularization function.

A yield meter dataset

We consider 3360 yield meter observations from an experimental field at Veddelev belonging to Risø National Laboratory. The field was grown with barley and different amounts of nitrogen (0, 54, 90, or 116 kg N/ha) were applied in parallel strips. Each of these strips was harvested separately. The driving tracks of the combine harvester are shown in Figure 1. The sampling locations are located on the driving tracks with approximately 6 m between each track and approximately 4 m between consecutive sampling locations on the tracks. The empirical treatment means/std.-deviations are 3.88/0.51, 5.11/0.46, 5.41/0.42, and 5.54/0.35, for 0, 54, 90, and 116 kg N/ha, respectively. These values indicate that the amount and uniformity of the yield is positively correlated with the amount of nitrogen applied.

Figure 1. Driving tracks of combine harvester, arrows indicate driving direction

In order to study the spatial distribution of the yield we first obtain residual yield meter values by subtracting the appropriate empirical treatment means from the observations. Figure 2 shows an interpolated map of residual yield meter values obtained using kernel smoothing with a subjectively chosen bandwidth equal to 15 *m*. The interpolated values range between -1.41 and 0.93. Choosing a smaller value for the bandwidth results in a noisier looking map whereas a bigger value results in a smoother map. This mapping technique does not take into consideration the smoothing of yield taking place in the combine harvester.

Figure 2. Interpolated residual yield meter values on a 5 m by 5 m grid. Light grey scales correspond to small values; see also the legend in the plot. The white line shows the outline of the field (i.e. the region occupied by the driving tracks in Figure 1).

Spatial modelling of yield meter data

Readers interested in the technical details of the statistical modelling of yield meter data may consult Tøgersen and Waagepetersen, 2002. Here it suffice to note that the unobserved yield is a random function, $Z(x)$, of a location x in the field. We assume that the residual process of $Z(x)$ forms a Gaussian field with constant variance and isotropic covariance function.

Furthermore, we assume that the yield meter value at x is given by

$$Y(x) = \int_{s \in D} \kappa_{d(x)}(x-s)Z(s)ds + \varepsilon(x),$$

where $\kappa_{d(x)}(\cdot)$ is the impulse response function, $d(x)$ denotes the driving direction at x , and $\varepsilon(x)$ is measurement noise. By numerical integration it is possible to calculate the anisotropic covariance function of $Y(x)$ when the functional form of the covariance function of $Z(x)$ and the impulse response function are known. For example, assume that the impulse response function is given by a product of functions taking care of smoothing across and along the driving direction. We may assume that the smoothing across the driving direction is the same across the cutting bar, thus using the density function of a uniform distribution of a width of the width of the cutting bar. Considering the delay/smoothing effect in the combine harvester

modelled by a convection-dispersion equation known from mass flow studies in hydrology and soil sciences (Jury, Gardner and Gardner, 1991), the solution to this may be used as smoothing kernel in the driving direction. Examples are shown in Figure 1. The length of the curve before the sudden increase is related to the delay effect (proportional to the distance between the cutting bar and the yield meter). The heavier tail to the right, the more mixing of grain in the combine harvester.

Figure 3. Plot of different smoothing kernels. The kernels are from the smoothing in the driving direction estimated in our case, from Whelan and McBratney 1999, and Pringle et al. 1999.

Using the Gaussian correlation function $\rho(r) = \exp(-r^2)$ as the covariance function of $Z(x)$ together with the above smoothing kernel enables us to calculate the covariance function of $Y(x)$. It turns out that this covariance function is anisotropic and depends on the driving directions of the sampling points; see Tøgersen and Waagepetersen, 2002 for the details.

Figure 4. Plots of the Gaussian covariance function $\rho(r) = \exp(-r^2/15^2)$, and corresponding functions K_0 and K_1 for specific values of the parameters of the smoothing kernel. Note the asymmetry of K_1 in the east-west direction.

Figure 4 shows plots of a Gaussian covariance function $\rho(r) = \exp(-r^2/15^2)$ and the corresponding functions K_0 and K_1 . The former is the covariance function for points with same driving directions whereas the latter is the covariance function for points of different driving directions. Note the asymmetry of K_1 in the east-west direction.

Parameter estimation

The parameters of the model for yield meter data are estimated using maximum likelihood. The sample sizes of yield meter datasets are very large compared to typical datasets in spatial statistics. In the Veddelev data, for example, the sample size is 3360 and this makes maximum likelihood estimation for the entire dataset impractical. We therefore consider the option of using only a subset of the yield meter data.

Assuming that both $Z(x)$ and $\varepsilon(x)$ are Gaussian distributed, $Y(x)$ is also Gaussian. It is thus possible to write up the likelihood function for a particular dataset assuming that $Z(x)$ and $\varepsilon(x)$ are independent. We may by a numerical optimisation routine find the values of the parameters that maximize the likelihood function. These values are, under the chosen model, the values that give the highest probability observing the data at hand. The details are given in Tøgersen and Waagepetersen, 2002.

Results for Veddelev yield meter data

For the Veddelev data we consider a model where the covariance function of $Z(x)$ is the sum of two Gaussian covariance functions. One of the Gaussian functions is taking caring of short range correlation, and another of the long range correlation.

Parameter estimates

Basing the parameter estimation on observations for a random subsample of size 840 of the 3360 sampling locations we obtain estimates of the parameters of the smoothing kernel. A plot of the smoothing kernel is shown in Figure 3 as the continuous line.

The first component of the covariance function of $Z(x)$ is practically a white noise with a high variance. The second component exhibits long range correlation. A plot of the sum is shown in Figure 5.

Figure 5. A plot of the sum of two Gaussian covariance functions. The components of the sum are those estimated from the Veddelev data.

Deconvolution

Since the first component of the yield process exhibits high frequency variation, which is not manageable with current precision farming technology, we restrict attention to the slowly varying second component. We predict this at 6958 locations on a 5 m by 5 m grid of dimension 98 by 71 covering the Veddelev field.

Figure 6 shows the prediction obtained by kriging the residuals obtained during the estimation process. The unknown parameters of the model are replaced with the maximum likelihood estimates. The predicted values of the slowly varying part of the random function range between -1.47 and 1.36. The variation is thus substantial compared with the fitted mean function (not mentioned here but see the paper by Tøgersen and Waagepetersen, 2002) of the yield Y and there is scope for using methods of precision farming to e.g. target fertilizer application at areas with low predicted values.

Incidentally, the kernel smoothed yield (Figure 2) obtained with a subjectively chosen bandwidth is visually rather similar to the prediction in

Figure 6. Note that the kernel smoother used for producing Figure 2 only provides values for locations closer than 15 meter to the field. However, the kernel smoothing does not provide the required deconvolution.

Figure 6. Predictions of the slowly varying part of $Z(x)$. Light grey scales correspond to low predicted values (see legend in plot). White line shows the outline of the field.

Discussion

We have established a method to kriging raw yield meter data taking into account the anisotropic behaviour of data due to the way data have been sampled. At the same time regularisation is performed to reduce the inherent noise sensitivity of deconvolution problems.

For computational reasons we restrict attention to a sum of Gaussian covariance functions. It is of interest to consider alternatives like spherical or Matérn covariance functions. However, for these covariance functions higher-dimensional numerical integration is required and this leads at present to unacceptable computing times.

Edge effects are ignored in our computations. Considering the heavy tail of the fitted impulse response function this may not be wise. In principle it is straightforward to calculate the marginal distribution for observations near the edges of the field but in practice a lot of extra programming is required. In our analysis we have also ignored the fact that the yield meter observations do not fall on perfectly straight and parallel lines. The deviations from straight lines are mainly due to GPS measurement error so it might be helpful to rectify the data by regressing the sampling locations onto straight lines.

The implications of using a particular parametric form of the impulse response function deserve further investigation. A too heavy tail dictated by the parametric model may e.g.

result in an amplification of the estimated yield variance and it would be interesting to consider also an impulse response function with bounded support.

Acknowledgements

We are grateful to Rasmus Nyholm Jorgensen at the Risø National Laboratory in Denmark for providing the Veddelev yield meter data. This work was supported by DINA – the Danish Network for Informatics in Agriculture, and by the Danish Natural Science Research Council.

References

- Cressie, N. (1993), *Statistics for Spatial Data*, New York: John Wiley & Sons.
- Jury, W. A., Gardner, W. R., and Gardner, W. H., 1991. *Soil Physics*, London: John Wiley & Sons.
- Lark, R. M., Stafford, J. V., and Bolam, H. C., 1997. "Limitations on the Spatial Resolution of Yield Mapping for Combinable Crops," *J. Agric. Eng. Res.*, 183-193.
- O'Sullivan, F., 1986. "A Statistical Perspective on Ill-posed Inverse Problems," *Statistical Science*, 1, 502-527.
- Pringle, M. J., Whelan, B. M., Adams, M. L., Cook, S. E., and Riethmuller, G., 1999., "Yield deconvolution - a wetted grain pulse to estimate the grain flow transfer function?," in *Proceedings of the Fourth International Conference on Precision Agriculture*, Robert, P. C., Rust, R. H., and Larson, W. E. (eds.), pp. 1177-1184.
- Ripley, B. D., 1981. *Spatial Statistics*, New York: John Wiley & Sons.
- Stein, M. L., 1999. *Interpolation of Spatial Data*, New York: Springer Verlags.
- Tøgersen, F. A. and Waagepetersen, R. P., 2002. "Statistical Modelling and Deconvolution of Yield Meter Data," *Scandinavian Journal of Statistics*.
- Whelan, B. M. and McBratney, A. B., 1997. "Sorghum grain flow convolution within a conventional combine harvester," in Stafford, J. V. (ed.) pp. 759-766.
- Whelan, B. M. and McBratney, A. B., 1999. "A Parametric Transfer Function for Grain Flow within a Conventional Combine Harvester," *Precision Agriculture*, 3, 123-134.
- Whelan, B. M. and McBratney, A. B., 2000. "An Approach to Deconvoluting Grain-flow within a Conventional Combine Harvester Using a Parametric Transfer Function," *Precision Agriculture*, 2, 389-398.

Spiser kvæg som en grønthøster, eller er de kræsne

Jens Henrik Badsberg¹, Lisbeth Nielsen² og Hanne H. Hansen³

¹ *Forskergruppe for Biometri, Afdeling for Husdyravl og Genetik, DJF*

² *Agro Business Park*

³ *Inst. for Husdyrbrug og Husdyrsundhed, KVL*

Baggrund og målsætning

I dette arbejde ønsker vi at undersøge, om stude under en ekstensiv græsning foretager en selektion mellem de planter, de spiser. Endvidere ønskes undersøgt, om denne selektion afhænger af belægningsgrad (høj/lav), studens alder (1. års / 2. års), tidspunkt på sæson og evt. lokalitet (øst/vest), samt år (1998/1999). Endelig ønsker vi at se, om selektionen kan forklares ved arternes fordøjelighed (fordøjeligt organisk stof, råprotein og NDF) samt udbredelse.

Den overordnede baggrund for projektet var at få nogle kvantitative beskrivelser af, hvorledes man med forskellige former for ekstensiv græsmarksdrift kunne bevare eller øge artsdiversiteten på ferske enge. Samtidig ønskedes at kvantificere konsekvenserne for den landbrugs-mæssige produktion på arealerne. Ferske enge varierer meget med hensyn til plantebestand, jordbundsforhold og fugtighed, og det er derfor væsentligt at kende disse faktorer, når man skal vurdere, hvilken produktion, der kan forventes på et givet engareal.

Når stude og får afgræsser på halvkulturarealer er der et større udvalg af arter end på intensive græsmarker. Desuden er der som regel en lavere belægning, hvilket betyder at dyrene har god mulighed for at selektere i det planteudbud, der er til rådighed. Målet med denne del af undersøgelsen var at få en beskrivelse af produktionens størrelse og fordeling over året, artssammensætningen, benyttelsens indvirkning på arterne, foderkvaliteten af enkeltarter og studenes fødevalg.

Resultaterne anført her er et uddrag af resultaterne givet i Nielsen et. al (2002a) med uddybning vedrørende de statistiske beregninger.

Metoder

Græsningsundersøgelserne i basisprojektet foregik med 1. års stude og 2. års stude, der afgræssede ved to afgræsningsintensiteter. Ved høj intensitet blev der stilet mod en afgræsningshøjde på 6 cm og ved lav intensitet var der halvt så mange stude per arealenhed, som ved høj intensitet. Arealet var inddelt i to blokke, vest og øst.

Planteproduktionen i græsningsfoldene blev målt gennem græsnings sæsonen i seks tilvækstperioder i perioden fra udbinding til indbinding. Efter hver tilvækstmåling blev næste måling udført i et nyt område i afgræsningsfolden. Foderkvaliteten blev analyseret i plantematerialet over 2 cm.

Den botaniske sammensætning blev fulgt ved at sortere planteprover pr. art, samt en fraktion med dødt materiale, og således få en beskrivelse af plantebiomassen på tørstofbasis. Analyser af den botaniske sammensætning i sommerperioden første år er vist i Nielsen et al. (2002b), hvor der er givet en beskrivelse af vest og øst, samt af "Læstenvej", der beskriver forholdene for 2. års stude i øst, hvor plantebestanden var noget anderledes end resten af øst-blokken. Som det fremgår, var der en relativt højere andel af græsser i øst og en relativt højere andel af urter i vest. For 2. års stude var der ikke så store forskelle mellem blokkene med hensyn til botanisk sammensætning som for 1. års stude. Enkeltarter blev indsamlet i juni, juli og september til foderkvalitets analyser.

Med henblik på at undersøge, hvad dyrene åd, blev planterester i vomfistulerede dyrs fækalie og vom-indhold undersøgt ved mikroskopi, og om det blev vurderet om der fandtes forskelle mellem vom- og fækalieprøver. Vom- og fækalieprøver blev udtaget en gang om dagen i fire dage én gang om måneden fra juni til september i 1998 og 1999. Prøver blev taget fra fire førstegangs græssende, vomfistulerede stude i 1998 og 1999 og fire vomfistulerede stude, der græssede på andet år i 1998. Studene blev grupperet parvist indenfor aldersgruppe og græssede sammen med andre dyr i enten vest- eller østfoldene. Fækalieprøver blev taget direkte fra dyrene samtidig med udtagningen af vomprøverne.

En referencesamling på 31 planter blev indsamlet fra Fussingø i 1997 og 1998. Karakteristikkerne fra denne referencesamling dannede grundlag for mikroskopiske identifikationer. Fotos af mikroskopiske udsnit af referencesamlingen kan ses på internettet på <http://www.kvl.dk./htm/hhh/text.htm>.

Prøverne blev tørret, formålet, rensat og anbragt på objektglas, og fem objektglas pr. prøve blev undersøgt. Fravær eller tilstedeværelsen af de refererede arter blev noteret i hvert af 20 felter på de 5 objektglas. Ud fra dette er en relativ frekvens for hver af de refererede arter udregnet.

For at kunne bestemme selektiviteten hos de græssende stude er de relative frekvenser af arterne fundet ved mikroskopi i vom- og fækalieprøverne sammenlignet med planternes frekvenser udregnet fra biomasseprøverne for den botaniske sammensætning på marken. Selektiviteten er en sammenligning af hvad der fandtes i dyrene med hvad der fandtes tilgængeligt i miljøet. Hvis dyrene ikke udviser selektivitet, vil den relative mængde, som findes i dyrene, ikke være signifikant forskellig fra den relative mængde, som findes i miljøet. Æder de forholdsvis mere af en planteart end, hvad der findes tilgængeligt, vil der være en positiv selektivitet for denne planteart, og omvendt vil der være en negativ selektivitet for de arter, som findes i dyret i signifikant mindre mængde end i miljøet.

De relative frekvenser blev, efter en logit transformation, modeleret ved en lineær normal model med flere varians-komponenter, i PROC MIXED, SAS version 8.2. I denne model bestemmes selektionen som kontrasten mellem de estimerede relative frekvenser af plante-materiale i hhv. biomasseprøver og vom/fækalieprøver. Middelværdien, og selektionen, af den enkelte art antages i modellen at afhænge systematisk af måned (dvs. tidspunkt på sæson), studens alder og græsningsintensitet, og sæsonvariationen i selektionen af studens alder og græsningsintensitet. I modellen er der varianskomponenter for at målingerne er korrelerede over sted og tid: Der er varianskomponenter for parcellen indenfor folden, for folden indenfor delblokken (ved aldersgruppe), og for blok (øst/vest). Endvidere er der varianskomponenter for år og måned indenfor år, med disse bidrag også afhængig af lokalitet (blok og delblok).

Udvidet beskrivelse af den statistiske metode

Notation

Målingerne er de relative frekvenser (med hensyn til (14 af) 25 arter) hhv. på marken (3 parceller) og i studen (2 stude med vom og fæces) på forskellige tidspunkter (2 år og 4-6 tidspunkter pr. år) og marker (8 stk. ved kombinationen af lokalitet, dyreart og intensitet):

- a : Art,
- i : Intensitet (høj/lav græsningsintensitet),
- d : Dyreart (1. år /2. års stude),
- l : Lokalitet (øst/vest),
- $\hat{a}$: År: 1998 eller 1999,
- t : Tidspunkt på sæson (maj/juni/juli/august/september).

For observationerne af arterne på marken:

- r : Ramme (1, 2 eller 3).

For observationerne af arterne i de fistulerede stude:

- p : Stud 1 eller 2;
- s : Sample (vom eller fæces).

Selektion

Vi ser på en en-dimensional variabel, f.eks. tørstof eller den relative frekvens af en enkelt art i fistuleret stud eller på mark. Hvis vi for en given art lader hhv. X og Y betegne hhv.

- $Y = X_m(a)$: Relativ frekvens af arten a fundet ved mikroskopisk undersøgelse af materiale fra fistuleret stud,

- $X = X_b(a)$: Relativ frekvens af arten a fundet ved analyse af botanisk sammensætning af materiale fra marken,

så kan selektionen udtrykkes ved $A = Y / X$ eller $S = \log(A) = \log(Y / X) = \log(Y) - \log(X)$. X og Y er relative frekvenser, sandsynligheder. $S = \log(A)$ kalder vi selektionen, således at vi får en selektion på 0, hvis arten findes med samme frekvens i studen som på marken. Selektionen er positiv, hvis frekvensen af arten er større i studen end på marken, og selektionen negativ, hvis frekvensen af arten er mindre i studen end på marken. Selv om studen spiser en art, men kun med den halve frekvens af hyppigheden af arten på marken, så er selektionen altså negativ, skønt studen også vælger at spise denne art.

Plots

Plot af den transformerede hyppighed $\log(Y)$ af arten i studen mod transformeret hyppighed $\log(X)$ af samme art på marken giver et billede af selektionen. I det enkelte plot indgår kun observationer for en art. Observationer over ramme i mark og over stud og sample (vom/fæces) er aggregeret for at få sammenhørende værdier. I plottet er et punkt for hver kombination af lokalitet, dyreart, intensitet, år og tidspunkt på sæson. Hvis punkterne fordeler sig om identiteten $y = x$, dvs. diagonalen spises arten med samme hyppighed som arten er på marken, og der er således ingen selektion. Af ovenstående formel ses, at en evt. selektion $S = \log(A)$ er skæringen med y -aksen. En positiv selektion vil således give punkter over diagonalen, mens negativ selektion giver punkter under diagonalen, i begge tilfælde stadig omkring linier parallel med diagonalen. Hvis punkterne ligger om en linie med hældning mindre end 1, så selekterer studen arten ved lave frekvenser på marken, men fravælger, når planten er hyppig. Ved en hældning større end 1 vil hyppigheden af arten variere mere i studens diæt end på marken.

Regression af hyppighed af art i stud på hyppighed på mark

Regression af den transformerede hyppighed $\log(Y)$ af arten i studens diæt på den tilsvarende hyppighed $\log(X)$ af arten på marken svare til antagelsen $\log(Y) = \alpha + \beta \log(X)$ omkring middelværdier. Skæringen α er her selektionen $S = \log(A)$, mens hældningen β burde være 1 under antagelsen $S = \log(A) = \log(Y / X)$. Ved tilbage transformation $A = Y / X = \exp(\alpha) X^{-(1-\beta)}$ har vi en større selektion ved lave frekvenser end ved højere frekvenser på marken, når β er mindre end 1.

Idet den primære fokus i denne analyse er, om der er en selektion eller ikke, og ikke så meget hvordan den øget hyppighed af en given art i diæten afhænger af hyppigheden på marken vil vi i første omgang undersøge modeller, hvor β er sat til 1.

Andre transformationer

I det foregående har vi transformeret hyppighederne af de enkelte plantearter med logaritme-transformationen. Derved har vi opnået lineære sammenhænge mellem de transformerede hyppigheder og selektionen. Endvidere får vi, at de transformerede hyppigheder med rimelighed kan antages normalfordelte, modsat de rå frekvenser.

Alternativer til logaritme transformationen er den komplementær log-log transformation, $\log(-\log(1-X))$, og logit-transformationen, $\log(X/(1-X))$, eller den inverse til en kumuleret fordelingsfunktion, f. eks normalfordelingen.

De tre transformationer $\log(X)$, $\log(-\log(1-X))$ og $\log(X/(1-X))$ ligger tæt på hinanden for lave frekvenser: $0 < X < 0.2$.

Ved logit transformationen og $S_{\text{Logit}} = \log(Y/(1-Y)) - \log(X/(1-X))$ får vi $\exp(S_{\text{Logit}}) = (Y/(1-Y)) / (X/(1-X)) = (Y/X) * (1-X)/(1-Y)$, altså et lidt større range af selektionen: Med $Y = X + \Delta$ ser vi et større estimat af selektion, når den relative frekvens er stor $(Y/X) * (1-X)/(1-Y) = ((X+\Delta)/X) * ((1-X)/((1-X)-\Delta))$.

Metode 1: Empiriske mål for selektion direkte afledt af sammenhørende værdier af hyppigheder til samme tidspunkt og samme sted

Vi kan bestemme selektionen til en given kombination af tidspunkt og position, når vi har sammenhørende værdier af begge frekvenser (for hhv. frekvensen af arten på marken og frekvensen af arten i studen) og disse ligger mellem 0 og 1. Vi aggregerer her inden selektionen beregnes observationerne af de botaniske sammensætninger over de 2-3 rammer på marken, og de mikroskopiske undersøgelser af planterester i de vomfistulerede dyrs fækali- og vom-indhold over stud og sample (vom/fæces).

Vi kan derefter i et andet trin i en normalfordelingsmodel, med korrelerede målinger undersøge, hvorledes disse tal afhænger af de relevante faktorer, dvs. intensitet, studens alder, lokalitet, tidspunkt på året, år, fordøjelighed, etc. Modellen svare til en regression af den transformerede hyppighed af plantearten i studen på hyppigheden af arten på marken, med begrænsningen, at hældningen skal være 1 således kun selektionen estimeres.

Frekvenser observeret til 0 (og 1) giver dog et lille problem.

Metode 2: Mål for selektion via model på transformerede hyppigheder:

Ved at modellere begge hyppigheder af plantearten (i hhv. botanisk sammensætning på mark og i mikroskopiske undersøgelser af planterester i vomfistulerede dyrs fækali- og vomindhold) i en og samme model med en kovariat for om prøven er fra botanisk analyse mark eller mikroskopisk undersøgelse af materiale fra vomfistulerede dyr, og afhængig af intensitet (høj/lav), studens alder (1. års/2. års), og lokalitet (øst/vest) og tidspunkt på sæson samt år (1998/1999), kan selektionen afledes som forskellen, kontrasten, mellem estimaterne for hhv. ”på mark” og ”i dyr”. Modellen kan give et samlet estimat af selektionen af den

enkelte art, under hensyn til, at selektionen evt. afhænger af intensitet og studens alder. Dette samlede estimat af selektionen af den enkelte art kan så senere relateres til de mål (for bl.a. fordøjelighed), der kun findes på artsniveau. Altså igen en 2-trins model.

Andre fordele ved modellen i forhold til modellen på de empiriske selektioner:

- Vi er tæt på de oprindelige observationer, og kan styre
 - o fordeling og,
 - o korrelation mellem målinger.
- Hvis en frekvens af en given art til given kombination af sted og tidspunkt ikke er observeret, vil "gennemsnit" af målinger relevant under modellen blot blive anvendt. F.eks. ved ingen interaktion mellem tidspunkt og sted, og botanisk analyse ("på mark") manglende for given art til givet tidspunkt og sted, så "anvendes" for denne kombination af tid og sted en passende kombination af hovedvirkninger af sted og tidspunktet. Vi behøver altså ikke indsætte en kunstig værdi, hvor frekvensen ikke er observeret eller observeret til 0. (Løsningen her er ikke "ideel", idet vi med 0-et jo ved, at værdien nok er lavere end de tilsvarende andre observationer for samme art.)
- Observationerne af hhv. botaniske analyser på mark og mikroskopisk analyse for dyremateriale behøver ikke ligge på samme tidspunkt, hvis tidspunkt (på sæson) indgår som kovariat.

Arter overhovedet ikke observeret, ved enten botanisk analyse eller mikroskopiske undersøgelser

Ingen af de to ovennævnte metoder vil kunne bruges (tilfredsstillende), hvis vi ønsker at beskrive, hvorfor en given art overhovedet (dvs. for alle kombinationer af sted og tid) ikke er observeret i studen, men findes på marken, eller er observeret i studen, men ikke på marken. Situationen er f.eks. at en art overhovedet ikke er observeret på nogen lokalitet over de 2 år ved botaniske analyser på marken, men arten er fundet i vom eller fæces.

- Metode 1: Mål for selektion direkte afledt af sammenhørende værdier: X er (kan være) for alle par af observationer (botanisk undersøgelse og mikroskopisk undersøgelse) sat til et lille "epsilon", for ikke at dividere med 0. Selektionen, middelværdien af $\log(Y) - \log(\text{"epsilon"})$ bliver så direkte proportional med $\log(\text{"epsilon"})$, og vi har så selv helt bestemt, hvad selektionen bliver.
- Metode 2: Mål for selektion via model på hyppigheder: Metoden vil bare ikke give et bud på selektionen af arten.

Problemet er helt tilsvarende, hvis arten overhovedet ikke er observeret i mikroskopiske undersøgelser af vom- eller fækalie-materiale, men er fundet ved botaniske sammensætning på marken. Vi har ganske vist ikke problemet allerede hvor vi "dividere", men vi kan ikke tage logaritme (angives som teknisk grund).

Men der findes naturligvis metoder for at håndtere sådanne problemstillinger, hvor responset nogle gange er en speciel værdi, og andre gange ligger i et interval, dvs. har en sammensat fordeling. Vi kan sige, at vi med en sandsynlighed (afhængig af vores forklarende variable) observere 0 (for hhv. botanisk sammensætning og mikroskopiske undersøgelser), og givet observationen er større end 0, så er en (logaritmisk) transformation deraf f.eks. normalfordelt.

Her er vi, når nogle arter overhovedet ikke er observeret (enten i botanisk sammensætning eller mikroskopiske undersøgelser) interesseret i at modellere, om arten er observeret. Derfor reducerer vi vores frekvens til en binær variabel; 0 hvis arten ikke er observeret, og 1, hvis frekvensen er større end 0. Vi kan så, ved kontingenstabeller, undersøge, hvorledes den simultane fordeling af de 2 binære variable for hhv. botanisk sammensætning og mikroskopiske undersøgelser afhænger af vores forklarende faktorer, intensitet, studens alder, sæson tidspunkt, lokalitet og år. De kontinuerte variable for fordøjelighed kan også indgå i kontingenstabellen - ved en omkodning til diskrete variable, eller vi kan anvende logistisk regression.

Vi kan også vælge at omkode frekvensen til en variabel med f.eks. 4-5 niveauer:

- 0: Fraværende, dvs. ren 0;
- 1: Svag, f.eks. mellem 0 og nedre kvartil (for arten);
- 2: Mild, f.eks. nedre kvartil og median (for arten);
- 3: Moderat, f.eks. median til øvre kvartil (for arten);
- 4: Kraftig.

Relation mellem selektion og tilgængeligt materiale:

Det er ikke sundt at lave en regression af selektion (bestemt ved $\log(Y) - \log(X)$ eller $\log(Y / (1-Y)) - \log(X / (1-X))$ (fra både metode 1 og metode 2) i ovenstående] på botanisk analyse, X , tilgængeligt materiale, når det er det samme X , der indgår for de enkelte observationer.

Selvom der ikke er en sammenhæng mellem udbredelse og selektion, så vil de 2 størrelser jo være korrelerede pga. den funktionelle sammenhæng mellem udbredelse og den beregnede selektion.

I ovenstående er der overvejelser om regression af hyppighed af art i mikroskopiske undersøgelser på hyppighed af art ved botanisk analyse, således selektionens afhængighed af hyppighed i botanisk analyse kan modelleres.

Man kan anvende inventering som mål for tilgængeligheden af den enkelte art, når der laves regression af selektion (bestemt ved stude-data i forhold til klip-data) på tilgængelighed. Inventeringen (Hald, 2002) er et andet mål for botanisk sammensætning, et mål, der ikke er beskrevet i denne artikel. Så er (burde være), givet sted og tidspunktet, den stokastiske støj på målet for selektionen uafhængig af støjen på målet for udbredelsen af arten. Alternativt kunne man anvende den ene stud og ramme 1 til estimation af selektion, og relatere denne selektion til den anden stud og ramme 2 (og evt. ramme 3), men der er stadig en korrelation.

Lidt overvejelser omkring korrelationsstruktur:

a) Botanisk sammensætning af materiale fra marken, klip-data:

Hvis en art et givet tidspunkt er til stede, ikke bare i forsvindende mængder, så er det også sandsynligt, at den ses ved observationen det følgende tidspunkt. Observationerne samme sted er altså korrelerede, og vi skal foruden målestøjen på den enkelte observation have varianskomponenter for lokalitet. Målinger i samme ramme indenfor samme mark er korrelerede, hvorfor vi har en varianskomponent for ramme i mark, og en varianskomponent for mark.

Klimaet virker tilfældig over de to år (giver på nogle tidspunkter gode betingelser for den enkelte art), hvorfor vi har en varianskomponent for tidspunkt, evt. afhængig af lokalitet.

Vi kan vælge at betragte målinger på samme lokalitet som gentagne målinger.

b) Mikroskopiske undersøgelser af materiale fra vomfistulerede stude:

Den enkelte stud kan have sine præferencer. Derfor en varianskomponent for stud, svarende til komponenten for ramme. Dog er samme stud anvendt på flere marker. En evt. forskel mellem vom og fæces kan betragtes som systematisk.

- o Ellers er der for stude samme varianskomponenter for mark og tidspunkt som ved data fra botaniske analyser på mark.

For at lette beregningen af selektion ønsker vi at aggregere målingerne på de 2-3 rammer indenfor samme mark til samme tidspunkt, og tilsvarende de 2 gange 2 målinger i hhv. vom og fæces i de 2 stude. Deraf følgende problemer pga. korrelation synes at være størst som følge af summering over de 3 rammer, mindst som følge af summering over vom-fæces, og summering over de 2 stude der imellem:

Hvis vi sammenligner målingerne over de 2 år taget på en mark med de tilsvarende målinger på en anden mark, så kan vi ikke afgøre, om en forskel skyldes forskel i behandlinger på de 2 marker, eller tilfældige forskelle mellem de 3 parceller på den ene mark med de 3 parceller på den anden. (F.eks. sammenligning af høj/lav stratificeret til blok og studens alder.) Ved

summering over stud er problemet tilsvarende, hvis vi begrænser materialet til 2 marker, og det er forskellige par af stude, der går på de 2 marker. Vi kan ikke afgøre, om en evt. forskel skyldes behandling (høj/lav) eller forskel mellem de 2 par af studes selektion. (Denne sidste sætning er vist "akademisk", idet det netop er samme par, der for givet blok (øst/vest) og alder er anvendt på de 2 intensiteter.)

Resultater, mål for selektion via model på hyppigheder (metode 2):

Resultater for modellering af logit-transformationen af de relative frekvenser i en og samme model PROC MIXED model i SAS:

Indenfor den enkelte art er målingerne af de relative frekvenser af plantearten korrelerede således:

1. Indenfor lokalitet:
 - a. Målingerne i samme "parcel" er korrelerede:
 1. Indenfor ramme for botaniske analyser på marken,
 2. Vom- og fækalie prøve indenfor stud, og stud på mark,
 - b. Målingerne på samme mark (her defineret til sammenlægning af de 2 delmarker høj og lav intensitet) er korrelerede, givet krydsklassifikationen øst/vest med 1./2. års stude,
 - c. Målingerne indenfor samme aldersgruppe og blok er korrelerede (således "Læstenvej" "resten i Øst" hver har deres tilfældige niveau), givet øst/vest,
 - d. Målingerne i samme blok er korrelerede (øst/vest).
2. Indenfor år:
 - a. Begrænset til alder og blok er målingerne fra samme år korrelerede,
 - b. Begrænset til blok er målingerne fra samme år korrelerede,
 - c. Målingerne samme år er korrelerede.
3. Måned indenfor år:
 - a. Begrænset til år, alder og blok er målingerne fra samme måned korrelerede,
 - b. Begrænset til år og blok er målingerne fra samme måned korrelerede,
 - c. Målingerne samme måned er korrelerede indenfor år.

Omkring de systematiske effekter: Der ser ud til at være en selektion mellem arterne.

Selektionen afhænger af måned, alder og intensitet. Endvidere ser det ud til at sæsonvariationen i selektion afhænger af de to faktorer alder og intensitet enkeltvis, samt at selektionen afhænger af kombinationen alder og intensitet.

Sæsonvariationen i selektion ser ikke ud til at afhænge af kombinationen alder og intensitet.

Modelkontrol

- Varianshomogenitet: Det største problem forventes knyttet til antagelsen af at mikroskopiske analyser har samme residualvarians som de botaniske analyser. De botaniske analyser har en spredning, der er ca. det dobbelte af spredningen for de mikroskopiske analyser. Det øvrige af kontrollen er derfor delt efter botanisk og mikroskopisk analyse. Variansen ser ud til at være homogen mht. art, år, måned indenfor år og sted.
- Normalfordeling: Det ser ikke helt skævt ud.

Yderligere aspekter, som kunne inddrages i modellen, ud over heterogenitet i botanisk / mikroskopisk analyse:

1. Anvendelse af den "additive logratio transformation", $\log(p/r)$, (se Barceló-Vidal, 2001 og Billheimer et. al., 1998) i stedet for den version af en "centered", $\log(p/(1-p))$, vi har anvendt, dvs. divider med frekvensen for "resten" i stedet for "1 minus frekvensen" før logaritmetransformationen.
2. De 14 målinger af de enkelte arter til samme tidspunkt og sted er antaget uafhængige. Det er givet en fejl.
3. Asymmetri mellem botanisk analyse og mikroskopisk undersøgelse:
 - a. Gentagelsen vom- og fækalie-prøve ved mikroskopisk undersøgelse,
 - b. Det samme par af stude er anvendt på 2 marker, og nye par af stude andet år, mens rammerne ligger samme sted begge år, dog således at klippefeltet flyttes indenfor rammen ved hvert klip.
4. Estimat af selektionen afhængig af tidspunkt med henblik på sammenstilling med tallene for fordøjelighed.

Forklaring af selektionen af arter ved fordøjelighed af den enkelte art:

Selektionen af den enkelte art, estimeret ved denne metode 2, er plottet mod fordøjeligt organisk stof, råprotein og cellevægsindhold (NDF). Der er som den enkelte arts mål for f.eks. NDF taget gennemsnit over de 6 målinger af denne størrelse, 3 for hvert af de 2 år.

(Selektionen burde estimeres afhængighed af tidspunktet, så der kan tages hensyn til at fordøjeligheden af den enkelte art afhænger af tidspunkt.)

Der ses her ingen sammenhæng mellem selektionerne af de enkelte arter og målene for fordøjelighed.

Resultater, empiriske mål for selektioner (metode 1):

Der er sideløbende foretaget analyse af "de empiriske mål for selektioner", dvs. selektion estimeret direkte ved målingerne på samme delmark og til næsten samme tidspunkt. Denne

metode, kan betragtes som en analyse af studenes indtag, vurderet ud fra mikroskopisk undersøgelse, givet den observerede botaniske sammensætning på marken. (Hvis måleusikkerheden på botanisk analyse er lille, så behøver vi ikke tænke på botanisk analyse som en stokastisk variabel, som ved lineær regression.)

Ud fra "fænomenets natur" er det muligt at give bud på, hvorledes mængden af de enkelte arter (både ved botanisk analyse og mikroskopisk undersøgelse) er korrelerede over tid og sted. Det blev gjort i afsnittet "Lidt overvejelser omkring korrelationsstruktur". Det synes ikke så let at give bud på, hvorledes studenes selektion er korrelerede, over tid og sted, givet mængden af de enkelte arter.

P-værdierne for test af de enkelte effekter vil jeg ikke lægge stor vægt på ved metode 1, da disse afhænger meget af valg af kovariansstruktur, og dette valg ikke er analyseret nøjere. Estimerne af selektion af de enkelte arter ser dog konsistente ud med estimer fundet ved metode 2.

Modellen er kørt med henblik på at inddrage mål for fordøjelighed.

Tilføjer vi disse mål for fordøjelighed, ved siden af art som faktor, så ser vi om fordøjelighed kan forklare yderligere variation i materialet, ud over det art forklarer, altså ikke hvorfor den enkelte art selekteres.

Resultaterne omkring fordøjelighed er de samme for de 2 foreløbige modeller.

Resultat: Diæt og selektion

Med hensyn til studenes diæt var de relative frekvenser af plantearter identificeret i vom- og fækalieprøver ikke signifikant forskellige, og derfor er resultaterne vist i tabel 1 gennemsnit af de to prøvetyper.

Selektiviteten af de 12 hyppigste arter samt resten, grupperet som "rest", er vist i tabel 2. Som det kan ses af tabellen, var selektiviteten for Alm. Kvik, Manna Sødgræs, Alm. Hønsetarm og Fløjlsgræs positiv og signifikant med selektiviteten for Alm. Kvik næsten dobbelt så stor som for Fløjlsgræs. Selektiviteten for Mose-Bunke, Rød Svingel, Knæbøjet Rævehale og Eng-Rapgræs var signifikant negativ, medens der ikke var nogen selektivitet for siv, Eng-Rottehale, Alm. Rapgræs, Lav Ranunkel og de resterende arter.

Tabel 1. Optagelse af de forskellige arter, som den gennemsnitlige relative frekvens af planter fundet i vom- og fækalieprøver (n=382).

Art	Gennemsnit	Standardafvigelse
Manna-Sødgræs (<i>Glyceria fluitans</i>)	21,96	18,02
Siv (<i>Juncus spp.</i>)	17,86	20,31
Fløjlsgræs (<i>Holcus lanatus</i>)	16,01	20,47
Alm. Kvik (<i>Elytrigia repens</i>)	12,78	8,46
Eng-Rapgræs (<i>Poa pratensis</i>)	8,17	9,50
Alm. Rapgræs (<i>Poa trivialis</i>)	7,25	9,50
Mose-Bunke (<i>Deschampsia caespitosa</i>)	5,26	9,34
Lav Ranunkel (<i>Ranunculus repens</i>)	3,84	3,11
Knæbøjet Rævehale (<i>Alopecurus geniculatus</i>)	3,19	3,83
Eng-Rottehale (<i>Phleum pratense</i>)	0,94	1,60
Blød Hejre (<i>Bromus hordeaceus</i>)	0,84	1,43
Rød Svingel (<i>Festuca rubra</i>)	0,76	1,50
Kløver (<i>Trifolium sp.</i>)	0,71	1,65
Hare Star (<i>Carex leporine</i>)	0,58	1,07
Alm. Hønsetarm (<i>Cerastium fontanum</i>)	0,25	0,77
Kær-Tidsel (<i>Cirsium palustre</i>)	0,08	0,36
Stor Nælde (<i>Urtica dioica</i>)	0,06	0,48
Rørgræs (<i>Phalaris arundinacea</i>)	0,05	0,34
Gåsepotentil (<i>Potentilla anserina</i> ,)	0,05	0,30
Sump-Kællingtand (<i>L. uliginosus ssp. uliginosus</i>)	0,03	0,26
Mælkebøtte (<i>Taraxacum sp.</i>)	0,01	0,22
Alm. Syre (<i>Rumex acetosa</i>)	0,005	0,07

Tabel 2. Planteselektivitet hos fistulerede stude på Fussingø. (Se tekst for beregningsgrundlag).

	Estimat af Selektivitet	Standard Error	Signifikans
Alm. Kvik	1.32	0.20	***
Manna-Sødgræs	1.17	0.20	***
Alm. Hønsetarm	0.74	0.34	*
Fløjlsgræs	0.67	0.19	***
Mose-Bunke	-1.17	0.26	***
Rød Svingel	-0.77	0.33	*
Knæbøjet Rævehale	-0.75	0.23	***
Eng-Rapgræs	-0.68	0.21	***
Siv	0.15	0.21	NS
Eng-Rottehale	-0.45	0.30	NS
Alm. Rapgræs	0.05	0.18	NS
Lav Ranunkel	-0.22	0.18	NS
Rest	-0.05	0.20	NS

***=p<0.001, **=0.001≤p<0.01, *=0.01≤p<0.05

Konklusioner vedr. selektion

- De fire arter som studene selekterede positivt var alle naturarter, nemlig Alm. kvik, Manna-Sødgræs, Alm. Hønsetarm og Fløjlsgræs,
- Af de arter som studene selekterede negativt var der to kulturarter, Rød Svingel og Eng-Rapgræs, samt to naturarter, Mose-Bunke og Knæbøjet Rævehale,
- Studenes selektion var ikke knyttet til kvalitet for foderværdi i form af fordøjeligt organisk stof, råprotein eller cellevægsindhold (NDF).

Litteratur

- Barceló-Vidal, C., Martín-Fernández, J.A. and Pawłowsky-Glahn, 2001. 'Mathematical Foundations of Compositional Data Analysis', in Proceedings of the Annual Conference of the International Association for Mathematical Geology, Cancun (Mexico).
- Billheimer, D., Guttorp, P., Fagan, W.F. ,1998.: Statistical Analysis and Interpretation of Discrete Compositional Data. NRCSE Technical Report Series No. 011.
- Hald, A.B., 2002. Botanisk udvikling. DJF intern rapport 154, 34-39.
- Nielsen, L., Hansen, H.H., Badsberg, J.H. & Søgaard, K., 2002a. Planteproduktion og fødevalg. DJF intern rapport 154, 7-13.
- Nielsen, L., Hoffmann, C.C. & Thamsborg, S.M. (eds.), 2002b. Ferske enge – ekstensiv landbrugsproduktion, natur og miljø. DJF intern rapport nr. 154, 60 pp.

Vurdering af forsøgsudslag ved hierarkiske designs og gentagne målinger

Erik Jørgensen

Forskergruppe for Biometri, Afdeling for Husdyravl og Genetik, DJF

Abstract

Næsten alle undersøgelser, både planlagte forsøg og ikke-eksperimentelle undersøgelser, gennemføres med en form for hierarkisk struktur i indsamlingen af data (plante, parcel, mark, bedrift). Således er der ofte flere målinger på samme forsøgsenhed/parcel. Med sådanne forsøgsdesign skal man være ekstra varsom, så eventuelle behandlingsudslag vurderes i forhold til den rette variation. Heldigvis har fremkomsten af programpakker til analyse af såkaldte mixed models givet større sikkerhed for at undgå faldgruberne.

Nøgleord: Mixed models, hierarkiske design, korrelationsstruktur, split-plot.

Introduktion

I langt de fleste forskningsområder indenfor Danmarks Jordbrugsforskning foretages forsøgsopgørelser ved hjælp af lineære normale modeller. Traditionelt har PROC GLM været arbejdshesten, der er blevet trukket på. Set i lyset af hvilke design der oftest anvendes, må det imidlertid anbefales at skifte til PROC MIXED, der som navnet antyder kan anvendes til modeller hvor modellens effekter både indeholder tilfældige og systematiske effekter. PROC GLM har også faciliteter til håndtering af tilfældige effekter, men sat på spidsen, virker disse faciliteter kun korrekt, når der er tale om balancerede forsøg.

Når dette har betydning, skyldes det, at statistisk analyse bygger på en række forudsætninger. I tilfældet med lineære normale modeller er det.

1. at afvigelserne fra modellen (residualerne) skal være normalfordelte.
2. at variansen på afvigelserne fra modellen er konstant.
3. at afvigelserne skal være uafhængige.

Ofte vil disse forudsætninger ikke være 100 procent opfyldte, så forsøgsopgørelse er det muliges kunst !. Imidlertid er betydning af manglende opfyldelse af normalfordelingsantagelsen den mindst betydende. Selv med ret store afvigelser vil den statistiske analyse føre til samme konklusioner. Antagelsen om konstant varians er mere afgørende. Hvis der er variansinhomogenitet kan det for eksempel føre til, at nogle betydende forsøgsudslag ikke kan konstateres signifikante, fordi målestokken til sammenligningen (middelfejlen) ikke er korrekt.

Men det er forudsætningen om uafhængighed, der er langt den vigtigste! Manglende uafhængighed vil typisk føre til at man fejlagtigt konkluderer, at der er behandlingseffekt. En af de bedste måder til at opnå denne uafhængighed er at anvende randomisering. Det vil sige at tildele behandling til den enkelte forsøgsenhed/-parcel udfra et lotteri, enten ved hjælp af tilfældige tal, eller for eksempel ved brug af en terning. Ofte er der imidlertid hierarkiske niveauer i designet, med split-plot forsøg som det typiske eksempel. Sådanne hierarkier introducerer afhængighed mellem forsøgets observationer, og det er her at det er vigtigt at anvende statistiske modeller, der håndterer afhængigheden, for eksempel PROC MIXED i SAS.

I det følgende vil vi meget kort introducere brugen mest som en appetitvækker.

Eksempel på fejlrisikoen

For at illustrere betydningen af at ignorere afhængighed tages der udgangspunkt i et eksempel. Eksemplet er fra en helt anden verden, nemlig husdyrforskningen: Vi starter med et citat fra Hoekstra & Jansen, *Applied Animal Behaviour Science* 16 (1986) 303-308

... After reading 50 of these papers in AABS-issues [Applied Animal Behaviour Science] of 1984 and 1985, we found that in about 25 cases statistical methods were used incorrectly.

The main defect was that observations entered into test statistics were not independent.

In a number of cases it was totally unclear how the authors made their computations

Som illustration af problemstillingen valgte forfatterne et forsøg udført i forbindelse med den hollandske adfærdsforskers Willem Schoutens Ph.D. afhandling. *Rearing conditions and Behaviour in Pigs*

Forsøget gik ud på at undersøge om pattegrises tidlige erfaring påvirker deres senere adfærd, I dette tilfælde om opstaldningsformen lige efter fødsel (faring) påvirkede deres senere adfærd. Til at undersøge dette blev der udført et eksperiment med hhv "golde" farestier ($2 \times 2 \text{ m}^2$) og "berigede" (28 m^2 med tildeling af rigelig halmstrøelse).

I forsøget blev anvendt ialt 8 søer, som parvist var søskende, og indenfor søskendepar blev behandling tilfældigt valgt. Kuldstørrelsen blev umiddelbart efter fødslen standardiseret til 8 grise, dvs ialt indgik 64 grise i forsøget. Under forsøget blev udført detaljerede adfærdsobservationer.

Bemærk at randomiseringen foregik på søniveau, at alle grise gik i samme sti og derfor efter alt at dømme påvirkede hinanden, og at deres adfærd tillige var påvirket af den adfærd som soen udviste. Derfor forekommer en antagelse om, at observationer på de enkelte grise var uafhængige ikke særlig velunderbygget. Alligevel blev analysen af forsøget foretaget med grise som forsøgsenhed, med store konsekvenser for konklusionen som vist i tabel 1.

Tabel 1: Resultater fra forsøget. Dels som rapporteret, dels den korrekte model, vha. kuld gennemsnit

	Rapporteret Model			Kuld Gennemsnit		
Effekt	df	SS	F	df	SS	F
Søskendepar	3	384.2		3	48.0	
Faresti System	1	893.5	5.14*	1	117.0	2.424
Residual	59	10253.7		3	138.2	
Total	63			7		

Som det fremgår af tabellen blev den oprindelige opgørelse foretaget med den enkelte gris som forsøgsenhed, det vil sige uden at tage højde for den afhængighed mellem observationerne, som det fælles stimiljø bevirke. F-testen blev udført med (1,59) frihedsgrader og gav tydeligt signifikant udslag. Den korrekte analyse, hvor effekten af stimiljø vurderes på baggrund af de uafhængige observationer fra hver enkelt faresti, var derimod langt fra at vise signifikans. F-testen blev i dette tilfælde udført med (1,3) frihedsgrader. I stedet for at anvende kuld gennemsnit ved opgørelsen kunne programsætningerne i Tabel 2 have været anvendt med samme resultat for behandlingssammenligningen. Effekten af farestisystem vil blive vurderet mod variationen mellem søer inden for søskendepar.

Tabel 2: Program til korrekt analyse af soforsøget med hhv. PROC GLM og PROC MIXED

PROC GLM ;	PROC MIXED ;
CLASS PAR SO SYSTEM ;	CLASS PAR SO SYSTEM ;
MODEL Y = SYSTEM PAR PAR*SO ;	MODEL Y = SYSTEM ;
RANDOM PAR PAR*SO / TEST ;	RANDOM PAR PAR*SYSTEM ;

(I tabellen er vist både GLM og MIXED. I det balancerede tilfælde vil de to analyser give identiske resultater. I det ubalancerede tilfælde vil kun PROC MIXED give de korrekte resultater.)

Eksemplet illustrerer problemet med falsk signifikante resultater når afhængigheden mellem observationer ignoreres, og forsøgsudslaget derved bliver vurderet med den forkerte målestok. Det er vigtigt at gøre sig klart, hvad falske signifikante resultater indebærer af besvær og omkostninger. På daværende tidspunkt var jeg mere direkte involveret i forskningen vedrørende grises adfærd, og kan tydeligt huske den interesse som Willems Schoutens resultater gav. Det var adskillige forsøg der blev iværksat på den baggrund, både i Danmark og i udlandet. Efterhånden som de nye forsøg ikke kunne eftervise de oprindelige resultater, dalede interessen naturligvis.

Planteforsøg

Selvfølgelig eksisterer denne problemstilling ikke kun for adfærdsforsøg med grisesøer. Grundlæggende er designet et såkaldt split-plot forsøg og det er jo velkendt indenfor plante-produktion.

Men det er værd at være opmærksom på, at problemstillingen i bund og grund er relateret til dataindsamlingsproceduren, og at problemet er blevet mere akut jo billigere prøveindsamlingen er blevet.

Lad os derfor se på eksempler fra planteproduktion.

1. Et forsøg med to behandlinger udlægges på 4 marker. Hver mark opdeles i to forsøgspareller og hver parcel tildeles tilfældigt een af de to behandlinger. Fra hver forsøgsparelle udvælges 8 planter, som analyseres for svampeangreb.
2. Et forsøg med to behandlinger udlægges på 4 marker. Hver mark opdeles i to forsøgspareller og hver parcel tildeles tilfældigt een af de to behandlinger. Udviklingsstadiet fastlægges på 8 forskellige tidspunkter gennem vækstsæsonen som gennemsnit af en stikprøve på 16 planter.
3. Et forsøg med to behandlinger udlægges på 4 marker. Hver mark opdeles i to forsøgspareller og hver parcel tildeles tilfældigt een af de to behandlinger. Fra hver forsøgsparelle tages 8 jordprøver, hvor N-indhold i 5 forskellige dybder analyseres.

Disse tre eksempler vil danne udgangspunkt for illustration af brugen af PROC MIXED.

Figur 1: Eksempel på udlægning af forsøgspareller på 2 marker med to behandlinger A,B og 8 målepunkter per parcel, som illustration af eksempel 3. En grænse mellem 2 jordbundstyper er skitseret vha. det grå område

Forsøgsdesign skal afspejles i model

Første punkt er at sikre sig at forsøgsdesignet afspejles i den statistiske model. Derved opnås, at den hierarkiske struktur kan indgå korrekt i beregningerne.

Typiske faktorer at huske på er:

- Valg af diverse blokinddelinger
- Valg af forsøgsenhed for behandling
- Valg af tidspunkt for observation
- Valg af position for prøveudtagning
- Valg af stratificering. (eks. Stikprøveudtagning med store planter vs. små planter)

I de tre eksempler er der klart et niveau, der hedder mark, som er den overordnede blokfaktor, mens $\text{parcel} = \text{mark} \times \text{behandl}$ er den egentlige forsøgsenhed.

Eksempel 1: Observationer fra de 8 planter indenfor hver parcel er uafhængige og der er ikke yderligere design struktur. Det fører til følgende modelspecifikation, fuldstændig analogt til eksemplet med soforsøget.

```
MODEL Y = Behand ;  
RANDOM Mark Mark*Behand ;
```

Eksempel 2: I dette tilfælde er de 8 tidspunkter en yderligere design faktor, hvor man typisk vil være interesseret i den systematiske effekt af såvel tidspunkt som eventuel vekselvirkning mellem behandling og tidspunkt. Dette kan opnås ved følgende udvidelse i modellen.

```
MODEL Y = Behand TID TID*Behand ;  
RANDOM Mark Mark*Behand ;
```

Bemærk, at man i stedet for at beregne gennemsnit af de 16 planter for hvert tidspunkt kunne have valgt at analysere på basis af enkeltobservationer. Men det ville så have krævet en modeludvidelse, så afhængigheden mellem de 16 målinger fra samme parcel og samme tidspunkt kunne inddrages.

```
MODEL Y = Behand TID TID*Behand ;  
RANDOM Mark Mark*behand mark*behand*TID ;
```

Hvis forsøget er balanceret med 16 observationer fra hvert tidspunkt vil opgørelsen med gennemsnit give samme vurdering af behandlingseffekter. Effekten af TID vil således blive vurderet ved hjælp af variationen mellem $\text{mark} * \text{behand} * \text{TID}$, hvilket svarer til variationen i gennemsnittene. Beregningsmæssigt vil modellen med gennemsnit være langt enklere og må derfor normalt foretrækkes.

Eksempel 3: De 8 positioner indenfor hver parcel vil normalt kunne antages at være uafhængige givet parcellens gennemsnitlige niveau. I lighed med eksempel 2 er der imidlertid endnu et niveau i designet, nemlig de 5 målinger i samme jordprøve. Her er der en syste-

matisk effekt, dybde, hvor igen både hovedvirkning og vekselvirkning med behandling har interesse.

```
MODEL Y = Behand Dybde Dybde*Behandl ;  
RANDOM Mark Mark*behandl mark*behandl*Position ;
```

Effekten af dybde bliver vurderet mod residualvariationen fordi effekten af dybde er bestemt for hver enkelt jordprøve inden for hver parcel.

Faldgruber

Som det fremgår af ovenstående kan anvendelsen af PROC MIXED være rimelig enkel. Næste punkt er at sikre sig, at SAS forstår forsøgsdesignet. Desværre er der en oplagt fejlmulighed her.

I PROC MIXED er der flere muligheder for at angive, hvilke frihedsgrader, der skal anvendes ved test af de systematiske effekter. Sandsynlighedsfordelingen for F-værdien vil ved balancerede forsøg følge en eksakt F-fordeling med kendte frihedsgrader, når nulhypotesen er sand. Ved ubalancerede forsøg gælder dette ikke længere, og F-testen bliver derfor en tilnærmet test. Tilnærmelsen kan ske på flere måder. Valget sker ved hjælp af en option, DDFM=metode, til MODEL sætningen.

Hvis der ikke angives en DDFM option vælges som standard den såkaldte *containment* metode. Denne metode minder i stor udstrækning om de frihedsgradsberegninger vi er vant til for eksempel ved brug af PROC GLM. Men der er en væsentlig ændring i forhold til GLM. Beregningerne foretages *alene* på baggrund af effekterne i MODEL og RANDOM sætningerne, og *ikke* på grundlag af forsøgsdesignet, som det fremgår af data,

Hermed opstår en yderst alvorlig fejlmulighed i opgørelsen!!!. Tilsyneladende uskyldige fejl i modelspecifikationen kan føre til grove fejl i opgørelsen. Fejl som ikke bliver bemærket af PROC MIXED !!!!

Effekten er, at forsøgsdesignet bliver mistolket, og forkerte frihedsgrader anvendt ved testene.

Et eksempel er vist i Tabel 3. Selv om effekterne mark*behand og mark*parcel entydigt identificerer den enkelte parcel, kan containment-metoden kun i det førstnævnte tilfælde se at behandling varierer inden for mark, men ikke inden for parcel.

Tabel 3: Program der giver korrekt hhv forkert frihedsgradsberegning ved brug af standard metoden i PROC MIXED

Korrekt	Forkert
PROC MIXED ; CLASS BEHAND MARK ; MODEL Y = Behand ; RANDOM MARK Behand*MARK ;	PROC MIXED ; CLASS BEHAND MARK PARCEL ; MODEL Y = Behand ; RANDOM MARK PARCEL*MARK ;

For at illustrere problemstillingen ud fra husdyreksemplet. Hvis Willem Schouten havde anvendt PROC MIXED i god tro med containmentmetoden, med samme random specifikation som i PROC GLM, ville han have endt med begå samme fejl i opgørelsen, som han oprindeligt gjorde ! (Rent faktisk var det først, da jeg skulle beskrive dette eksempel, at jeg konstaterede, at jeg havde gjort samme fejl i formuleringen af tabel 2, men nu skulle den være rigtig).

På grund af disse fejlmuligheder kan det *ikke* anbefales at benytte standard metoden til at beregne frihedsgrader. I stedet kan anbefales at anvende *Satterthwaite's* metode, som vælges ved hjælp af:

```
MODEL Y = Behand/DDFM=SATTERTH ;
```

Satterthwaite metoden tager udgangspunkt i forsøgsdesignet, som det fremgår af data, og er derfor ikke så følsom for modelspecifikationen, som containmentmetoden. Der følger en ekstra fordel med. Som nævnt er F-testene ikke eksakte ved ubalancerede forsøg.

Satterthwaite's metode tilnærmer til de korrekte test-størrelser ved at ændre på frihedsgraderne i testene. (Dette medfører at frihedsgraderne ikke længere er heltallige, og at frihedsgraderne kan være forskellige for forskellige afhængige variable.)

Udvidet modellering af afhængighed mellem observationer

PROC MIXED giver mulighed for en mere detaljeret modellering af afhængigheden mellem observationer. Her skal blot illustreres nogle af mulighederne.

I eksempel 2 med målinger over tid er det ikke sikker at den beskrevne model er den optimale. Modellen indebærer at alle målinger på samme parcel er lige korrelerede. En anden model ville være at afhængigheden mellem to målinger på parcellen ville være afhænge af hvor lang tid, der var forløbet mellem målingerne. En måde at håndtere dette på, er at anvende såkaldte gentagne målinger.

```
MODEL Y = Behand TID TID*Behand ;  
RANDOM Mark ;  
REPEATED TID /subject=Mark*Behand TYPE=UN ;
```

Vi har nu reduceret random sætningen til kun at indeholde mark-effekten, mens modellen for korrelationen mellem målingerne på parcellen nu håndteres ved hjælp af repeated sætningen. Repeated sætningen indeholder tre elementer. En indikator variabel TID, som angiver rækkefølgen af de gentagne observationer. En subject specifikation, der fortæller på, hvilken design-enhed de gentagne målinger foretages, i dette tilfælde på parcellniveauet. Og endelig, en type specifikation, som angiver, hvilken modeltype for korrelationen, der skal vælges. UN står for unstruktureret som er den mest fleksible, idet den simpelthen estimerer hele kovariansmatricen for de 8 observationer over tid (hvilket kræver hele 28 parametre). Der findes en

lang række typer for korrelationsstrukturer. Med afhængighed over tid kunne man overveje $TYPE=AR(1)$, som er en simpel autoregressiv proces,

I designet i eksempel 3, som illustreret i Figur 1 kunne en naturlig model for korrelationsstrukturen være at korrelationen var afhængig af afstanden mellem to målepunkter, jo længere afstand, jo lavere korrelation. Også dette vil kunne håndteres. Der findes mulighed for en række spatiale korrelationsstrukturer (eksempel $TYPE=SP(EXP)$) som udfra positionskoordinater beregner korrelationerne.

Før man begynder på de mere komplicerede modeller for korrelationsstrukturer skal man være forberedt på problemer. Dels er data ofte af sådan beskaffenhed, at det er svært at skelne mellem forskellige korrelationsmodeller. Dels er der en risiko for ende med en for kompliceret model. I de nævnte eksempler, er det forholdsvis tynde data materiale ikke tilstrækkeligt til at gå i dybden med modelleringerne. Og endelig vil de mere komplicerede modeller ofte føre til beregningsmæssige vanskeligheder med at finde løsninger. Dette afspejler sig både i beregningstid og i en risiko for at der ikke kan foretages en estimation.

Det er vigtigt at bevare fokus på behandlingssammenligningerne. For eksempel vil systematiske ændringer, som de to jordbundstyper i Figur 1 påvirke korrelationen mellem to målepunkter. Typisk vil man også kunne forvente, at behandlingerne påvirker korrelationsstrukturerne. Sådanne ændringer er svære at modellere i et standard setup som PROC MIXED. Og i sidste instans kræves der mange uafhængige gentagelser af behandlingerne (parceller i de eksempler vi har diskuteret) for at vurdere de forskellige strukturer.

Diskussion

Med tilgængeligheden af programpakker som PROC MIXED til analyse af hierarkiske design skulle man kunne undgå problemer ved vurderingen af om forsøgsudslag er signifikante. Den hyppige fejl med at ignorere afhængighed mellem observationer vil derved kunne undgås, hvis man er omhyggelig med at forsøgsdesignet er afspejlet i den statistiske model.

Samtidig er der mulighed for en mere detaljeret modellering af afhængigheden mellem observationerne, for eksempel ved at inddrage afstand i tid eller rum. En sådan modellering indebærer mulighed for en mere effektiv udnyttelse af de indsamlede data.

I indlægget er anvendt PROC MIXED som udgangspunkt. Dette er blot een blandt mange muligheder. Forskellige programpakker har forskellige styrker og svagheder. For komplicerede problemstillinger kan det derfor være en fordel at vælge en programpakke der er mere skræddersyet end PROC MIXED.

En afsluttende kommentar. Mange problemer ved analyse af hierarkiske design kan klares ved blot at beregne gennemsnit på de laveste niveauer i hierarkiet. Derfor, gør det til en vane at starte med at spørge: Er der nogen grund til *IKKE* at beregne gennemsnit?

Modern chemometrical methods in relation to agricultural sciences – is it the light at the end of the tunnel?

Søren Højsgaard

Forskergruppe for Biometri, Afd. for Husdyravl og Genetik, DJF

Introduction

The mathematical tools on which chemometrics is based have a long story with some methods dating back to the early years of the 20th century. The use of these tools have become more widespread in the agricultural sciences and food sciences over the past two decades and for good reasons: (1) Measurement of multivariate data becomes more and more common due to modern sensory and measurement techniques and (2) Windows based software, allegedly user friendly, together with modern cheap computers provide the frame for working with these data.

Traditional (mathematical) statistics has to some extent developed into a deductive science. Chemometrical methods, on the other hand, is much more exploratory in its nature. Both approaches certainly have their place in applied statistics, but the choice of method clearly depends on the questions asked in a particular context. In this connection there are two (not always very clear) distinctions to make: whether the purpose is (1) prediction of future observations or (2) understanding the biological mechanisms generating the data. In the first case chemometrical methods often work very well, whereas they may not be the best choice in the latter.

In this paper we first illustrate a few chemometrical methods and outline (1) when they solve problems which can not directly be dealt with using “classical statistical methods” and (2) sketch how the results are interpreted. Then afterwards, we discuss the shortcomings of these methods, and try to point towards new directions of development.

Example: Heat development when concrete hardens

Table 1 shows measurements of the heat development (y) during hardening and the chemical composition of concrete. In a statistical analysis one might investigate the significance of the x s in explaining the strength y .

Table 1. Heat development (y) (calories per gram concrete) during hardening and contents of

$x_1 = 3CaO \cdot Al_2O_3$, $x_2 = 3CaO \cdot SiO_2$, $x_3 = 4CaO \cdot Al_2O_3 \cdot Fe_2O_3$, and $x_4 = 2CaO \cdot SiO_2$ in %.

y	x1	x2	x3	x4	Σ	y	x1	x2	x3	x4	Σ	y	x1	x2	x3	x4	Σ
78.5	7	26	6	60	99	109.2	11	55	9	22	97	115.9	21	47	4	26	98
74.3	1	29	15	52	97	102.7	3	71	17	6	97	83.8	1	40	23	34	98
104.3	11	56	8	20	95	72.5	1	31	22	44	98	113.3	11	66	9	12	98
87.6	11	31	8	47	97	93.1	2	54	18	22	96	109.4	10	68	8	12	98
95.9	7	52	6	33	98												

the x s: The “sum” column in the table contains the sum of the x measurements for each sample, and the sum is nearly 100 (%). Hence if three of the x ’s are known then so is the fourth (approximately) too.

A tempting starting point is a multiple linear regression model, i.e.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i.$$

Important aspects of a statistical analysis are (1) estimation of the regression parameters, (2) drawing inference about the significance of explanatory variables and (3) making predictions of future cases. The collinearity poses several problems in this connection:

1. **Estimation:** A classical way of estimating the regression parameters is by ordinary least squares (OLS) where the estimates of $\beta_0, \dots, \beta_4$ are obtained by minimizing the sum of squares $\sum_i (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}))^2$.

In this example, the parameters can be estimated this way, but the estimates are not very good in the following sense: The variance of the estimates is very high meaning that by pure randomness, the parameters are likely to be far away from their true value.

2. **Prediction:** If the goal is prediction, then the collinearity also poses a problem. For the reasons mentioned above, the parameters may be poorly determined, meaning in practice that the predictions are not necessarily very good.
3. **Significance:** The collinearity has the following consequence:
 - If the significance of the regression coefficients $\beta_1, \dots, \beta_4$ is tested, then one finds that none of them are significant. It turns out, that if one of the x ’s are deleted from the model, then the remaining three regression parameters become significant – and that is irrespective of which of the x ’s that one removes.
 - This coupling of the x s (they sum to approximately 100) suggests that it does not necessarily make much sense to ask the question whether one of the x ’s are irrelevant for explaining y

Example: Prediction from IR data

From 17 milk samples, the contents of the responses (y)s Fat, Protein, Lactose and Dry matter are measured. Moreover, 150 IR measurements (explanatory variables, x s) are made at wavelengths from 963 to 1542 cm^{-1} . The IR measurements are plotted against wavelength in Figure 1. From the figure it follows, that collinearity is also an issue for these observations: Similar wavelengths tend to give similar measurements. Plotting the correlation between the y s and the x s as in Figure 2 shows that there are groups of wavelengths strongly correlated to each response, but no single (distinct) wavelengths. Thus we are in a more extreme situation than in Section 2: From a practical point of view, it is not even possible to use OLS to estimate the 150+1 regression parameters from only 17 observations (the 1 comes from the intercept).

Figure 1. IR measurements at different wavelengths for 17 milk samples

Figure 2. Correlation between the responses that the IR measurements plotted against wave number.

PCA, PCR and PLS

We now describe very briefly three classical chemometrical methods: Principal component analysis (PCA), principal component regression (PCR) and partial least squares regression (PLS).

Principal component analysis

Principal component analysis (PCA) can be illustrated by plotting the e.g. the first and the fifth wavelengths against each other as in Figure 3, left. Most variation is in the direction of the bold line. Thus if the data points are projected onto that line, then these new points on the line tells nearly the same story as the original data do. Yet, an important dimension reduction has taken place: The original data points are 2-dimensional, whereas the line is 1-dimensional. The projection corresponds representing data in the new coordinate system is given by the bold and the dashed line in the figure. The original data points are denoted by (x_1, x_2) and the new ones (in the new coordinate system) by (x'_1, x'_2) . If the angle of the bold line is say $\theta = 40^\circ$ then classical geometry gives that

$$\begin{aligned}x'_1 &= \cos\theta \times x_1 + \sin\theta \times x_2 = 0.76 \times x_1 + 0.64 \times x_2 \\x'_2 &= -\sin\theta \times x_1 + \cos\theta \times x_2 = -0.64 \times x_1 + 0.76 \times x_2\end{aligned}$$

So the new variables are weighted sums of the old ones. The x_1 s are the first principal component (PC1), and the x_2 s are the second, PC2. In Figure 3, right, PC2 is plotted against PC1, and by considering the scale, it is clear that PC1 explains far more variation than PC2. More generally, the principal components are obtained as weighted sums of the original data, i.e.

$$s_{i1} = x_{i1}w_{11} + \dots + x_{ip}w_{1p}$$

$$s_{i2} = x_{i1}w_{21} + \dots + x_{ip}w_{2p}$$

$$s_{i3} = \dots$$

and so on. The values s_{il} constitute the l th principal component (also called the l th *score*) and the coefficients $w_{1l}, \dots, w_{pl}$ are called the *loadings* for that component. The coefficients can be chosen such that (1) the principal components are uncorrelated and (2) the first component explains most variation in the data, the second component explains second most variation and so on. Often it is so that the first 2 or 3 components explains a high proportion of the total variation (say 70-80%) and one would then take those components as representing a dimension reduction of the data. The next step is to try to interpret these components. This is illustrated later.

Figure 3. Illustration of PCA. Left: The first and fifth IR measurement. Right: The principal components PC1 and PC2 obtained by rotating the original data.

Principal component regression

Principal component regression is based on the following idea:

1. First extract (usually a few) principal components.
2. Next use the components, $s_1, \dots, s_R$ as explanatory variables in a multiple regression with y as response.

What is gained from using this method compared with classical multiple regression? First, for the milk data multiple regression is not an option. Secondly, it is often the case, that PCR regression will produce far more precise predictions than a regression based on the original data.

Figure 4. Left: Observed versus predicted values (based on cross-validation). Right: The regression coefficients for each wavelength.

The number of components to extract is usually determined by *cross validation* methods. For the milk data, it turns out that 3 components is optimal. A plot of the observed versus predicted values and of the regression coefficients are shown in Figure 4.

Partial least squares regression

Partial least squares (regression) – PLS – is very similar in spirit to PCR regression. The difference essentially lies in how the components are chosen: With PLS, the response y is involved when deriving the components. In practice, PLS and PCR regression often give very similar results, with no method being generally superior to the other.

Example: Children's choice of apples

In this example, 146 children of both genders and from 3 different ages (years in school) have tasted 6 different varieties of apples and have graded them on a scale from 1 to 5 to reflecting how much they like the apples. A taste panel with 10 members have also evaluated the apples with respect to various criteria such as taste, firmness, crispness etc. The contents of e.g. malic acid is also measured.

Here we focus only on the preferences of the children. First we calculate the first principal components, and plot PC2 versus PC1 as in Figure 5, left, which is called a *score plot*. The picture is then that the apple varieties separate out in two groups on PC1, and that may be interpreted as if there are two qualitatively distinct groups of apples. (The Granny Smiths appear towards the right in the plot).

The next thing is to plot the loadings for PC2 against the loadings for PC1, Figure 5, right, called a *loadings plot*. The loadings are specific to each child, and therefore we may label the points with e.g. the age or level of the child. There does not appear to be any pattern in the plot in the sense, that there is no apparent grouping of the children according to sex or age.

The interpretation of the plots in Figure 5 must necessarily be based on subjective judgments of the form “there seems to be several groups”. Yet what this method does not allow for is to answer questions of the form “is there a statistically significant difference between boys and girls with respect to variety preference”. Although perhaps theoretically questionable, it could be tempting in practice to continue the by analyzing the scores and loadings using more classical statistical methods.

Figure 5. Left: Score plot. Right: Loadings plot. The numbers indicate sex, and the different colours the age of the children.

Some (classical) problems where the projection methods may not be appropriate

In many areas of agricultural research there is a hierarchical structure to data. One example is split-plot experiments, where e.g. one treatment, say harvesting time is applied to large plots (whole-plots) and where another treatment, say fertilizer, is applied to sub-plots within the large plots. Suppose a high dimensional response of yield, sugar content, resistance etc. is measured. It is well known, that the effect of harvesting time should be judged against the whole plot variation, otherwise one typically find too significant results. Such a problem is not easily dealt with by chemometrical methods.

Along similar lines suppose the interest is in studying heritability. In that case the genetic variation needs to be separated from total variation. Likewise, the genetic correlation may be of interest, e.g. to investigate whether the relation between yield and fungi resistance is hereditary. Such a problem is also not easily dealt with by chemometrical methods.

Software for chemometrical analyses

There is a variety of tools facilitating chemometrical methods. Free tools are provided by the `pls.pcr` package and the `prcomp` functions in the R statistical package, see www.r-project.org. A very popular program for chemometrics is the (commercial) Unscrambler program, www.camo.no which is tailor made for chemometrical analyses. SAS (also commercial) provides some methods, especially `proc pls` and `proc princomp`, see www.sas.com.

Discussion

We have illustrated two classical chemometrical tools, PCR and PCA. From a prediction perspective, PCR (and PLS) are often found to work well in practice. Moreover, we have also seen that looking at the scores (and loadings) may help in defining clusters in the data, i.e. they may be helpful in trying to understand the mechanisms behind the data. Yet, these methods do not provide methods for asking classical questions about the significance of various effects. Neither do they (directly) provide measures of precision of estimated effects.

Acknowledgements

Anette Kistrup Thybo, Danish Institute of Agricultural Sciences is acknowledged for providing the apple data, and Carsten Ridder, Lattec I/S for providing the IR data.

Modellering af QTL-effekter over tid samt deres vekselvirkning med miljøfaktorer øger statistisk styrke og biologisk forståelse

Mogens Sandø Lund

Sektion for Bioinformation og Biometrisk Genetik, Afd. for Husdyravl og Genetik, DJF

Sammendrag

I de seneste år har der været en betydelig forskningsindsats for at kortlægge kromosomområder med gener der påvirker kvantitative egenskaber (QTL). Formålet er dels at øges vores viden om den molekylærgenetiske baggrund for fænotypisk variation og dels at muliggøre udvikling af DNA-tests, som kan udnyttes i avls forædling ved Markør Integreret Selektion. I QTL-studier reduceres gentagne målinger over tid oftest til en enkelt egenskab. Et eksempel er mælkeydelse hos malkekvæg. Igennem laktationen gennemgår dyrene imidlertid store fysiologiske forandringer. Det må derfor forventes at effekten af det identificerede QTL ændres med forandringer i andre proteiner, hormoner osv. over tid. Hos planter er der stor evidens for, at effekten af identificerede QTL afhænger af det miljø de udtrykkes i. Ofte vil denne interaktion kunne kvantificeres ved sammenhæng imellem QTL-effekterne og forskellige kvantitative miljøfaktorer som temperatur, nedbør, jordbundsforhold osv.

I denne artikel foreslås en model, som modellerer effekten af QTL og polygene effekter som en funktion af tid eller anden kvantificerbar miljøfaktor. Vi viser ved analyse af simulerede data og feltdata at denne model har to fordele i forhold til de traditionelle modeller. For det første vil man øge sandsynligheden for at identificere QTL'er. I et af de simulerede scenarier øgedes sandsynligheden for at identificere et QTL fra 6 % til 98% og i analyse af feltdata identificeredes fire QTL i forhold til et QTL med en traditionel model. For det andet vil modellens parametre give biologisk information af det enkelte QTLs effekt over tid eller interaktion med forskellige miljøfaktorer.

Keywords: QTL, gentagne målinger, genotype miljø vekselvirkning

Introduction

In QTL mapping studies, traits have often been defined as one record even though it is a function of several measurements recorded over a time period. An example is the 305 days milk yield, which is a weighted sum of a number of measurements recorded over the lactation. Using the sum may be reasonable if the genetic influence is constant over time. However, studies have shown that correlations between measurements differ over the lactation (e.g. Meyer et al. 1989; Pander et al. 1992; Swalve 1995; Jakobsen, et al. 2000). One

reason could be that physiological state of animals change greatly over the lactation, such that different genes are limiting factors, and therefore, are particularly important in certain periods of the lactation. Therefore, it can be important to model the different genotypes' deviation from a fixed curve over time. This will provide important biological information on the expression pattern/effect over time, and thereby assess the relevance for selection purposes. Furthermore, the statistical power to detect QTL is expected to be higher compared to a traditional univariate analysis, particularly if the effect of a QTL varies over time.

A model that allows the covariance between measurements to change gradually over time can be defined using covariance functions. A linkage model that is particularly suited in this framework is the random QTL model (Fernando and Grossman 1989; Grignola et al. 1996) because it models the (co)variances between QTL allelic effects. These random allelic effects can be regressed on the curve parameters in the same way as additive polygenic effects, such that curve shape depends on QTL genotypes.

The objectives of this study were to (i) develop a linkage analysis method for Random Regression models, (ii) assess the ability of the approach to fit different patterns of QTL effects over time, (iii) assess the power to identify a QTL compared to a traditional univariate method.

Material and Methods

Random regression model

The random regression mixed model including QTL effects can be written as:

$$\mathbf{y} = \boldsymbol{\beta}(\mathbf{t}) + \mathbf{u}(\mathbf{t}) + \sum \mathbf{q}_i(\mathbf{t}) + \mathbf{p}(\mathbf{t}) + \mathbf{e}, \quad (1)$$

where $\mathbf{y}$ is a records at different time points or values of environment the QTL interact with $(\mathbf{t})$, $\boldsymbol{\beta}$ are the fixed effects including the fixed mean curve of the population, $\mathbf{u}(\mathbf{t})$, $\mathbf{q}(\mathbf{t})$, and $\mathbf{p}(\mathbf{t})$ are random gaussian processes with mean zero and covariances between any two timepoints t and s is described by the covariance functions $U(t,s)$, $Q(t,s)$, and $P(t,s)$. The covariance functions can be modelled in many ways. Here we will use Legendre polynomials.

The random regression coefficients $\mathbf{p}$, $\mathbf{u}$, and $\mathbf{q}$, and $\mathbf{e}$ are assumed to be multivariate normal distributed as: $\mathbf{p} \sim \text{MVN}(0, \mathbf{P}_0 \otimes \mathbf{I})$, $\mathbf{u} \sim \text{MVN}(0, \mathbf{G}_0 \otimes \mathbf{A})$, $\mathbf{q}_i | \mathbf{M}, \mathbf{p}_i \sim \text{MVN}(0, \mathbf{K}_{0i} \otimes \mathbf{Q}_{i|M, \mathbf{p}_i})$, and $\mathbf{e} \sim \text{MVN}(0, \mathbf{I} \sigma_e^2)$, where $\mathbf{P}_0$, $\mathbf{G}_0$, and $\mathbf{K}_{0i}$ are variance-covariance matrices among random regression coefficients. $\mathbf{A}$ is the additive genetic relationship matrix and $\mathbf{Q}_{i|M, \mathbf{p}_i}$ is the gametic relationship matrix of the allelic effects at the i 'th QTL conditional on marker data ($\mathbf{M}$) and the position ($\mathbf{p}_i$) on the chromosome. The gametic relationship matrix was calculated by the recursive algorithm in Wang et al., 1995.

Simulation of data

The simulated pedigree was based on a small granddaughter design consisting of 20 unrelated grandsires each having 20 sons (referred to as sires). The linkage map consisted of 11 biallelic marker loci with 10 cM between each locus. A biallelic QTL was positioned in the midpoint between the third and fourth marker. In all loci, allele frequencies were assumed to be 0.5. For each sire, daughter yield deviations (DYDs) were calculated at 55 time points. DYDs were based on 100 daughters, which each had 11 test-day records with 30-day intervals. Of the 100 daughters, 20 had their first test-day on days 5, 10, 15, 20, and 25. The DYDs for sire i on day j (DYD_{ij}) were simulated using the following model that assumes the fixed curve to be constant over time: $DYD_{ij} = 1/20 \left(\sum_{l=1}^{20} \frac{1}{2} a_{li} \times LP_{lj} + m_{li} LP_{lj} + w_{ijk} \times f(j) + e_{ijk} \right)$. The terms $a_i \times LP_j$ and $m_i \times LP_j$ are the polygenic effects (deviation from the fixed curve) of sire i and the mendelian effect of daughter l at time j . The vectors $\mathbf{a}_i = (a_{0i}, a_{1i}, a_{2i}, a_{3i})$ and $\mathbf{m}_i = (m_{0i}, m_{1i}, m_{2i}, m_{3i})$ include polygenic regression coefficients, and $LP_j = (LP_{0j}, LP_{1j}, LP_{2j}, LP_{3j})$ the coefficients of a normalised fourth order Legendre polynomial at time j . The term $w_{ijk} \times f(j)$ is the effect of the paternally inherited QTL allele of daughter l at time j . The function $f(j)$ describes the allelic effect over time.

Three different scenarios were simulated which differed in the pattern of the QTL effect over time. In the first scenario, the QTL effect was constant over time (Figure 1 a). In the second scenario, an initially large effect declined gradually, and in the last half of the time period the effect was minimal (Figure 1b). In the third scenario, the effect of the initially positive allele declines gradually, and becomes negative in the last half of the time period, while the initially negative allele becomes positive (Figure 1 c).

Each scenario was replicated 100 times.

Figure 1. Effect of the two alleles in the biallelic QTL in the three scenarios. Effects are expressed as deviations from the fixed curve over time (days)

Figure 2a shows the QTL variance in the three scenarios. Polygenic and residual variances were the same for all scenarios, and are shown in Figure 2b.

Figure 2. Variance around the fixed curve over time (days) in the three simulated scenarios due to the QTL (a) and residual and polygenic variances (b).

Protein yield in dairy cattle

Based on all data in the Danish Cattle database two types of DYD's for Danish Holstein bulls were produced. The first were based on 305 days records and produced one DYD per sire. The second were based on test day records and produced time dependent DYDs in 10 day intervals. The current analyses were based on a granddaughter design with 19 grandsires and 1394 sons of the breed Danish Holstein. Seven chromosomes (BTA 9, 10, 14, 20, 23, 26, and 27) were scanned with the use of 4-12 markers per chromosome.

Data analysis

Data analyses included the field data for protein yield and each of 100 replicates of the three simulated scenarios. In each analysis two likelihood ratio tests were performed to test if the QTL could be identified using a Random Regression model and a traditional univariate model. Under both models, the marker haplotypes were assumed known for grandsires, when the gametic-relationship matrix was calculated. The restricted log likelihoods were maximised using an Average Information REML procedure (Jensen et al. 1997).

Results

Power: The statistical power of the tests was calculated as the proportion of the tests that were significant of the 100 simulated replicates within each scenario and model type (Table 1). When the QTL effect was constant over time the power of the two models were comparable. However, in scenario 2 and 3, where the QTL effect varies over time the difference were substantial.

Table 1. Statistical power of tests over different scenarios and models.

	Scenario 1	Scenario 2	Scenario3
Univariate	62	16	6
Random Regression	67	95	98

In the analysis of field data models with different number of parameters in the polynomial for QTL effects were fitted (Table 2). At the chromosomal position with the highest likelihood the full model was reduced to lower fits of the QTL effects. One chromosome reached significance with a one parameter Legendre polynomial, four with a two parameter polynomial but none with more than two parameters. This is due to the many parameters in the model. Compared to traditional 305-day models the use of random regression models identified more QTL (4 vs. 1).

Table 2. Likelihood ratio test statistics of different models (305 days model and random regression models with 1-4 parameter Legendre polynomials) at the maximum of each chromosome

BTA	305 d	RR1	RR2	RR3	RR4
9	0.9	0.0	0.6	9.3	12.4
10	1.1	0.1	8.6*	8.8	10.4
14	3.9*	7.9**	8.4*	nc	7.2
20	0.2	1.1	9.8*	10.0	10.2
23	0.1	0.0			5.2
26	2.5	3.4	5.4	11.9	13.6
27	0.0	1.1	9.8*	nc	10.6

Variance estimation: Given the estimated parameters, the variances can be calculated over the trajectory of time as well as covariances between any combination of time points (Kirkpatrick et al. 1990). Figure 3 shows the average of the estimated curves of QTL variance over time over 100 replicates in the three scenarios, as well as the curves based on simulation input parameters.

Figure 3. Curves of simulated (squares) and mean of estimated (triangles) variance of QTL effects over the time period in scenario 1 (a), scenario 2 (b), and scenario 3 (c).

In the analysis of field data the variance explained at different stages of lactation were calculated at the most likely position on each chromosome. While some QTL had a constant effect others changed over the lactation (Figure 4).

Figure 4. QTL variance function over time (percent of genetic variance) at maximum likelihood estimates for the seven chromosomes

Discussion

Results show that the novel QTL-mapping method, based on a Random Regression model, provides an appropriate and a very powerful tool to detect QTL affecting traits that are measured repeatedly over time or interacts with environmental factors.

The power to detect QTL increase substantially compared with a univariate analysis, especially when the QTL-allelic effects change over time. This is because the effect is averaged over the whole period in the univariate model, and is thereby not large enough to be detected. The Random Regression model estimates the QTL variance at all time points and utilises the large effect in parts of the period. In the analysis of field data the random regression models identified more QTL (4 vs. 1) compared to traditional 305-day models. One chromosome reached significance with a one parameter Legendre polynomial, four with a two parameter polynomial but none with more than two parameters. This is due to the many parameters in the model. Other functions might fit the QTL curves with less parameters and might thereby improve the method further.

The method can provide important biological information on the expression pattern/effect over time. In the field data some QTL have high effect in the beginning of lactation (contributing to physiological stress) but have very little effects over 305 days. These might be candidates to select against, which must be verified in multiple trait analyses with functional traits. Others have higher effects throughout lactation and are candidates to select for. Furthermore, knowledge of QTL-shape parameters are potentially needed to predict breeding values.

Generally, the shapes of the estimated curves agree very well with the simulated curves. The exceptions are a slight overestimation of QTL variance in scenario 1, and a slight underestimation in the end of the time period in scenario 3. In each case, the polygenic variance showed a bias in the opposite direction. This indicates the difficulty in decomposing the genetic variance into two additive components, which is enlarged when the model is mis-specified as in scenario 1 where QTL effects were fitted with a four parameter Legendre polynomial while simulated with a constant (one parameter polynomial).

References

- Fernando, R.L., and Grossman, M., 1989. *Genet. Sel. Evol.* 21: 467-477.
- Grignola, F.E., Hoeschele, I. and Tier, B., 1996. *Genet. Sel. Evol.* 28: 479-490.
- Jensen, J., Mantysaari, E., Madsen, P., and Thompson, R., 1997. *Jour. Ind. Soc. Ag. Stat.* 49: 215-236.
- Jakobsen, J., 2000. PhD Thesis, Royal Veterinary and Agricultural University, Copenhagen, Denmark.
- Kirkpatrick, M., Lofsvold, and Bulmer, M., 1990. *Genetics* 124: 979-993.
- Meyer, K., Grasser, H.U. and Hammond, K., 1989. *Livest. Prod. Sci.* 21: 177-199.
- Pander, B.L., Hill, W.G. and Thompson, R., 1992. *Anim. Prod.* 55: 11-21.
- Swalve, H.H., 1995. *J. Dairy Sci.* 78: 929-938.
- Wang, T., Fernando, R.H., van der Beek, S. and Arendonk, J.A.M., 1995. *Genet. Sel. Evol.* 27: 251-274.

Reducing the size of large experiments and treating missing observations

Rodrigo Labouriau¹ and Peter K. Theil²

¹ *Biometry Research Unit, Dept. of Animal Breeding and Genetics, DIAS*

² *Dept. of Animal Nutrition and Physiology, DIAS*

Dansk resumé

Hvordan kan man reducere et stort forsøg og hvordan kan man håndtere tabte observationer. Der bliver diskuteret et alternativ for forsøgsreduktion vha. et lille forsøg kørt sammen med et stort forsøg hvor en associeret egenskab bliver målt. Som eksempel bliver det vist hvordan søernes mælkeproduktion kan estimeres ved at måle mælkeoptagelsen hos kun to pattegrise per kuld meget præcist og måle pattegrisenes tilvækst for alle i kullet. En sådan reduktion af forsøgsstørrelsen kan lade sig gøre ved anvendelse af en moderne statistisk teknik: estimation vha. EM-algoritmen for behandling af ukomplette data. Et stort simulationsstudium (mere end 100.000 simulationer) viser, at metoden kan yde rimelige resultater i konkrete praktiske situationer. Den samme teknik kan bruges til at mindske forsøgsstørrelsen i andre anvendelsesområder og til at behandle forsøg hvor nogle observationer går tabt.

Abstract

We discuss an alternative for experiment reduction by using a small experiment coupled with a large (non-expensive) auxiliary experiment. The ideas presented are illustrated by the estimation of sow milk. The total sow milk production could be estimated by very accurately determining the individual milk intake by the D2O dilution technique of only two piglets per litter. This drastic experiment reduction is obtained by using a modern statistical technique: the EM-algorithm-based imputation of incomplete data. A large simulation study (more than 100, 000 simulations) is reported and indicates that the proposed methods can produce reasonable results in concrete practical situations. The same technique can be used for reducing the experimental size in other areas and for treating experiments with missing observations.

Key words: Experiment-reduction, Incomplete-data, EM-algorithm, Missing-data, Milk-production, Sow, D₂O-dilution-technique and Water-turnover.

Introduction

Suppose that you have to perform an experiment that involves a very expensive or time-consuming measurement or laboratory determination. You would probably be tempted to

perform a reduced version of the experiment where the expensive measurements are performed only a few times. But, by doing so you might get very imprecise or even inconclusive results, since the experiment you perform might be too small. All the work and investment would be then in risk. After thinking for a while on the problem you realize that there is another measurement that you can use instead of the original expensive determination. This new measurement is not expensive, is somehow related with the original determination, but, unfortunately, does not measure exactly what you want to study. You are then faced with three possibilities: 1) carrying out a large and expensive experiment anyway (ruining your budget), 2) performing a small experiment (under the risk of getting inconclusive results and therefore wasting resources and time) and 3) giving up and making the experiment only based on the discount measurement which is perhaps only vaguely related with the question you want to answer (and therefore also wasting time and resources).

Stories like the one above occur often in the experimentation practice. Indeed we will discuss here a real example of this hypothetical situation: the determination of sow's milk production using the heavy water (D_2O) dilution method. Our main point is that there are often other choices apart from the three possibilities mentioned in the last paragraph. We will show that in some cases it is possible to carry out a relatively large experiment using the non-expensive (but not ideal) determination for all the observations and simultaneously perform the precise but expensive measurements only for a few units. The whole experiment becomes less expensive or less time consuming and, perhaps surprisingly, it might still be possible to obtain accurate results, provided that appropriate statistical and mathematical techniques are used. Before going further with this general explanation of our proposition let us consider a concrete example.

Example: Determination of sow milk production by the heavy water dilution method

The sow milk production can be determined by the heavy water (D_2O) dilution method. This modern technique can produce very accurate determinations and posse's advantages over the classic methods, which typically are based on successively weighting piglets (see Theil et al, 2002). The heavy water dilution method consists essentially of injecting a known amount of D_2O in each piglet at the beginning of the experiment and determining the individual milk intake of each piglet by using the decay of the concentration of D_2O during the observation period. Although the D_2O dilution technique is ingenious and very accurate, it is clearly expensive and time consuming (specially if one thinks on it as a routine determination for litters typically containing 10-12 piglets in a large animal breeding program). Alternatively, one might use the weight gain of the piglets in the period for which we want to determine the sow production. This is inexpensive and definitively a straightforward alternative technique. However, although weight gain is expected to be related with milk intake, it might be influenced as well by many other factors that have nothing to do with milk intake, for example, individual variations in the basal metabolic activity (both genetically or

environmentally induced), individual variations in absorption and metabolism of nutrients, diseases, variations in the milk composition, not to mention others. As a matter of fact, sometimes one can even observe negative weight gains even though the piglet is known to intake milk. Moreover, as we will see in our data, the relation between milk intake and weight gain depends on the age and development stage. In summary, weight gain is related to milk intake but it cannot replace satisfactorily the determination of milk intake. This is precisely the situation that we described in general terms above.

The main idea behind the method that we will discuss here is to consider the large experiment, for which we only observed the non-expensive measurement, as being indeed a large experiment intended to be based on the expensive determinations, but for which we have many missed observations (the observations for which we have not performed the expensive determinations). This point of view allow us to use the whole machinery developed originally to treat experiments with incomplete or missing observations to reduce the number of the expensive determinations (and still get something useful). There are many methods for treating incomplete observed experiments that can potentially be used in this context, but we chose one technique (probably biased by the example we had originally in mind or by the authors taste), namely: EM-algorithm like imputation with an auxiliary variable (see Dempster *et al.*, 1977). This method takes advantage of the association between the two determinations (the expensive and the simple). We use the common observations to estimate a suitable statistical model relating the two determinations (say, a regression like model). Next, the values that we would have observed of the expensive determination in the large experiment (if we had made this measurement for the whole large experiment) are predicted using the results of the non-expensive measurement and the model relating the two determinations. We then replace the non-observed values of the expensive determination in the large experiment by the predicted values for it. We call this new data the *imputed data*. A new model relating the expensive determinations to the non-expensive determination is fit, now based on the imputed data. The originally non-observed values of the expensive determination are replaced by the predicted values from the new model (but not the originally observed data for the expensive variable) generating a new imputed data. The process is repeated until no essential changes are observed (we define that more precisely latter). Although the process described above might seem to be complicated, it can easily be implemented using standard statistical software. Moreover, many theoretical and practical considerations show that this technique produces reasonable results (we return to this point below).

Back to the example of sow milk production

In our example on sow milk production the “expensive measurements” are the D₂O dilution determinations and the “non-expensive measurements” are the weight gain measured during the observation period. The experimental set-up used involved four different sows and

measurements made in three different age periods. The weight gain in each period was measured for all piglets, but the milk intake measured by the D₂O dilution method was determined only for some piglets.

Figure 1. Scatter plot of the individual milk intake (g/day) against the weight gain (g/day) in the weeks 1 and 3. The two lines drawn are the predicted values of a quadratic regression fit separately to the first week (continuous line) and the third week (interrupted line).

Figure 1 display a scatter plot of the D₂O dilution determinations against the observed gain weight for two age periods (for the piglets for which we had both observations). On the basis of some explorative analyses (not shown here) we decided to use a quadratic relation between the milk intake and the weight gain, that is, we describe the milk intake by a polynomial of second degree applied to the weight gain (one different polynomial for each week, as it is clear from Figure1), i.e.

$$Y_{ijk} = \alpha_i + \beta G_{ijk} + \gamma (G_{ijk})^2 + \varepsilon_{ijk}, \quad (1)$$

where Y_{ijk} is the milk intake (g/day) and G_{ijk} is the weight gain (g/day) of the k^{th} piglet ($k = 1, 2, \dots$) in the j^{th} litter ($j = 1, 2, 3, \text{ and } 4$) in the i^{th} period ($i = 1, 2, \text{ and } 3$). Here α_i (depending on the week), β and γ are the regression parameters and ε_{ijk} are regression residuals. In fact, we used a more complicated model taking into account the dependency present in the data, but this is irrelevant for now (we describe the model in more details in the material and methods). Note, moreover, that the model regression lines drawn in Figure 1 are obtained from separate adjustments of a quadratic regression for each week, therefore the line would not necessarily be parallel, as one would expect for the predicted values of the model given by

(1). However, it can be seen that the lines are almost parallel, confirming that this assumption on parallelism is reasonable. We perform indeed other formal tests and model control techniques to ensure that the model used is adequate.

The idea behind the proposed method is to insert the observed weight gain in the equation (1) to obtain estimates of the values of the individual milk intake of the piglets for which we have not observed the milk intake. Replacing that values we form a new data set, termed the imputed data set. A new regression is fit using the imputed data (i.e. we obtain new values for the parameters α , β and γ) and then, using this new regression, we construct a new version of the imputed data by replacing the values of the milk intake for the piglets for which we did not know originally the milk intake. A provisory estimate of the sow's milk production is obtained by summing the milk intake of all piglets in the litter (both the originally known values and the estimated values). The process is then repeated until no more changes are observed. This will produce the final estimates of the sow's milk production.

The aim of this text is to illustrate the general methodology proposed for reducing (or at least partially reducing) experiments. The details of the example used are irrelevant here in this context and are described and studied in the specific literature of sow milk production (see Theil *et al.*, 2002 and Theil, 2002). The reader is then asked to keep in mind this purpose when reading further.

Our experience in applications of methods for treatment of data with missed or incomplete observation shows that a common (and comprehensible) reaction of the not specialized reader is to be sceptical with respect to these methods. The imputation using the EM-algorithm might sound as pure data manipulation and it might be difficult to believe that this kind of method works properly. Of course it is not convincing (and not valid) to argument that this method is in fact theoretically well funded and has been successfully applied in a wide variety of situations. Therefore we present in the rest of this text the results of a large simulation study. We will use an almost complete experiment, where both the milk intake measured by the heavy water dilution method and the weight gain are measured for all the observations. This data will allow us to simulate the situation of incomplete or reduced experiment described above. The completeness of the example used will allow us to have a standard for comparing the results obtained with deliberately artificially reduced experiments, with the results obtained with the complete experiment. We will show that it is possible to reasonably reconstitute the results of the complete experiment when using only three or even only two piglets per litter (a typical litter has 10-12 piglets). This might be a more convincing (and reasonable) argument for advocating the use of the proposed methods.

The rest of the text is organized as a standard article. In the material and methods section we describe with more technical details the statistical methods and models applied, the simulation study performed and the measures of quality used to evaluate the results of the simulation study (specifically adapted notions of accuracy and precision). A reader not interested in

technical details can skip this section in a first reading. The section of results presents partial results of the large simulation study (the complete simulation study will be published somewhere, hopefully soon). Some concluding remarks are provided in the last section.

Material and methods

The experimental procedure

Four multiparous sows (Danish Landrace x Yorkshire) were randomly selected from the herd at the Danish Institute of Agricultural Sciences. Their milk productions were determined in weeks one, two, and three of lactation by means of D₂O dilution in the piglets (for details see Theil et al., 2002 and Theil, 2002). The method of D₂O dilution requires that at least one piglet per litter is used as a control and therefore its milk intake was not determined. We eliminated deliberately these piglets from the data we analyse here in order to obtain complete reference data. This is not correct in practice, but is irrelevant for the demonstration purposes of the present study. Note however that the weight gain is available for the control piglets, which would allow us to use them for estimation with the proposed method.

The EM-algorithm for imputation of missed data and estimation of the sow milk production

The observations for which both the milk intake and the weight gain were known are termed here the *complete observations* and the observations for which only the weight gain was known are called the *missing observations*. The following mixed covariance analysis model was successively fitted in the algorithm described below and used to estimate the milk intake of the missing observations. Let Y_{ijk} and G_{ijk} represent the milk intake and the weight gain for the k th piglet ($k = 1, 2, \dots$) in the j th litter ($j = 1, 2, 3, \text{ and } 4$) in the i th week ($i = 1, 2, \text{ and } 3$). The model is given by

$$Y_{ijk} = \mu + \alpha_i + \beta G_{ijk} + \gamma (G_{ijk})^2 + \upsilon_j + \tau_{jk} + \varepsilon_{ijk} ,$$

where α_i , β and γ are the regression parameters, υ_j and τ_{jk} are normal distributed random components (with mean zero and each with its own variance) of sow and piglets within sow, respectively, while ε_{ijk} are normal distributed residuals. The random variables υ_j , τ_{jk} and ε_{ijk} are independent.

The steps of the proposed algorithm are:

1. Construct a data set constituted only by the complete observations and call it the *current full data*
2. Estimate the parameters of the predictive model described above based on the *current full data*. Call these estimates the *current values of the parameters*
3. Evaluate the *current predicted milk intake of the missing observations* by the predicted values obtained from the predictive model, using the current values of the parameters and the known weight gains of the missing observations
4. Construct the new *current full data* by replacing the values of the milk intake of the originally missed observations in the *current full data* by the *current predicted milk intake of the missing observations*
5. If the changes of the parameter estimates were less than a prespecified tolerance limit, then declare that the algorithm converged and stop the process; otherwise, return to step 2.

This algorithm is an adaptation of the general algorithm for imputation of incomplete data proposed in Dempster et al. (1977). The algorithm converges under very mild regularity conditions.

The simulation study

We simulated the situation where the milk intake was determined by the D2O dilution only for a reduced number of piglets per litter by deliberately eliminating the information of milk intake for some piglets and considering them as missing observations. In this way we generated data sets where the information on the milk intake was available only for 2,3,..., 12 piglets per litter. For each of these types of data we generated 10,000 data sets by sampling with reposition which piglets would keep their observed milk intake. All in all we generated $(11 \times 10,000)$ 110,000 data sets. For each of those data sets we estimated and recorded the sows' milk production by using the estimation procedure described above.

The measures of quality of the estimates

The relative accuracy and the precision of the procedure were then characterized as bellow. Here the relative accuracy is understood as a measure of how close the results of the estimation procedure with the incomplete data were from the estimation with the complete data (see Kotz and Johnson, 1982). Accuracy can be measured by many ways. We choose the relative bias, defined as the mean difference between the estimates obtained with the incomplete data sets and the estimates obtained with the complete data set. In order to standardize the measures of the accuracy we divided the mean difference above by the estimates based on the complete data and multiplied by 100, obtained the accuracy as a (signed) percentage of the estimates with the complete data. Note that here the estimates

based on the complete data play the role of “true value” of the quantity to be measured in the classic definition of accuracy. Therefore the notion of accuracy used here was termed relative accuracy. Furthermore, we adopted a sampling schema with reposition in the simulation study, which implies that the mean differences used are the so-called bootstrap estimates of the differences between the estimates from the incomplete data sets and the estimate based on the complete data set. That is, the operational definition of accuracy here is in fact a bootstrap estimate of the theoretical accuracy. The bootstrap estimates are known to possess reasonable properties such as consistency (i.e. approximate the right quantity), see Davison and Hinkley (1997) and the vast literature referred therein.

The precision, understood as the resolving power of the estimation procedure, might be measured by the sample variance of the values of the estimates based on the incomplete data. In order to standardize the results we used the coefficient of variation instead of the variance, i.e. we divided the square root of the variance by the sample mean and multiplied by 100. Analogously to the case of the accuracy, our estimate of the precision can be viewed as a bootstrap estimate of the actual precision, due to the sampling schema used.

Results

The (partial) results of the simulation study are presented in Table 1 and Figures 2 and 3 given below.

Table 1. Estimated relative accuracy (relative bias in % of the complete estimation) and precision (coefficient of variation in %) for the four sows in the three weeks obtained with only two piglets per litter with information on milk intake (i.e. with the worst simulated case).

Litter	Coefficient of variation (%)			Relative bias (%)		
	Week 1	Week 2	Week 3	Week 1	Week 2	Week 3
1	4.14	3.69	3.09	-0.51	2.18	-1.38
2	4.21	3.95	3.73	-5.30	-4.65	0.52
3	3.82	3.62	6.75	7.91	-0.42	-3.60
4	5.98	3.62	3.21	-1.85	2.60	-0.86

Figure 2. Estimated relative accuracy (relative bias in %) for the fourth litter and the three weeks for different numbers of piglets per litter for which the information on the milk intake was used.

Figure 3. Estimated precision (coefficient of variation in %) for the fourth litter and the three weeks for different numbers of piglets per litter for which the information on the milk intake was used.

Discussion and conclusion

We presented a large simulation study demonstrating that the proposed technique for experiment reduction and estimation can produce reasonable and reproducible results, even when the experiment reduction is as drastic as using the complete information for only two piglets per litter. There is bias and larger dispersion of the estimates when a significant part of the experiment is reduced, as one might expect. However, the lost in accuracy and precision is kept still in a reasonable order of magnitude, even in the most extreme case - eliminating the information of milk intake for all but two piglets per litter – where the larger coefficient of variation encountered was close to 7% and the larger relative bias was around 8%. Both the accuracy and the precision tended to increase when the number of completely observed piglets per litter increased.

The method proposed is expected to be of general applicability, not necessarily restricted to the example presented here. However, it is not to be expected that spectacular experiment reductions with insignificant lost in the accuracy and the precision will be always obtained. It seems to be essential the auxiliary determination (i.e. the non-expensive measurement) to be strongly associated with the main determination (i.e. the expensive measurement) for obtaining good results with the proposed methodology. Moreover, the predictive model relating the auxiliary and the main determinations should be carefully chosen. The quadratic mixed model used here is suitable for these data, but not necessarily appropriate even for other similar data on sow milk production.

The methods presented here are applicable also for treating incomplete experiments with relatively many missing observations (since they were originally designed for that purpose!) with the obvious modifications. The use of the proposed techniques to reduce experiments will probably depend on the creativity and imagination of the researchers in the field. The necessary techniques are at the top of our fingers.

References

- Davison, A.C., and Hinkley, D.V., 1997. Bootstrap methods and their application. Cambridge University Press, UK.
- Dempster, A.P., Laird, N.M. Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* 39, 1-38.
- Kotz, S. and Johnson, L.N., 1982. *Encyclopaedia of Statistical Sciences*. Vol.1. John Wiley & Sons, New York.
- Theil, P. K. 2002. Energy and protein metabolism in pregnant and lactating sows. Ph.D. Diss., The Royal Veterinary and Agricultural University and Danish Institute of Agricultural Sciences, Denmark, 140 pp.

Theil, P.K., Nielsen, T.T., Kristensen, N.B., Labouriau, R., Danielsen, V., Lauridsen, C and Jakobsen, K. 2002. Estimation of milk production in lactating sows by determination of deuterated water turnover in three piglets per litter. *Acta Agric. Scand., Sect. A*, 52, 221-232.

Construction of homogeneous zones for field experimentation

Ulrich Halekoh¹ and S.E. Olesen²

¹ *Biometry Research Unit, Dept. of Animal Breeding and Genetics, DIAS*

² *Dept. of Agroecology, DIAS*

Abstract

In designed experiments on field level it is advantageous to use as blocks areas which are homogeneous with respect to within field variability which is described by measurements easily available on a dense grid like soil electrical conductivity (EM38), former yield experience or topological landscape characteristics. The construction of such blocks is an unsupervised classification task based on the continuous multivariate measurements. The construction is based on a probabilistic model that allows incorporation of the spatial aspect in order to insure that the blocks are spatially connected. The usefulness of classifications obtained by such surrogate measurements to more expensive but possibly more relevant soil variables like clay or soil organic matter is discussed.

Keywords: Unsupervised classification; Hidden Markov model; soil electrical conductivity; EM38.

Introduction

For plant experiments on larger fields it may be advantageous to characterize the variability of the response variable with respect to the within field variation. Sub-areas where the potential response of the plants (e.g. yield in its different facets as total plant mass or number of germinable seeds) is homogeneous can be used as blocking units in the experimental design in order to control for variability due to different soil properties or topological characteristics.

The task is to combine several such field characteristics to construct a classification of the field in homogeneous zones which are as much as possible different from each other. The success of the final classification for the use of blocking units depends of course on the extent of the relation of the yield to the variables used in the classification process (cf. Østergaard (1997) in the context of N fertilization)

In the case of a single field characteristic as for example the slope of the field, choosing the zones as having similar value with respect to slope would aim to minimize the within block variability and thus minimize the variability of the treatment difference. A similar approach applied to multivariate measurements from each location is known as k-means clustering. The disadvantage of this classification algorithm is that it does not incorporate the spatial aspect of the data. In this article we use an approach based on a common stochastic model (called

Hidden Markov model) for the continuous multivariate measurements at the locations and the locations class membership indicator. The algorithm for the construction of the zones can be derived from statistical estimation principles. A single parameter is used to control the spatial smoothness of the classification and - indirectly- the size of the zones.

Alternative methods to incorporate the spatial coordinate into the classification process are for example by using them as additional variables appropriately weighted in the analysis (c.f. Berry, 1966), imposing spatial contiguity constraints (Lebart, 1978) or using a modified dissimilarity matrix based on estimation of variograms (Oliver and Webster, 1989).

In the example we will use three easily available variables for the construction of homogeneous zones: soil electrical conductivity (SEC), wetness index (WET) and former yield maps. Soil electrical conductivity measured as electromagnetic induction (EM38) is nowadays easily and cheaply available on a dense spatial grid (cf. Nehmdahl and Greve (2001), Lück and Eisenreich 2001). It is known to be related to soil temperature and water content and to be positively correlated with soil characteristics as content of clay and –to a lesser extent- the content of soil organic matter (SOM). The wetness index (or compound topographic index) is a measure combining local and global topological information.

The present study was not primarily conducted to analyse the construction of experimental blocking units on a field but for the construction of management zones for precision agriculture. The goal is similar to the experimental setting as one wants to identify zones which are homogeneous to some need (e.g. phosphorus) in order to apply a management activity (e.g. fertilisation) at a constant rate. In the experimental setting one might have stronger requirements with respect to the minimal size of a zone or its shape (e.g. one could require it to be rectangular).

Material and Methods

We consider the construction of management zones as an unsupervised classification or classification problem. Multivariate measurements at a dense grid in the field like soil conductivity and wetness-index are used to compute a classification for the locations in a field. The management zones are spatially contiguous subsets of the locations belonging to the same class.

We take account of the spatial relation between the measurements by using a hidden Markov model. Given the class membership (called label) of each location, we assume the measurements to be independent multivariate normal with a separate mean vector and diagonal covariance matrix for each class. The unobserved labels itself are distributed according to a Strauss-Pott Markov random field, i.e. knowing the labels of the neighbors at a location s the probability that this location has the label k is proportional to

$\exp(\beta \cdot [\text{number of neighbors with label } k])$ (Ripley, 1991). The larger the parameter β the stronger is the requirement that neighboring locations should belong to the same class and the more spatially connected are the zones within the classes. From the distributional specifications one can calculate the probability of a class given the measurements. One possible estimate of the classification is based on the marginal posterior mode of a location to belong to a classification which is called the MPM estimate. The estimation of the mean vectors and the covariance matrices of the multivariate normal distributions are included in the iterative maximization process. Dang and Govert (1998) use an EM-algorithm to calculate the estimates of these parameters. Alternatively the estimation of the class membership can be based on the maximal posterior probability (called MAP estimate) which is calculated via the ICM-algorithm of Besag (1986). The computation of the classification is based on a program developed by M. Dang. It is available at www.hds.utc.fr/mdang. Instead of trying to estimate β , we consider it more as a tuning parameter with which one can control the spatial contiguity of the zones within the classes.

Data from a test field was used in this study. The 10 ha test field is located on a young moraine landscape of loamy and clay till, glaciofluvial sand and peat. Elevation ranges from 54 to 67 m. The natural drainage is good except for a small peat basin in the central part. Soil data was collected in a 20 x 20 m grid at soil depth of 0 - 25 cm and 25 - 75 cm. The clay content varies from 3 to 24%. Generally the clay content for sub soil samples is slightly higher than for topsoil samples. The plant available potassium content (K) was measured for both soil depths. Apart from the peat basin soil organic matter (SOM) content varies between 1.3 and 3.3 %. SEC data were collected in an app. 5 x 10 m grid using a Geonics EM38 sensor. The sensor was towed behind an all-terrain-bike and data from the EM38 was merged by a GPS-unit. A digital elevation model (DEM) was calculated on the basis of height measurements in a 20 m grid. The wetness index (WET) (Burrough, Mc Donnell, 1998) combines local and global topological information: $WET = \log(AC / \tan \alpha)$, where AC is the catchment area and α the flow slope. WET was calculated from the DEM. Average yield from the years 1993 to 1995 defined the YIELD variable.

Because the different variables were not measured exactly at the same locations they were kriged to a 10m times 10m grid prior to the analysis. Each variable was standardized to have zero mean and a variance of 1. To describe the extent to which a classification explains the variation of a feature variable x , we computed the ratio of the weighted sum of the within variances to the total variance:

$$\sum_k \sum_{s \in C(k)} (x_s - \bar{x}(k))^2 / \sum_s (x_s - \bar{x})^2 ,$$

where $C(k)$ are the locations of classes k , $\bar{x}$ is the overall mean and $\bar{x}(k)$ the mean in class k .

Table 1. Partial correlation of soil properties with the variables used for classification

		Partial correlation			Unexplained Variance
		variable			
		EM38	WET	YIELD	1-R ² (%)
Property	Soil depth				
SOM	25 cm	0.40	0.41	-0.42	66.08
	75 cm	0.39	0.39	-0.36	69.47
K	25 cm	0.19	0.09	-0.08	95.73
	75 cm	0.52	-0.02	-0.15	70.32
CLAY	25 cm	0.68	0.01	0.14	39.51

We constructed for the number of classes ranging from 2 to 5 ‘dummy’ classifications solely based on the spatial coordinates such that a location belonging to class k had minimal Euclidean distance to its class center compared to the other centers. These spatially strongly contiguous classes with nearly convex shape were used as a reference classification for the ones derived via the hidden Markov model.

Results

Classifications were calculated for the test field based on the three variables EM38, WET and YIELD for different degrees of smoothing, i.e. different β varying the number of classes from 2 to 5. It turned out that for β larger than 1.5 the resulting classifications were nearly the same, whereas for β smaller than one many small management zones were created.

The classification method has not the property of monotony that the classification based on n classes may be derived from the rougher classification into only $n-1$ classes by the split of classes. This means especially that management zones from a classification into n classes are not subdivisions of management zones from a classification into $n-1$ classes.

For the algorithm to obtain the optimal solution it was necessary to use a good initial solution. Taking the k-means classification as initial solution led to a solution with much larger objective functions than using random start classifications.

The marginal posterior mode principle (MPM) yielded classification less dependent on the initial classification than the classification based on the algorithm with respect to the maximum posterior probability.

Figure 1. Steps for the classification into 5 classes with $\beta=2.0$ based on three dimensional measurements (MPM principle)

Figure 1 depicts the process of the classification for five classes based on three variables EM38, YIELD and WET. The three maps on the left are the original measurements, the upper map gives the initial classification based on k-means classification. The lower map gives the final classification based on the MPM principle with the choice of $\beta = 2.0$.

Table 2. Unexplained variance for soil properties at 2 different soil depths based on classifications with different numbers of classes.

Property	Variables	Unexplained Variance							
		Soil depth							
		25				75			
		Number of classes				Number of classes			
		2	3	4	5	2	3	4	5
SOM	EM38, YIELD, WET	0.99	0.91	0.62	0.27	1.00	0.94	0.45	0.43
	EM38, YIELD	0.83	0.85	0.95	0.91	0.87	0.88	0.96	0.93
	Dummy	1.00	0.91	0.89	0.91	0.99	0.92	0.90	0.91
K	EM38, WET, YIELD	0.99	0.93	0.92	0.92	0.97	0.84	0.84	0.79
	Dummy	0.99	0.98	0.89	0.87	1.00	0.99	0.89	0.84
CLAY	EM38, WET, YIELD	0.73	0.55	0.53	0.44	0.85	0.73	0.72	0.62
	EM38	0.44	0.49	0.46	0.46	0.63	0.67	0.63	0.63
	Dummy	1.00	0.91	0.68	0.67	1.00	0.95	0.78	0.80

The ability of the classification based on the three variables EM38, WET and YIELD was analysed with respect to the extent it represents the spatial variation of the soil characteristic feature variables clay content, SOM, and K and in 25 and 75 cm soil depth.

To get a first insight into how much the variables EM38, WET and YIELD allow prediction of the soil parameters the Table 1 gives the partial correlation of the soil parameters with these variables, the correlation giving an impression on how important the corresponding variable is for predicting the soil parameter if one has already corrected for the other variables. The $1 - R^2$ of a simple regression of the soil variables on EM38, WET and YIELD, gives a rough indication on the unexplained variation with respect to using the variables themselves and should be about a lower limit for the unexplained variation achievable by a classification with a non excessive number of classes based on these variables.

In Table 2 the unexplained variation for the soil variables with respect to classifications with 2 to 5 classes based on the MPM principle with $\beta = 2.0$ are given. With the exception of SOM an area of high organic content was excluded from the area to be classified.

For clay the high partial correlation for EM38 and the low ones for the other variables indicate that after accounting for EM38 the variables WET and YIELD do not considerable add to the prediction of clay. The unexplained variation based on the classification of all three variables (Table 2) attains values lower than 0.6 for more than 3 classes for clay at 25 cm depth. Using EM38 alone for the classification, the explained variation at soil depth of 25 cm is even lower for less than 5 classes.

For SOM the absolute value of the partial correlations for all three variables are nearly the same. Dropping the wetness index from the computation of the classification increases the unexplained variation for classification with higher number of classes. The classification cannot be distinguished with respect to the unexplained variation from the dummy classification.

For potassium the classification is not different with respect to explained variation by the dummy classification (Table 2). Only for K in 75 cm the variable EM38 shows appreciable partial correlation and the unexplained variation by the classification is smaller than for the 25 cm depth.

Discussion

The Hidden Markov random fields approach provides an automatic method for the construction of blocking zones using continuous multivariate surrogate measurements for yield on a field. The spatial context is incorporated by the specification of a simple probabilistic model for the common class membership of neighbouring locations and the degree of spatial smoothness is controlled by a single parameter. Because it is an unsupervised classification the success of the classification may depend on the actual type of yield used in the experimentation. Taking clay content, soil organic matter and potassium concentration as variables representing directly a yield response satisfactory classification could be obtained for clay and SOM. Furthermore, it turned out that including more variables for classification does not necessarily improve the explained variation - for clay just using EM38 gave similar results to using additionally WET and YIELD in the classification. On the other hand, excluding variables from the construction of the classification may decrease the explained variation, as was the case with SOM.

For K, no appreciable amount of variation could be explained by the classification by any number of classes between 2 and 5.

For clay and SOM results were in accordance with a simple correlation analysis. For soil clay the predictors YIELD and WET had no predictive power after adjusting for EM38, whereas for SOM all three variables had similar partial correlations.

In the present implementation of the Hidden Markov classification algorithm we used a very simple model to control the spatial appearance of the zones. Stronger requirements on the shape of the zones, e.g. that they should be of rectangular shape of about a certain size, can be incorporated into the approach via the prior distribution for the labels.

References

- Ambroise, C., Dang, M. and Govaert, G., 1997. Clustering of spatial data by the EM algorithm. IN: *geoENV I- Geostatistics for Environmental Applications*, A. Soares, J. Gómez-Hernandez and R. Froidevaux (eds.), 493-504. Quantitative Geology and Geostatistics, vol. 9. Kluwer academic Publishers.
- Berry, B.J.L., 1966. Essay on commodity flows and the spatial structure of the indian economy. Research paper 111. University of Chicago, Department of Geology.
- Besag, J., 1986. On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. B*, 48, 259-302.
- Burrough, P. A. and McDonnell, R.A., 1998. *Principles of Geographical Information Systems*. Oxford: Oxford University Press.
- Dang, M. and Govaert, G., 1998. Spatial Fuzzy Clustering using EM and Markov Random Fields. *System Research and Information Systems*, 8, 183-202.
- DeGrujter and McBratney, 1988. A modified fuzzy k-means for predictive classification. In: Bock, H.H. (ed.) *Classification and related Methods of Data Analysis*, Elsevier Science, Amsterdam, pp. 97-104.
- Lebart, L., 1978. Programme d'aggrégation avec constraints. *Cahiers de l'analyse des données* 3, 275-287.
- Lück, E. and Eisenreich, M., 2001. Electrical Conductivity mapping for Precision Agriculture, *Proceedings of the 3rd European Conference on Precision Agriculture* (ed. Grenier, G. and Blackmore, S.), agro Montpellier, Montpellier, p. 425-430.
- Østergaard, H.S., 1997. Agronomic consequences of variable N fertilization. In: *Proceedings of the 1st European Conference on Precision Agriculture*, ed. Stafford, J.V. BIOS Scientific Publishers, Oxford, 1, 315-320.
- Nehmdahl, H. and Greve, M., 2001. Using soil electrical conductivity measurements for delineating management zones in highli variable soils in Denmark *Proceedings of the 3rd European Conference on Precision Agriculture* (ed. Grenier, G. and Blackmore, S.), agro Montpellier, Montpellier, p. 461-466.
- Oliver, M.A. and Webster, R., 1989. A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology* 21, 15-35.
- Ripley, B.D., 1991. *Statistical Inference for spatial processes*. Cambridge: Cambridge University Press.
- Seber, G.A.F., 1984. *Multivariate Observations*, New York: John Wiley & Sons.

Anvendelse af ufuldstændige blokforsøg

Kristian Kristensen¹, Jakob Willas², Lise Nistrup Jørgensen³ og Rene Gislum⁴

¹ *Forskergruppe for Biometri, Afd. for Husdyravl og Genetik, DJF*

² *Afd. for Sortsafprøvning, DJF*

³ *Afd. for Plantebeskyttelse, DJF*

⁴ *Afd. for Plantebiologi, DJF*

Sammendrag

I forsøg med mange forsøgsled, forsøg på uensartede arealer og forsøg, som af praktiske grunde skal opdeles på mindre enheder, vil det ofte være fordelagtigt at anvende blokke, som ikke indeholder alle forsøgsled. Sådanne ufuldstændige blokke kan konstrueres efter flere forskellige principper. I enfaktorielle forsøg anvendes oftest metoder, som sikrer, at alle par af forsøgsled kan sammenlignes med den samme (og bedst mulige) sikkerhed, mens man i flerfaktorielle forsøg oftest anvender metoder, som sikrer, at de mest interessante effekter (hoved- og/eller vekselvirkninger) kan estimeres bedst muligt. Artiklen viser tre eksempler på anvendelse af ufuldstændige blokforsøg: 1) Sortsforsøg i ærter (enfaktoriel), 2) Frøavlsforsøg med strandsvingel (trefaktoriel) og 3) Forsøg med svampemidler i byg (tofaktorielt split-plot). For hver af disse vises den benyttede plan samt effekten af at benytte planen i stedet et randomiseret blokforsøg med fuldstændige blokke.

Nøgleord: Generaliserede lattice forsøg, α -forsøg, effektivitet, Lsd, enfaktoriel, flerfaktoriel, split-plot.

Introduktion

En del forsøg har mange forsøgsled. Hvis sådanne forsøg skal anlægges med fuldstændige blokke betyder det at blokkene skal indeholde mange parceller, og blokkene dermed bliver store. Da jordens variabilitet (oftest) stiger med blokstørrelsen, betyder dette, at den tilfældige variation vil blive stor, når der er mange forsøgsled, og det kan så blive besværligt at påvise forskelle mellem forsøgsløddene. Der er flere måder, hvorpå man kan forsøge at løse dette problem.

I nogle tilfælde kan man gøre blokkene mindre ved at dele det nødvendige antal planter på flere parceller (gentagelser). F.eks. kunne man vælge at have 6 parceller á 10 planter for hver behandling i stedet for 3 parceller á 20 planter. Der er dog en grænse for, hvor langt denne proces kan fortsætte i praksis; dels fordi forsøget vil blive dyrere jo flere gentagelser man deler det nødvendige antal planter på, og dels fordi arealet i de enkelte blokke efterhånden

ikke reduceres nævneværdigt, da en stor del af denne nu vil blive brugt til værn og gangarealer.

Man kan forsøge at begrænse forsøgets størrelse ved at dele det i flere mindre forsøg. Herved begrænses størrelsen af de fuldstændige blokke, og derved opnås en sikrere sammenligning af de forsøgsled, som forekommer i samme forsøg. Samtidig har man dog gjort det umuligt at sammenligne forsøgsled, som ikke forekommer i samme forsøg. Derfor er en sådan opdeling af forsøget i flere mindre forsøg (kaldes af nogen serier) kun acceptabel, hvis forsøgsleddene kan deles på disse mindre forsøg, sådan at alle interessante sammenligninger m.m. kan foretages indenfor forsøget. Ved at medtage enkelte forsøgsled i alle forsøg kan man dog muliggøre sammenligning af forsøgsled i forskellige forsøg, men usikkerheden på sådanne indirekte sammenligninger er væsentlig større end sammenligning af forsøgsled i samme forsøg.

Alternativt kan man benytte forsøg med ufuldstændige blokke. Her indeholder blokkene ikke alle forsøgsled og bliver derved mindre. F.eks. kunne et forsøg med 80 forsøgsled med 3 gentagelser lægges ud i 12 ufuldstændige blokke med 20 forsøgsled i hver, eller i 30 ufuldstændige blokke med 8 forsøgsled i hver. Det betyder selvfølgelig, at hver blok kun vil indeholde en del af forsøgsleddene. Hvis forsøgsleddene fordeles tilfældigt på de ufuldstændige blokke, vil man ikke opnå en bedre sikkerhed - tværtimod, forsøget kan i værste tilfælde blive værdiløs. Fordeles forsøgsleddene på de ufuldstændige blokke så de interessante sammenligninger optimeres, kan anvendelsen af ufuldstændige blokke ofte være meget fordelagtigt.

Man skelner normalt mellem to grupper af ufuldstændige blokke:

1. I enfaktorielle forsøg anvendes oftest planer, hvor man sikrer sig, at alle par af forsøgsled kan sammenlignes med samme og (og størst mulig) sikkerhed.
2. I flerfaktorielle forsøg anvendes oftest planer, hvor man sikrer sig, at de interessante effekter (hoved eller vekselvirkninger) kan estimeres og testes med en stor sikkerhed, mens de mindre interessante (eller uinteressante) effekter så bliver estimeret med en lav sikkerhed (eller slet ikke kan estimeres).

Enfaktorielle forsøg

Ufuldstændige blokforsøg har været anvendt i mange år. Metoden går tilbage til Yates (1939), og har gennem årene vist sig at være meget effektiv (se, f.eks. Le Clerg, 1966 eller Patterson and Hunter, 1983). I litteraturen findes mange forskellige typer af ufuldstændige blokforsøg. Her vil kun blive omtalt en type: generaliserede lattice forsøg (også kaldet α -forsøg), som blev udviklet af Patterson and Williams (1976). Disse forsøg er valgt, dels fordi de er "resolvable" (de ufuldstændige blokke kan samles så de danner hele gentagelser), og dels

fordi de kan konstrueres for alle mulige antal forsøgsled og gentagelser samt for et meget bredt spektrum af blokstørrelser.

Eksempel

I sortsafprøvningen med ærter var der i 2003 34 sorter, som skulle sammenlignes. Ved anvendelse af sædvanlige fuldstændige blokforsøg vurderedes det, at blokkene ville blive for store. Derfor blev det valgt at benytte et generaliseret lattice forsøg (α -forsøg). Det blev valgt at benytte en plan med 4 gentagelser og 3 ufuldstændige blokke i hver gentagelse. To af de tre ufuldstændige blokke får 11 parceller, mens den sidste får 12 parceller (figur 1). Planen er her vist så de ufuldstændige blokke ligger i hver sin kolonne, men de kan placeres anderledes i marken, f.eks. i forlængelse af hinanden, så der bliver en kolonne for hver gentagelse.

Gentagelse 1			Gentagelse 2			Gentagelse 3			Gentagelse 4		
34	6	14	27	15	7	33	15	16	1	4	8
2	29	20	33	30	19	7	14	34	13	11	2
28	4	33	8	13	17	29	25	2	34	16	30
27	13	23	14	12	2	31	24	32	31	18	17
18	1	7	28	18	29	23	22	30	9	32	12
30	26	25	3	10	26	27	21	9	3	15	10
22	31	15	4	22	32	17	3	26	24	14	19
8	16	9	5	11	21	1	19	8	20	26	27
3	21	5	1	31	25	4	13	20	29	6	21
24	12	17	9	16	24	18	28	6	22	25	23
11	19	32	6	20	34	11	10	12	7	28	33
		10		23				5	5		

Figur 1. Fordeling af sorter på gentagelser og blokke for ærtforsøget. Hvert tal udgør en parcel og angiver sortsnummeret. Hver kolonne er en blok.

Af planen fremgår det at alle par af sorter forekommer sammen i samme blok 0, 1 eller 2 gang (der er ingen par af sorter, som forekommer sammen mere end 2 gange). Der optimale er, at alle par af sorter forekommer sammen lige mange gange (f.eks. 1), men det er i praksis sjældent opnåeligt. Det fremgår endvidere, at de ufuldstændige blokke er samlet i hele gentagelser, eksempelvis indeholder de tre blokke længst til venstre alle sorter præcis 1 gang.

Ved placering i marken kan følgende 4 trin anvendes:

1. Vælge et så ensartet areal som muligt.
2. Inddel arealet i 4 lige store områder (4 gentagelser), således at variationen i hver gentagelse er så lille som muligt (= variationen mellem de 4 gentagelser skal være så stor som muligt).
3. Inddel hver af disse 4 gentagelser i 3 ufuldstændige blokke, således at variationen i hver ufuldstændig blok bliver så lille som muligt.
4. Inddel hver ufuldstændig blok i det nødvendige antal parceller, således at variationen mellem parceller bliver så lille som muligt

Analysen af et sådant ufuldstændigt blokforsøg kan foretages med den samme model (model 1 nedenfor), som for et forsøg med fuldstændige blokke (men for computeren er det lidt mere besværligt). Normalt vil man dog opdele effekten af blokke i 2 komponenter nemlig, den som skyldes de hele gentagelser, og den som skyldes de ufuldstændige blokke inden for gentagelser (model 2). I mange tilfælde er det rimeligt at antage, at effekten af de ufuldstændige blokke er tilfældig (model 3), og i så fald vil det ufuldstændige blokforsøg altid være lige så effektivt som et forsøg med fuldstændige blokke.

- (1) $Y_{vb} = \mu + \alpha_v + \gamma_b + E_{vb}$ Effekt af blokke antages systematisk
 (2) $Y_{vrb} = \mu + \alpha_v + \beta_r + \gamma_{rb} + E_{vrb}$ Effekt af blokke indenfor gentagelser antages systematisk
 (3) $Y_{vrb} = \mu + \alpha_v + \beta_r + C_{rb} + E_{vrb}$ Effekt af blokke indenfor gentagelser antages tilfældig
 E_{vb} , E_{vrb} og C_{rb} antages uafhængige og normal fordelte med konstant varians, σ_E^2 og σ_C^2

Ved anvendelse af ufuldstændige blokke forekommer alle forsøgsled, som nævnt, ikke i alle blokke. Dette bevirker, at en sort, som har været i højt ydende blokke vil have et gennemsnit som er for højt, mens en sort, som har været i lavt ydende blokke vil have et gennemsnit, som er for lavt. Derfor kan simple gennemsnit ikke benyttes til at sammenligne sorterne. I stedet må man benytte de sortsestimater som fås ved udtrykket: $\hat{\mu} + \hat{\alpha}_v$. Man kan betragte dette udtryk som simple gennemsnit korrigeret for, om den pågældende sort har været placeret i ufuldstændige blokke, hvor udbyttet i gennemsnit har været lavere end gennemsnittet eller højere end gennemsnittet.

Effektivitet

Sortsforsøget med ærter blev udført på 4 lokaliteter. I tabel 1 er Lsd-værdierne vist for hver af disse forsøg sammen med den Lsd-værdi man ville have fået, hvis man havde anvendt sædvanlige fuldstændige blokke. Lsd-værdierne blev reduceret noget i to af forsøgene, meget i et forsøg og slet ikke i et forsøg. I tre af forsøgene har man haft en fordel af at benytte α -design. I forsøgene på lokalitet 3 og 4 kunne man have opnået den samme Lsd-værdi med sædvanlige fuldstændige blokke ved at øge antal gentagelser med 1, mens man på lokalitet 2 skulle have anvendt 4 ekstra gentagelser (d.v.s. i alt 8) for at reducere Lsd-værdien med 29%.

Tabel 1. Lsd-værdier for sammenligning af sorter på hver af de 4 lokaliteter ved anvendelse af to forsøgstyper samt opnået reduktion ved anvendelse af α -forsøg

Lokalitet	Fuldstændige blokke	α -forsøg gens. (min – max)	Relativ reduktion, %
1	4.8	4.8 (4.8-4.8)	0
2	6.9	4.9 (4.7-4.9)	29
3	6.4	6.0 (5.8-6.0)	9
4	3.7	3.4 (3.3-3.5)	8

Fordele og ulemper ved ufuldstændige blokke i enfaktorielle forsøg

Fordelene er først og fremmest at de gør det muligt at have mange forsøgsled (sorter, pesticider m.m.) i samme forsøg og stadig være i stand til at sammenligne alle forsøgsled med god sikkerhed. Fordelene er især store på marker, hvor det kan være svært at placere tilstrækkelig ensartede fuldstændige blokke. De ufuldstændige blokforsøg kan også benyttes som en slags forsikring mod uventet store variationer i marken, da de altid (med antagelsen om tilfældige blokeffekter) vil være lige så effektive som tilsvarende forsøg med fuldstændige blokke (Yates, 1940). I marken vil de ufuldstændige blokforsøg ikke kunne skelnes fra forsøg med fuldstændig blokke.

Ulemperne er, dels en lidt mere besværlig proces ved konstruktion og planlægning, og dels at alle forsøgsled sjældent vil blive sammenlignet med præcis den samme sikkerhed, idet nogle par af forsøgsled vil forekomme sammen oftere end andre par. Dette vil eksempelvis betyde, at man ved publiceringen ofte ikke blot kan anføre en Lsd-værdi, da denne vil afhænge af, hvilket sortspar man vil sammenligne. Forskellene mellem mindste og største Lsd-værdi er dog oftest kun nogle få procent (se tabel 1). Endelig skal nævnes, at ufuldstændige blokforsøg er lidt mere følsomme overfor manglende værdier samt for ombytning af forsøgsled (ved fejlsåning, forkert behandling m.m.).

Flerfaktorielle forsøg

Forsøg med flere faktorer får hurtigt mange forsøgsled. Eksempelvis vil man i et forsøg med 4 faktorer, som hver har 3 niveauer, få 81 forsøgsled. I princippet kunne her anvendes samme metode som ovenfor, men oftest vil det være mere fordelagtigt at prioritere de enkelte effekter i planen, således at man sikrer en god og sikker estimation af de interessante effekter på bekostning af en mindre sikker estimation af de mindre interessante (eller uinteressante) effekter. En effekt kan være mindre interessant, f.eks. fordi den ikke er vigtig for at besvare det spørgsmål der skal belyses, eller fordi den kun er væsentlig, hvis den giver et meget stort udslag. Også denne type af forsøg har været benyttet i mange år og går helt tilbage til Yates (1935). For forsøg, hvor alle faktorer har 2^p niveauer eller 3^p niveauer er konstruktionen af sådanne planer rimelig simpel (se f.eks. Cochran & Cox, 1957). I andre tilfælde må man enten finde planen i litteraturen eller konstruere den ved en eller anden ad hoc. metode.

Eksempel

I et forsøg med strandsvingel til frøavl ville man undersøge effekterne af følgende tre faktorer samt deres tovejs vekselvirkninger: 2 dækafgrøder, 4 niveauer af N-tildeling om efteråret og 4 niveauer af N-tildeling om foråret. Forsøget blev udlagt med 4 gentagelser. Hvis forsøget skulle udlægges som et almindeligt randomiseret blokforsøg ville der blive 32 parceller i hver

blok. Med den variation man sædvanligvis har i den type forsøg blev det anset for nødvendigt at reducere antallet ved at anvende ufuldstændige blokforsøg efter følgende principper:

- De hele gentagelser skulle underinddeles i 2 blokke á 16 parceller.
- Vekselvirkningen mellem de 3 faktorer var mindre interessant og kunne kobles til blokke.

De 32 behandlinger blev i dette forsøg delt på 2 blokke ved - fra tre-vejs vekselvirkningen - at udvælge 4 ortogonale kontraster med hver 1 frihedsgrad, og derefter koble disse til blokke i hver sin gentagelse. De effekter, der er knyttet til disse 4 kontraster (een i hver gentagelse) bliver således bestemt dårligere end de øvrige, da de nu kun kan estimeres i de 3 gentagelser, hvor de hver især ikke er koblet (figur 2).

Ved gennemgang af figuren finder man, at alle kombinationer af 1 og 2 faktorer forekommer lige mange gange i hver blok. Eksempelvis forekommer hver af de 4 niveauer af N-tildeling om foråret præcis 4 gange i hver blok. Derimod findes kombinationen "*dækafgrøde 1, N-tildeling 1 om efteråret og N-tildeling 1 om foråret*" 1 gang i blokken længst til venstre mens kombinationen "*dækafgrøde 1, N-tildeling 1 om foråret og N-tildeling 2 om efteråret*" ikke forekommer i denne blok.

Gentagelse 1	Gentagelse 2	Gentagelse 3	Gentagelse 4
232	224	143	231
221	124	222	221
111	114	244	144
233	113	133	113
144	143	142	243
244	112	112	144
144	121	121	144
241	112	113	114
232	144	121	214
131	121	124	212
131	222	131	134
214	231	134	242
143	132	232	223
133	242	121	122
213	121	241	141
122	213	232	133
122	242	114	231
223	132	112	111
241	211	123	233
142	141	131	221
123	214	234	212
124	134	211	141
114	243	112	111
221	211	214	234
111	141	211	143
234	123	222	213
224	234	213	132
113	244	241	132
212	233	223	123
142	223	142	122
243	222	233	224
134	133	243	242

Figur 2. Markplan for forsøget med strandsvingel til frøavl.
Hvert 3-cifret tal udgør en parcel og cifrene angiver den aktuelle behandling i rækkefølgen dækafgrøde, niveau af N-tildeling om efteråret og niveau af N-tildeling om foråret.
Hver kolonne er en blok.

Ved placering i marken bruges de samme principper som ovenfor. Randomiseringen er udført i 3 trin. Først er placeringen af de 4 gentagelser randomiseret, derefter er placeringen af de 2 blokke randomiseret i hver gentagelse, og til sidst er rækkefølgen af de 16 behandlinger i hver blok randomiseret.

Analysen af forsøget foretages også her med en model, som næsten er identisk med den, der anvendes for et forsøg med fuldstændige blokke.

Effektivitet

I forsøget blev der udført telemålinger. Planen forbedrede kun sikkerheden på de ukoblede effekter med ca. 3% for denne variabel (tabel 2). For de ukoblede effekter er effektiviteten, som forventet, uafhængig af om blokeffekterne antages systematiske eller tilfældige. Når blokeffekterne antages systematiske er nogle af sammenligningerne i tre-vejs tabellen, som forventet, mindre sikre ved anvendelse af ufuldstændige blokke end ved anvendelse af fuldstændige blokke. For de koblede parvise sammenligninger i tre-vejs tabellen er effektiviteten bedre når blokeffekterne kan antages at være tilfældige. I eksemplet er der imidlertid kun 3 frihedsgrader til bestemmelse af varianskomponenten for ufuldstændige blokke, hvorfor det kan være betænkeligt at anvende antagelsen om tilfældige blokeffekter ved estimation af de koblede effekter.

Tabel 2. Lsd-værdier for sammenligning af 3 effekter på telemåling ved anvendelse fuldstændige blokke, ufuldstændige blokke med systematiske blokeffekter og ufuldstændige blokke med tilfældige blokeffekter.

Effekt	Fuldstændige blokke	Systematiske blokeffekter	Tilfældige blokeffekter	
Dækafgrøde	0.0068	0.0066	0.0066	
N-niveau efterår	0.0096	0.0093	0.0093	
Dækafgrøde × N-niveau efterår	0.0135	0.0132	0.0132	
3-vejs vekselvirknings tabel	0.0270	0.0263	0.0263	48 par
	0.0270	0.0269	0.0260	384 par
	0.0270	0.0274	0.0269	64 par

Split-plot

Split-plot forsøget er et særtilfælde af ovenstående metode. Eksempelvis et forsøg, hvor man skal sammenligne effekten af 6 svampebehandlinger på 4 bygsorter i 3 gentagelser. Her har man 3 effekter at prioritere imellem: 1) svampebekæmpelse 2) sorter og 3) vekselvirkning mellem disse. Hvis formålet med forsøget er at vurdere de 6 svampebehandlinger, vil det ikke være så interessant at sammenligne hovedeffekten af sorter, og det kan vælges at koble denne til blokke. Her er benyttet et forsøg med 4 blokke á 6 parceller i hver gentagelse (figur 3). Kan blokke indenfor gentagelser antages at være tilfældig, hvad vi oftest gør, kan effekten af sorter estimeres og testes, men med en større usikkerhed, dels fordi der kun vil være 6 frihedsgrader til bestemmelse af variationen mellem blokke indenfor gentagelser, og dels fordi denne variation sikkert vil være ret stor.

Gentagelse 1				Gentagelse 2				Gentagelse 3			
4f	3f	2f	1f	2a	1c	3b	4a	3e	4e	1b	2e
4e	3e	2e	1e	2d	1f	3f	4b	3b	4b	1e	2a
4d	3d	2d	1d	2c	1q	3a	4f	3f	4c	1c	2f
4c	3c	2c	1c	2f	1d	3d	4e	3c	4d	1d	2b
4b	3b	2b	1b	2e	1e	3c	4c	3d	4a	1a	2d
4a	3a	2a	1d	2b	1b	3e	4d	3a	4f	1f	2c

Figur 3. Fordeling af sorter og behandlinger på gentagelser og blokke for bygforsøget. Tallet angiver sortsnummeret, mens bogstavet angiver behandlingen. Hver kolonne er en blok.

Effektivitet

Planens effektivitet er vurderet ved at sammenligne Lsd-værdierne med dem man ville have fået, hvis man havde anlagt forsøget som et sædvanligt randomiseret blokforsøg (tabel 3). Som det ses af tabellen, halveres Lsd-værdien for sammenligning af svampebehandlinger, hvis det er tilstrækkeligt at betragte hovedvirkningen (d.v.s. ingen vekselvirkning). Er det nødvendigt at sammenligne behandlingseffekterne ved hver enkelt sort, halveres Lsd-værdien næsten. Derimod bliver det ved anvendelse af split-plot planen meget mere usikkert at sammenligne de gennemsnitlige sorteffekter, idet Lsd-værdien for denne sammenligning stiger med næsten 70%.

Tabel 3. Lsd-værdier for sammenligning af behandlinger og sorter ved anvendelse af et randomiseret blokforsøg og det benyttede split-plot forsøg.

Effekt	Randomiseret blokforsøg	Split-plot forsøg
Svampebehandling	4.6	2.3
Sorter	3.7	6.2
Vekselvirkning	9.1	4.7 (samme sort)
	9.1	7.5 (forskellige sorter)

Fordele og ulemper ved anvendelse af ufuldstændige blokke i flerfaktorielle forsøg

Også her er den største fordel, at man kan medtage mange forsøgsled (kombinationer af flere faktorer) og stadig kunne anvende passende små blokke, og dermed undgå at variationen mellem parceller (i store blokke) bliver for stor.

Ulemperne er, at man må prioritere og dermed acceptere at visse effekter enten ikke kan estimeres eller kun kan estimeres med en stor usikkerhed. Desuden skal nævnes, at denne type ufuldstændige blokforsøg er mere følsom overfor antagelsen om at blokeffekterne er additive.

Ikke additive blokeffekter for en effekt (f.eks. hovedeffekten A) kan i visse tilfælde give falske udslag for en anden effekt (måske vekselvirkningen $B \times C$).

Konklusion

De tre eksempler viser, at der i nogle tilfælde - men ikke alle - kan opnås betydelig sikrere estimerer ved at anvende ufuldstændige blokforsøg, når der er mange forsøgsled. Hvis blokeffekterne kan antages tilfældig vil forsøg med de viste typer af ufuldstændige blokke aldrig være dårligere end tilsvarende forsøg med fuldstændige blokke. I flerfaktorielle forsøg med systematiske blokeffekter vil de ukoblede effekter aldrig være dårligere. Da der oftest kun er lidt ekstra arbejde i forbindelse med planlægningen, konstruktionen og analysen af sådanne forsøg anbefales det at overveje om man skal bruge forsøg med ufuldstændige blokke, hvis man har mange forsøgsled eller af anden grund vil få store fuldstændige blokke.

Referencer

- Cochran, W.G., and Cox, G.M., 1957. Experimental Designs. Second Edition. John Wiley & Sons, Inc. New York. 611 pp.
- Le Clerg, E.L., 1966. Significance of Experimental Design in Plant Breeding. Plant Breeding. 243-313.
- Patterson, H.D., and Williams, E.R., 1976. A new class of resolvable incomplete block designs, Biometrika 63, 83-92.
- Patterson H.D. and Hunter E.A., 1983. The efficiency of incomplete block designs in National List and Recommended List cereal variety trials. Jour. Agric. Sci. Camb. 101, 422-433.
- Yates, F., 1935. Complex experiments. Jour. Roy. Stat. Soc. Suppl. 2, 181-247.
- Yates, F., 1939. The recovery of inter-block information in variety trials arranged in three-dimensional lattices. Annals of Eugen. 9, 136-156.
- Yates, F., 1940. The recovery of inter-block information in balanced incomplete block designs. Annals of Eugenies. 10, 317-325.