Epidémiol. santé anim., 1997, 31-32

EXPERIENCES IN APPLYING MULTI-LEVEL MODELS FOR ASSESSING THE DISTRIBUTION OF ANIMAL DISEASES

McDermott J.[1,2,3], O'Callaghan C.[1,4], Kadohira M.[5]

Les modèles hiérarchiques sont particulièrement utiles lorsqu'on analyse simultanément des facteurs à l'échelle de l'animal et de l'élevage. L'importance des différents niveaux dépend du type de la maladie : les maladies à vecteur sont fortement influencées par les facteurs régionaux, les maladies de production par les facteurs de gestion du troupeau. Les sous unités ou les regroupements dans les troupeaux doivent parfois être considérés en fonction de l'objectif de l'étude. Un problème particulier dans les études à plusieurs niveaux est la définition de la maladie et des différentes catégories de facteur de risque. Souvent les facteurs d'élevage et régionaux sont confondus, par exemple les pratiques de pâturage et la pluviométrie. Les logiciels disponibles utilisés par les auteurs sont listés et leurs méthodes d'estimation sont brièvement décrites. Nous recommandons vivement d'utiliser plusieurs logiciels et comparer les résultats avant d'accepter le modèle final. L'interprétation du modèle dépend essentiellement des objectifs de l'étude. Dans les études cliniques ou dans les essais de terrain les effets fixes sont plus importants, alors que dans les études d'observation les effets aléatoires sont privilégiés. Cependant, la structure hiérarchique des effets aléatoires a une grande influence sur le modèle et doit être la plus proche du modèle naturel de l'apparition de la maladie. Pour les données d'observation, l'analyse à plusieurs niveaux est très utile pour caractériser les tendances entre les régions. A ce jour les analyses sont focalisées sur les niveaux spatiaux, il reste à développer l'approche permettant de tenir compte des structures temporelles.

BACKGROUND

The distribution of animal diseases can be influenced by factors that occur at either animal, farm, or area (environmental or farming system) levels. The relative importance of factors at different levels varies between diseases. Diseases strongly influenced by farm management such as mastitis vary most from farm-to-farm while vector-borne diseases usually vary most by factors influencing the occurrence of the vector in an area. Interactions between levels can also occur, for example when vector-density is sufficiently low that disease control practices at the farm level can have an effect.

In planning animal disease control programmes, multi-level models can help to determine the proper unit of interest(s), estimate the effects of risk factors at different levels, and to assess the success of control programs at farm, area and national levels. While these many potential applications make the use of multi-level modelling particularly attractive, there are a number of potential dangers and pitfalls at all stages: study design, sampling, data collection and analysis. In this paper we describe our experiences in considering the distribution of diseases at multiple levels while conducting field studies of tropical animal diseases. We will comment on features we have found important in conducting, analyzing and interpreting these studies.

STUDY METHODS

The choice of individual-animals and herds as important levels of interest is usually obvious. Animal age is invariably an important determinant (or at least proxy) of disease occurrence. Definitions become more difficult at the herd level in some farming systems, either because there are important subunits of herd (litters, pens) or because individual herds are managed with other herds, particularly in pastoralist systems. If litters or pens are important units of interest they and their associated risk factors should be included as subunits within herds. For larger pastoralist groups, we have found significant difference in brucellosis occurrence between individually-owned herds even if they are herded together for considerable periods (Kadohira et al., 1997a). Thus, our advice would be to always include herd as a unit of interest and to include either subunits or aggregates of herds as additional levels as appropriate.

Area and other higher units of aggregation need to be defined based on both study objectives and knowledge of disease distribution. For vector-borne diseases, agro-ecological zones (AEZ) have been useful in highlighting differences in descriptive studies. However, if the objective is to assess a disease control program, control district would be an obligatory level, perhaps refined by AEZ data. For trypanosomiasis, both micro (grazing-level) and macro (AEZ-level) determinants can be important. The precision in defining ecological levels in such cases will depend on study objectives and available resources. We have not used formal methods to define sample sizes at multiple levels and do not know any software available to do this. Our usual approach is to first estimate the relative costs and variability at different levels of interest. Often at this stage we reduce the problem to a 2-stage

[1] Dept. of Population Medicine, University of Guelph, N1G 2W1, Canada
[2] Dept. of Public Health, University of Nairobi, Box 29053, Nairobi, Kenya
[3] CIRAD-EMVT, BP 5035, 34032 Montpellier, cedex 1, France
[4] EAMG, Dept. of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK
[5] Dept. of Disease Control, University of Zambia, Box 32379, Lusaka, Zambia

(area and herd) cluster sampling problem (since for small and moderate sized herds costs of sampling animals within herds are small) for which formula exist (Cochran, 1977).
An important problem that must be considered in multi-level studies is the difficulty of consistent diagnosis (e.g. variations in specificity of diagnostic tests) and risk factor measurement across areas. In addition, categories of important risk factor may not be available in all areas and important herd risk factors may be confounded by area factors. For example Fig. 1 displays the confounding between farm size and grazing on IBR farm sero-prevalence in 3 districts of Kenya. Grazed herds were larger and had higher and less variable IBR sero-prevalence than zero-grazed herds However this standard epidemiological problem can be further complicated in multi-level studies by area factors. For example, in this data, large (>30 cattle) herds were only found in one district and small herds dominated in the other 2. In the presence of such area and herd confounding it is very difficult to attribute changes in sero-prevalence to different risk factors at different levels.

SOFTWARE OPTIONS

We have listed in Table I some software that we have used in multi-level analysis. Both EGRET and GEE cannot be used for more than 2 levels of clustering but have been included because we like to compare their results to the 2-level results of the other multi-level programmes. Other published multi-level analyses have been conducted in the human health field. The best known to us is the work of Katz et al. (1993) who analyzed their data using a hierarchical logistic regression model (alternating logistic regression; Carey et al., 1993). Most others have used either the software listed or written programmes of there own.
Because there are many parameters to be estimated, maximum likelihood estimation of specific mixture distributions require intensive computer resources. Given that the exact distribution is rarely known the sensible tendency has been to develop algorithms using approximate method such as penalized quasi-likelihood or pseudo-likelihood approaches which are thought to be robust. In our experience, the different methods listed agree relatively well but given the large number of parameters that need to be estimated, often sparse data for some categories (particularly at higher levels), and the relatively recent development of most software we think it is very important to compare and check results from a number of different 2-level and multi-level approaches for consistency.

Table I
Some software used to analyze variations in disease distribution at 2 or more levels

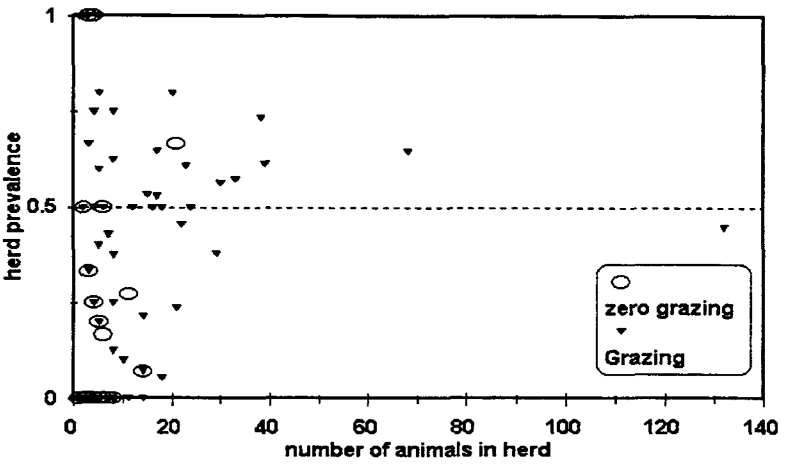
Software	Levels	Estimation procedure	Features
EGRET	2	ML estimation of mixture distributions (e.g. Beta-binomial, logistic normal)	Good graphics and residual plots Likelihood tests between all models
SAS[a] GEE (IML)	2	Solving of GEE Score functions (Karim and Zeger, 1986)	Robust and fast but currently not programmed for >2 levels
Schall (IML)	n	Penaized quasi-likelihood (PQL) estimation (Schall, 1991)	Fast and easily implemented. No graphics or tests for random-effects
Proc Mixed	n	GLMM pseudo-likelihood approach (Wolfinger and O'Connel, 1993)	Available within SAS Proc Mixed
Min	n	PQL estimation or ML for fixed-effects (Goldstein, 1995)	Multi-level residuals plots and test of simple and complex random-effects

[a] Can also be programmed in other packages such as S-Plus (Chambers and Hastie, 1993). Additionally, methods such as the jacknife and Gibb's sampling procedures could also be programmed.

MODEL RESULTS AND INTERPRETATION

In any mixed (fixed and random) effect multi-level modelling, the parameters to be estimated and model interpretation will depend on the objectives. If the objective is to assess a disease control program delivered at multiple levels, then the fixed program effects will be of primary interest and random effects for the levels of secondary interest. In observational studies of disease patterns and risk factors, particularly in descriptive phases, random effects will be of primary or at least equal interest. In our experience in analyzing observational data, we have noted that both fixed and random parameter estimates vary depending on what other fixed and random parameters are in the model (Kadohira et al., 1997b) and, even more importantly, on the hierarchical structure of random effects chosen (McDermott et al., 1997). In choosing a hierarchical structure for random effects, it is important to explore whether single or multiple farm (or area) variance components should be fit. For example, in Fig.1 it is obvious that the farm-level variance of IBR sero-prevalence is a function of farm size. A constant variance component for farm poorly model the data. Adding variance components for farm size classes greatly improves the fit but the best fit is obtained by having separate farm variance components for each district (production system). The correlation of farm size with grazing (and other production system descriptors) shows why this random-effect structure fits better, and why it is important to consider the pattern of IBR farm sero-prevalence within a specific production system. Thus, in addition to the usual regression problems of correctly measuring and including important fixed effects, it is crucial in multi-level models to develop a random-effects hierarchy which reflects the natural pattern of disease occurrence.

Figure 1
Variability of IBR sero-prevalence by farms and grazing pattern



We have found that multi-level analyses can be very helpful in describing broad patterns of disease, particularly in characterizing high risk units and risk factors. Once high risk ecological or production systems are defined, it is more profitable to reduce the geographic scale covered and concentrate on herd and animal level analyses in these areas. In this regard, we have noted that serological profiles by age can differ markedly across areas and farm types and are particularly useful in describing the force of infection. Also as noted above, farm clustering of brucellosis varies considerably by production system. In smallholder systems without communal grazing the incidence of brucellosis is low and its spread from farm-to-farm is essentially random. But in pastoralist systems, brucellosis incidence is higher and there is considerable clustering by herd, even those grazing together. There must be herd-level risk factors such as propensity to exchange or purchase animals which contribute to the large herd-to-herd differences seen, but we have yet to measure herd risk factors with sufficient precision in pastoral areas to explore this.

Our experience to date has been in analyzing disease patterns across multiple spatial levels of organization using prevalence data or incidence data reduced to the occurrence of new infections within a given time period. However, refining the temporal aspects of disease incidence within a multi-level spatial context is important to improve models of epidemic infectious diseases. We see this as an important area for the future development of multi-level models.

REFERENCES

Carey V., Zeger S., Diggle P., 1993. Modelling multivariate binary data with alternating logistic regressions. Biometrika, 80, 517-526.

Chambers J., Hastie T., 1993. Statistical models in S. Chapman and Hall, New York, 608 pp.

Cochran W., 1977. Sampling techniques. 3rd ed. Wiley, New York, 428 pp.

Goldstein H., 1995. Mutilevel Statistical Models. 2nd ed. Arnold, London, 192 pp.

Kadohira M., McDermott J., Shoukri M., Kyule M., 1997a. Variations in the prevalence of antibody to brucella infection in cattle by farm, area and district in Kenya. Epidemiology and Infection, 118: 35-41.

Kadohira M., McDermott J., Shoukri M., Thorburn M., 1997b. Assessing infections at multiple levels of aggregations. Preventive Veterinary Medicine, 29(3), 161-177.

Karim, M.R., Zeger, S., 1986. GEE, A SAS macro for longitudinal data analysis. Tech. Rept. 674. Dept. of Biost., Johns Hopkins University, Baltimore MD, USA, 11pp.

Katz J., Carey V., Zeger S., Sommer A., 1993. Estimation and design effects and diarrhoea clustering within households and villages. American Journal of Epidemiology, 138: 994-1006.

McDermott J., Kadohira M., O'Callaghan C., Shourki M., 1997. A comparison of different models for assessing variations in the sero-prevalence of IBR by farm, area and district in Kenya. Prev.Vet.Medicine., in press.

Schall R., 1991. Estimation in generalized linear models withe random effects. Biometrika, 78, 719-727.

Wolfinger R., O'Connell M., 1993. Generalized linear mixed models: a pseudo-likelihood approach. Journal of Statistical Computation and Simulation, 48: 233-243.