TECHNICAL
REPORT

0313

Correcting for inter-observer effects in a geographical oral health study

Samuel M. Mwalili, Emmanuel Lesaffre and Dominique Declerck

IAP STATISTICS
NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Abstract: In this paper, we present an approach for correcting for inter-observer systematic differences in a geographical oral health study. The scoring variability of different examiners complicated the identification of a geographical trend in a recent study on caries experience in Flemish children (Belgium). Classically the factor "examiner" would be included in the regression model to correct for its confounding effect. However when applied here, the geographical trend was removed. Instead, we modified the (logistic) regression model by introducing the conditional classification probabilities of a score by an examiner given a score by a 'gold standard'. These probabilities were estimated from a calibration exercise. Our model was fitted on data from the Signal-Tandmobiel® study using the SAS procedure NLMIXED and the WINBUGS program.

Keywords: calibration exercise; inter-observer agreement; systematic bias;

1 A geographical oral health data analysis for caries experience

The Signal-Tandmobiel® study is a 6 year longitudinal oral health study started in Flanders (Belgium) in 1996 involving 4468 children. Data were collected on oral hygiene level, gingival condition, dental trauma, prevalence and extent of enamel developmental defects, fluorosis, tooth decay, presence of restorations, missing teeth, stage of tooth eruption and orthodontic treatment need, all using established criteria. The children were examined annually for a period of six years (1996-2001). Their average age at entry was 7.1 years ($sd = 0.4$). In our paper the measurement of interest is the *dmft* index, which is the sum of the number of decayed (d), missing due to caries (m) and filled (f) teeth. The *dmft* index is a measure for the so called caries experience. Sixteen dental examiners were involved. They were trained and examination methods were calibrated at baseline and re-calibrated half yearly.

From a dental point of view it is of interest to examine the geographical trend in caries prevalence caries experience in Flanders, *see* e.g. Declerck, Lesaffre, Mwalili and Vanobbergen (2002). For our geographical analysis we have taken the data of the first year of the study, and hence the Signal-Tandmobiel® study will be referred to below as the cross-sectional dental study. Further, an ordinal response of caries experience for child i, scored by examiner j at school k was constructed from the *dmft*-score as follows:

$$y_i = \begin{cases} 0 & \text{if the } dmft \text{ score for the } i\text{th child is } 0 \ , \\ 1 & \text{if the } dmft \text{ score for the } i\text{th child is } 1 \ , \\ 2 & \text{if the } dmft \text{ score for the } i\text{th child is in } (1,4] \ , \\ 3 & \text{if the } dmft \text{ score for the } i\text{th child is in } (4,20] \ . \end{cases} \tag{1}$$

When this ordinal score is obtained from the jth examiner, the notation y_{ij} will be used below when appropriate. This ordinal score will be referred to below as the "degree of caries experience". The categorization of the score was motivated by the skewed nature of data, lack of possible near normality transformation, and primarily because of the easy interpretation of ordinal model in dentistry. It was, therefore, more sensible to fit the ordinalized response than the original data.

In a preliminary analysis, we applied an ordinal logistic regression model with a random school intercept. The geographical trend in the degree of caries experience was examined in 2 ways: (1) by including a dummy for the four of the five provinces of Flanders (Antwerp as reference class); (2) by including the $x-$ and $y-$coordinate of the municipality of the school to which the child belongs. However, this is not a spatial model by itself as we do not incorporate either geostatistical or lattice modeling components. Here the $x-y$ coordinates or the regions, as explained above, take the geographical components into account.

Additionally age and gender were included as covariates. The most popular ordinal regression model, with logit link, is the *cumulative logit* model. A random effect version has the expression, *see* e.g. Hartzel, Agresti and Caffo (2001):

$$log\left(\frac{\pi_{ik1}+\cdots+\pi_{ikr}}{\pi_{ik,r+1}+\cdots+\pi_{ik4}}\right)=\lambda_r+\boldsymbol{x}_i'\boldsymbol{\beta}+u_k, \qquad r=1,2,3 \tag{2}$$

where $\boldsymbol{x}_i$ is a d-dimensional vector of covariates pertaining to the ith child and $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients (fixed effects). It is assumed here that the effect of covariates is the same for all logits. This is called the *proportional odds* assumption. π_{ikr} is the probability of child i in school k being classified in category r of the ordinal caries response. Further the random intercept u_k pertains to the kth school and we assume that $u_k \sim N(0, \sigma^2)$. λ_1 is the intercept and (λ_2, λ_3) are the ordered category cut-off parameters, which satisfy $\lambda_2 < \lambda_3$. Below the vector $(\lambda_1, \lambda_2, \lambda_3)'$ will be denoted as $\boldsymbol{\lambda}$.

Model (2) was applied to the cross-sectional caries experience data, the results are shown in Table 1. The point estimates and the standard errors of the category cut-offs are practically the same in the two geographical models. Further, the two geographical models gave similar estimates for the regression coefficient of age and of gender. More importantly, the results clearly indicate a significant East-West gradient in the degree of caries experience, being higher in the province of Limburg (*see also* Figure 1).

[TABLE 1 ABOUT HERE]

However, there were 16 examiners involved in this study. Despite the half-yearly calibration exercises (for scoring different aspects) they still differed in their assessment of caries experience. Further, in Figure 1 it is clearly seen that each examiner was active in a relatively restricted geographical area. Thus, a legitimate question was whether the geographical trend in the degree of caries experience was due to the different scoring behaviour of the examiners. In other words, was the East-West trend due to biased scoring of some examiners (compared to a gold standard) or due to a genuine geographical effect? Here, the gold standard is an experienced dentist, who is assumed to be error-free for measuring the dmft score.

[FIGURE 1 ABOUT HERE]

A classical way to take a confounder into account is to include it in the (logistic) regression model. Adding "examiner" into the ordinal logistic model (2) as a fixed effect gave the estimates shown in Table 2. Clearly, controlling for examiner removed the geographical East-West trend. The same conclusion could be drawn when the examiner was included in the model as a random effect (results not shown). We argue that correcting for examiner in this way is not appropriate because it does not take into account the scoring variability of the examiners.

[TABLE 2 ABOUT HERE]

On the other hand, model (2) assumes that the probability of scoring k on y is the same for all examiners, and hence ignores possible different scoring behaviour of the examiners. However, it became evident during the conduct of the Signal-Tandmobiel®

study that some examiners consistently over- and other examiners consistently under-scored the degree of caries experience with respect to the gold standard. To properly take the examiners' effect into account, we opted for an *external correction* using the calibration data. This resulted in an estimated model mimicking the case where all children had been scored for caries experience by only one examiner (the gold standard).

2 Taking into account examiner's effect

2.1 Calibration exercises

Much attention was paid to the selection and training of the 16 dentist-examiners for the project. In order to maintain a high level of intra- and inter-examiner reliability, *calibration* exercises were carried out twice a year for all examiners involved. During the study period (1996-2001) four of these calibration exercises were devoted to scoring of caries experience. A minimum of 12 children was included in each exercise. The children selected for this exercise were screened a priori to ensure that a variety of pathologies was present, including untreated caries, recurrent caries and fillings. Additionally a "gold standard" was present. The scores on caries experience of each of the 16 dental examiners were compared with the scores obtained by the gold standard. Due to the relatively small number of children used per calibration exercise we combined the data from at most four caries calibration exercises. Observe that by pooling the data of the calibration exercises we actually underestimated the possible systematic bias of the examiners since it would be expected that the examiners became better calibrated in due time.

[TABLE 3 ABOUT HERE]

Table 3 shows the classification matrices and conditional probabilities for the examiners 12, 13, and 16, which had mild, extreme and perfect classifications respectively. The conditional classification probabilities were estimated from the classification matrix $\boldsymbol{M}_j$ as explained in Section 2.4.2. The closeness of the diagonal elements of the conditional classification matrix for examiner j to 1 implies perfect classification of the examiner j against the gold standard.

2.2 A classical way to take inter-observer disagreement into account

A classical way to express the difference between the gold standard and the 16 examiners is to show a measure of agreement like the weighted kappa(κ_w) (Agresti 1990, p. 367).

$$\kappa_w = \frac{\sum_a \sum_b w_{ab}\pi_{ab} - \sum_a \sum_b w_{ab}\pi_{i+}\pi_{+b}}{1 - \sum_a \sum_b w_{ab}\pi_{a+}\pi_{+b}}$$

where $\pi_{a+} = \sum_a \pi_{ab}$, $\pi_{+b} = \sum_b \pi_{ab}$ and the weights $w_{ab} = 1 - (a-b)^2/(I-1)^2$; $a, b = 1, \cdots, 4$; $I = 4$. The weighted kappas for the 16 examiners using the pooled calibration data of at most four caries calibration exercises are given in Table 4.

[TABLE 4 ABOUT HERE]

Based on the scheme of agreement levels proposed by Landis and Koch (1977) all examiners had an excellent agreement with the gold standard (κ_w above 0.80) except for the examiners 1, 3 and 13 who had "only" a substantial agreement (κ_w between 0.60 and 0.79). Note that the upper bounds of the estimated 95% confidence intervals from the SAS procedure FREQ for some kappas are greater than 1 because of the asymptotic approximations. However, Kappa is an omnibus index of agreement, i.e. it does not make distinctions among various types and sources of disagreement. Further, the kappas do not help us in evaluating whether or not the geographical trend exists, when taking the possible systematic bias of the examiners into account. We will now show how the results of the calibration exercise can be used directly to correct for possible examiner bias.

In this paper, we propose both Bayesian and Frequentist approaches to the ordinal response measured with error problem. The Bayesian approach uses the posterior distribution of the misclassification parameters whereas the Frequentist approach uses the sample estimates of the misclassification parameters. In this analysis, we assume a non-differential measurement error, that is, conditional on the covariates, the true response is independent of the observed response. Our goal is, therefore, to present an approach for correcting for errors in the misclassification of the ordinal response. This model derives from the probability model of the true response given the observed response, which is based on the external data from a calibration study. Both the true outcome ($\tilde{\mathbf{z}}$) and the proxy outcome ($\mathbf{z}$) are recorded in the calibration study, while on the main study only the proxy or surrogate measures of the response ($\mathbf{y}$) are available for each child. In this paper, we are indeed correcting for the errors in the ordinal response measured by examiners with different scoring variability.

2.3 A frequentist approach

Model (2) and a model expressing the score of the jth examiner as a function of the score of the gold-standard can be combined to yield a prediction model for the gold standard. The first model is based on the prevalence of caries data from the cross-sectional dental study, while the second model will be based on the calibration data. For clarity we wish to distinguish the scores assigned in the calibration exercises, from the scores given in the cross-sectional dental study. Let z_{ij} be the ordinal *dmft* score (discretized as in equation (1)) given by examiner j to the ith child in the calibration exercise, and $\tilde{z}_{ij}$ be the corresponding score given by the gold standard.

Further, let m_{jab} be the number of times the jth examiner scores $z_{ij} = a$ and the gold standard scores $z_{ij} = b$. For the examiner j, the numbers m_{jab} are collected in a 4×4 matrix $\boldsymbol{M}_j = (m_{jab})$ $a, b = 1, \cdots, 4$, here referred to as the classification matrix. Further, let $\boldsymbol{\gamma}_j = (\boldsymbol{\gamma}_{jab})$, $a, b = 1, \cdots, 4$ be the corresponding matrix of the conditional classification probabilities, where γ_{jab} denote the conditional probability of classifying a discretized *dmft* score in the ath category by examiner j given it is classified in the bth category by the gold standard. Hence sample estimates of $\boldsymbol{\gamma}_{jab}$ equals $m_{jab}/\sum\limits_{d} m_{jad}$. It satisfies $\sum\limits_{b} \gamma_{jab} = 1$, that is, the column total sum to 1. Similarly, y_{ij} and $\tilde{y}_{ij}$ denote the score by the jth examiner and the corresponding score by the gold standard on the cross-sectional dental study. Then the probability that the ith child from school k is scored as a by the jth examiner can be written as a function of γ_{jab}'s and the probability model of the gold standard, namely,

$$\begin{aligned}
\Pr(y_{ij} = a|\boldsymbol{x}_i,\, u_k) &= \Pr(y_{ij} = a|\tilde{y}_{ij} = a)\Pr(\tilde{y}_{ij} = a|\boldsymbol{x}_i,\, u_k) + \\
&\quad \Pr(y_{ij} = a|\tilde{y}_{ij} \neq a)\Pr(\tilde{y}_{ij} \neq a|\boldsymbol{x}_i,\, u_k) \\
&= \Pr(y_{ij} = a|\tilde{y}_{ij} = a)\Pr(\tilde{y}_{ij} = a|\boldsymbol{x}_i,\, u_k) + \\
&\quad \sum_{b \neq a} \Pr(y_{ij} = a|\tilde{y}_{ij} = b)\Pr(\tilde{y}_{ij} = b|\boldsymbol{x}_i,\, u_k).
\end{aligned} \tag{3}$$

In (3) we assume that the probabilities $\Pr(y_{ij} = a|\tilde{y}_{ij} = b)$ do not depend on the covariates and the school. We further assume that $\Pr(y_{ij} = a|\tilde{y}_{ij} = a') = \Pr(z_{ij} = a|\tilde{z}_{ij} = a')$. Applying the latter assumption to equation (3) we obtain:

$$\begin{aligned}\Pr(y_{ij} = a|\boldsymbol{x}_i,\, u_k) = {} & \Pr(z_{ij} = a|\tilde{z}_{ij} = a)\Pr(\tilde{y}_{ij} = a|\boldsymbol{x}_i,\, u_k) + \\ & \sum_{b\neq a}\Pr(z_{ij} = a|\tilde{z}_{ij} = b)\Pr(\tilde{y}_{ij} = b|\boldsymbol{x}_i,\, u_k). \end{aligned} \tag{4}$$

Inserting γ_{jab}'s then (4) becomes

$$\begin{aligned}\Pr(y_{ij}{=}\, a|\boldsymbol{\gamma},\, \boldsymbol{x}_i,\, u_k) &= \gamma_{jaa}\Pr(\tilde{y}_{ij} = a|\boldsymbol{x}_i,\, u_k) + \sum_{b\neq a}\gamma_{jab}\Pr(\tilde{y}_{ij} = b|\boldsymbol{x}_i,\, u_k) \\ &= \gamma_{jaa}q_{ikr} + \sum_{b\neq a}\gamma_{jab}q_{ikh} \\ &= \textstyle\sum_{d=1}^{4}\gamma_{jad}q_{ikd}, \end{aligned} \tag{5}$$

with $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_1,\, \boldsymbol{\gamma}_2,\, \cdots,\, \gamma_{16}\}$ and

$$q'_{ik} = \begin{pmatrix} F(\lambda_1 + \boldsymbol{x}'_i\boldsymbol{\beta} + u_k) \\ F(\lambda_2 + \boldsymbol{x}'_i\boldsymbol{\beta} + u_k) - F(\lambda_1 + \boldsymbol{x}'_i\boldsymbol{\beta} + u_k) \\ F(\lambda_3 + \boldsymbol{x}'_i\boldsymbol{\beta} + u_k) - F(\lambda_2 + \boldsymbol{x}'_i\boldsymbol{\beta} + u_k) \\ 1 \qquad - F(\lambda_3 + \boldsymbol{x}'_i\boldsymbol{\beta} + u_k) \end{pmatrix}.$$

From equation (5) it follows that

$$\Pr(y_{ij} \leq a|\boldsymbol{\gamma},\, \boldsymbol{x}_i,\, u_k) = \sum_{c=1}^{a}\sum_{d=1}^{4}\gamma_{jcd}q_{ikd}, \tag{6}$$

holds, where the dependence in (5) and (6) on $\boldsymbol{\gamma}$ is highlighted. It also follows that (6) yields random effects logistic model (2) when all examiners would score exactly like the gold standard.

The SAS procedure NLMIXED (SAS Institute Inc.) can be used to estimate the unknown parameters $(\boldsymbol{\lambda},\, \boldsymbol{\beta},\, \sigma^2)$ of model (6) when for $\boldsymbol{\gamma}$ an estimated value is imputed, say from the calibration exercise. Indeed, the SAS procedure NLMIXED can fit generalized linear mixed models by maximizing a marginal likelihood. However, this approach does not take into the account the uncertainty with which the $\boldsymbol{\gamma}_j's$ are estimated. This can be achieved by the following Bayesian approach.

2.4 A Bayesian approach

2.4.1 Likelihood and prior for the cross-sectional data

The likelihood for the cross-sectional data if all caries experience scores were obtained from the gold standard is obtained from model (5). We denote model (5) also as $f(\boldsymbol{y}|\gamma,\, \boldsymbol{\beta},\, \boldsymbol{\lambda},\, \sigma^2)$, where $\boldsymbol{y}$ is the total vector of ordinal caries experience responses over all children. In a Bayesian context the likelihood needs to be combined with a prior distribution of the parameters. Here, we combined a vague normal prior for β_s $(s = 1, \cdots, d)$, i.e. $\beta_s \sim N(0,\, 10^{-6})$ $(s = 1, \cdots, d)$ with a vague inverse-gamma prior for σ^2, i.e, $\sigma^2 \sim IG(10^{-3},\, 10^{-3})$. Further, we assume a truncated normal prior for the category cutoffs, i.e. $\lambda_2 \sim N(0,\, 10^{-6})I(,\, \lambda_3)$ and $\lambda_3 \sim N(0,\, 10^{-6})I(\lambda_2,\,)$, and a vague normal prior for λ_1 , i.e. $\lambda_1 \sim N(0,\, 10^{-6})$. No prior distribution for $\boldsymbol{\gamma}$ is given here, instead we specify the distributional aspects of $\boldsymbol{\gamma}$ using the calibration data.

2.4.2 Likelihood and prior for the calibration data

The classification matrix $\boldsymbol{M}_j$ from the calibration data provides information for estimating $\boldsymbol{\gamma}$. We assume that the distribution of the (total vector) of the calibration data $\boldsymbol{z}$, $f(\boldsymbol{z}|\boldsymbol{\gamma})$, is obtained from

$$M_{j(b)} \sim \mathrm{Multinomial}\big(\mathrm{m_{j+b}}, \mathrm{w_j v_b} \,+\, (1-\mathrm{w_j})\boldsymbol{\gamma}_{\mathrm{j(b)}}\big) \tag{7}$$

where A(b) is the bth column of matrix A, $m_{ja+} = \sum_{a=1}^{4} m_{jab}$, w_j is an examiner specific random deviate taking values in $[0,1]$ and v_b is a row vector of size 4 with the bth element equal to 1 and 0 otherwise, e.g., v_3=(0,0,1,0). This ensures that the multinomial probability sums to 1. Observe that model (7) locates each examiner in-between the gold standard (w_j) and the average score ($\boldsymbol{\gamma}_{j(h)}$).

Further, each of the w_j are assumed to have a prior Beta(η_{j}, ν_{j}) distribution. We assign hyperprior distribution to (η_j, ν_j) on a uniform grid in the range $[-2.5, -0.5] \times [0.5, 3.5]$ which is centered at (-1.5, 2.0), that is, $(\eta_j, \nu_j) \approx (1.3, 6.0)$. This grid ensures that all forms of shape of the Beta densities are accomodated. Finally, a Dirichlet prior with parameters (1,1, 1, 1) is taken as prior for $\boldsymbol{\gamma}_{j(h)}$. As a result the posterior distribution p($\boldsymbol{\gamma}|\boldsymbol{z}$) is obtained.

2.4.3 Calculating the posterior distribution p($\boldsymbol{\lambda}$, $\boldsymbol{\beta}$, $\sigma^2|\boldsymbol{y}$, $\boldsymbol{z}$)

For a given $\tilde{\boldsymbol{\gamma}}$ the Bayesian analysis on the cross-sectional data yield p($\boldsymbol{\lambda}$, $\boldsymbol{\beta}$, $\sigma^2|\boldsymbol{y}$, $\tilde{\gamma}$), and the posterior estimates obtained by WINBUGS (Spiegelhalter *et al.*) are conditional on the imputed value for $\tilde{\boldsymbol{\gamma}}$. On the other hand, the calibration data result in the posterior distribution p($\boldsymbol{\gamma}|\boldsymbol{z}$), where $\boldsymbol{z}$ is the vector of ordinal caries experience responses over all children in the calibration exercise. This posterior distribution could be used as a prior distribution for $\boldsymbol{\gamma}$ in the Bayesian analysis on the cross-sectional data. However, it was not immediately clear how to do this in an elegant way using WINBUGS. Instead, we opted to process the cross-sectional data and the calibration data simultaneously. That is, at each iteration of the Markov Chain of the calibration data we obtained an estimate of $\boldsymbol{\gamma}$ and imputed this estimate into the Markov Chain pertaining to the cross-sectional data. This procedure enabled us in a simple way to take into account the variability with which $\boldsymbol{\gamma}$ is estimated from the calibration data. In fact our procedure samples from

$$\mathrm{p}(\boldsymbol{\lambda},\, \boldsymbol{\beta},\, \sigma^2|\boldsymbol{y},\, \boldsymbol{z}) = \int \mathrm{p}(\boldsymbol{\lambda},\, \boldsymbol{\beta},\, \sigma^2|\boldsymbol{y},\, \gamma) \cdot \mathrm{p}(\gamma|\boldsymbol{z})\mathrm{d}\gamma.$$

3. Application to the Signal-Tandmobiel$^{\circledR}$ study

3.1 Frequentist approach

The results of fitting the corrected multinomial logit model (6) by the SAS procedure NLMIXED are displayed in Table 5. We used the adaptive Gaussian quadrature procedure and the dual quasi-Newton optimization technique to estimate the parameters. There is a small to mild effect on the estimated regression coefficients of the provincial terms after external correction compared to those of Table 1. Further, all the standard errors slightly increased.

[TABLE 5 ABOUT HERE]

The East-West gradient in the degree of caries experience is somewhat more pronounced in this model. However, due to an increase in standard errors, the province Limburg does not have a significantly higher degree of caries experience any more than the other provinces. The geographical model in terms of the $x-$ and $y-$coordinate still shows a significant East-West trend. We must realize, though, that this (Frequentist) approach does not take into account the sampling variability of conditional classification probability estimates. Therefore, this is perhaps a too naive approach for estimating the parameters of the corrected model. To get more realistic estimates of the variability of the parameter estimates, one needs to incorporate the uncertainty over the estimates of the conditional classification probabilities $\boldsymbol{\gamma}_j's$.

3.2 Bayesian approach

We used WINBUGS to simultaneously estimate the parameters of model (6) and (7), respectively. The estimates of $\boldsymbol{\gamma}_j's$ are used in estimating the parameters of model (6) by parallel processing the calibration data and the cross-sectional data. This is accomplished by having two independent MCMC loops for these models in one run. Namely one for the calibration data, sampling the conditional classification probabilities $\boldsymbol{\gamma}$, and another one where the parameters $\boldsymbol{\lambda}$, $\boldsymbol{\beta}$, and σ^2 are sampled employing the sampled $\boldsymbol{\gamma}$ from the first loop (*See* Appendix for the WINBUGS program). Five initially overdispersed chains were initiated, with a burn-in sample of 6000 iterations and an extra 9000 MCMC iterations run for each chain. The posterior summary statistics of the regression coefficients from the Bayesian procedure are shown in Table 6. The Gelman & Rubin shrinkage factors quickly drop to 1 for all geographical coefficients, suggesting convergence of the posterior density (*see* Best *et al.*).

[TABLE 6 ABOUT HERE]

The estimated geographical regression coefficients are slightly larger in absolute value compared to the estimated regression coefficients from NLMIXED, but overall the estimates of the NLMIXED and WINBUGS are in line, given their estimated variability. Further, a sensitivity analysis by e.g changing the prior distribution of the regression coefficients from normal to t-distribution with 4 degrees of freedom gave practically the same results. However, the standard errors of the estimated regression coefficients are somewhat more increased, due to the variability with which $\boldsymbol{\gamma}$ is estimated from the calibration data. But more importantly, the East-West gradient remains important in both geographical models.

We also used the Bayesian approach to estimate the expected values of the random deviates $w's$ governing the average precision of each examiner with respect to the gold standard.

[TABLE 7 ABOUT HERE]

Table 7 shows the posterior summary statistics of the w's . Our results largely confirm the conclusion obtained from the kappa statistics.

4 Discussion

In this paper, we have shown that a random effect logistic regression model (2) can be modified to a gold standard model (6), which accommodates the examiners systematic bias to the (discretized version of) $dmft$ score. This model confirmed the East-West gradient in the degree of caries experience in Flanders. Various other models were fitted to check the confounding effect of other covariates (*see also* Declerck et al., 2002). Models (2) and (6) were also considered without the random school component, but we observed almost no difference in the geographical components, compared to the models fitted in this paper. Further, the same East-West gradient was obtained for the prevalence of caries (*see* Declerck et al., 2002).

Finally, our approach assumes a gold standard. It remains to be investigated how the external correction would work if no gold standard were available. This issue forms the basis of future research.

Acknowledgements

This work is partially funded by research grant OT/00/35 of the Catholic University Leuven. We thank Tommi T. Härkänen for his input in this work and Jacky Vanobbergen for preparing the calibration data sets.

References:

Agresti, A. (1990).*Categorical Data Analysis.* New York: Wiley

Best, N., Cowles, K. M. and Vines, K. (1996) *Convergence Diagnosis and Output Analysis Software for Gibbs sampling output Version 0.30.* Cambridge, UK.

Carrol, R.J, Ruppert, D. and Stefanski, L.A (1995) *Non-linear Measurement Error Models.* London: Chapman and Hall.

Carrol, R. J., Spiegelman, C. H., Lan, K. K. G., Bailey, K. T. and Abbott, R. D. On errors-in-the variables for binaryr rergession model. *Biometrika* **71** : 643 − 648 (1984).

Declerck, D., Lesaffre, E., Mwalili, S. and Vanobbergen, J. (2002) Geographical variations in caries experience in the primary dentition of Flemish children. (In preparation)

Gelman, A., Carlin, B. J., Stern, S. H. and Rubin, B. D. (1995) *Bayesian Data Analysis.*

Chapman and Hall, London.

Hartzel, J., Agresti, A. and Caffo, B. (2001) Multinomial logit random effects models. *Statistical Modelling* **1** : 81 − 102.

Landis, J. R. and Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics* **33** : 159 − 174.

Liu, X. and Liang, K. J. (1990) Adjustment for non-differential misclassification error in generalized linear models. *Statistics in Medicine* **10**:1197-1211

Richardson, S. and Gilks, W. R. (1993a) Conditional independence models for epidemiological studies with covariate measured with error. *Statistics in Medicine* **12**:1703-1722

Richardson, S. and Gilks, W. R. (1993b) A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology* **138**:430-432

SAS Institute Inc. ©1999-2001, *The SAS System for Windows.* Cary, NC, USA.

Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996) *Bayesian inference Using Gibbs Sampling Manual (version ii).* Cambridge, UK.

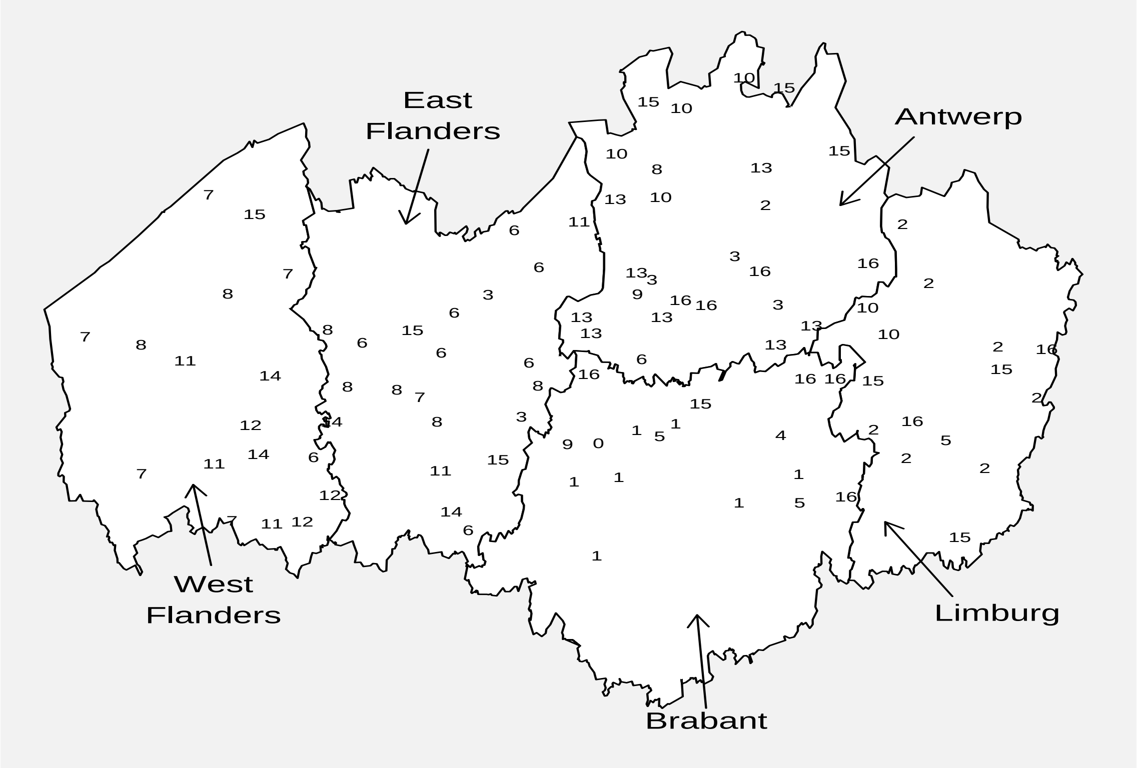

Figure 1. Plot showing the distribution of the examiners examining at the different schools (numbers indicate examiners that scored the children of a particular school located at the corresponding x-y co-ordinates) in the five provinces of Flanders.

Table 1. Parameter estimates from the random-intercept multinomial logit (2) predicting the degree of caries experience, controlling for geographical effects fitted by the SAS procedure NLMIXED.

	Provinces				x-y coordinates			
	Estimate	Conf. Limits			Estimate	Conf. Limits		
Parameter	(SE)	Lower	Upper	P-value	(SE)	Lower	Upper	P-value
λ_1	-0.522(1.138)	-2.768	1.724	0.6473	-0.576(1.245)	-3.033	1.881	0.6441
λ_2	0.453(0.045)	0.365	0.541	$< .0001$	0.452(0.045)	0.364	0.540	$< .0001$
λ_3	1.587(0.085)	1.420	1.755	$< .0001$	1.587(0.085)	1.419	1.7546	$< .0001$
Brabant	0.170(0.206)	-0.237	0.577	0.4116				
Limburg	0.412(0.200)	0.017	0.807	0.0409				
E_Fland	0.148(0.185)	-0.218	0.514	0.4267				
W_Fland	-0.295(0.202)	-0.693	0.103	0.1455				
x-coordinate					0.004(0.001)	0.001	0.006	0.0029
y-coordinate					-0.003(0.004)	-0.010	0.004	0.4140
Gender	0.033(0.127)	-0.218	0.283	0.7980	0.032(0.127)	-0.219	0.282	0.8047
Age	0.090(0.159)	-0.224	0.403	0.5725	0.112(0.158)	-0.199	0.424	0.4774
σ^2 †	0.035(0.071)				0.039(0.075)			

† Confidence limits are omitted because negative values for the 95% confidence lower limits are reported by SAS

Table 2. Parameter estimates from the random-intercept multinomial logit (2) predicting the degree of caries experience, controlling for geographical and examiners' effect fitted by the SAS procedure NLMIXED.

	Provinces				x-y coordinates			
	Estimate	Conf. Limits			Estimate	Conf. Limits		
Parameter	(SE)	Lower	Upper	P-value	(SE)	Lower	Upper	P-value
λ_1	1.235(1.354)	-1.438	3.907	0.3631	-0.489(1.438)	-3.327	2.349	0.7342
λ_2	0.464(0.045)	0.375	0.552	<.0001	0.460(0.045)	0.372	0.549	<.0001
λ_3	1.617(0.083)	1.452	1.782	<.0001	1.615(0.083)	1.451	1.780	<.0001
Brabant	0.043(0.246)	-0.443	0.529	0.8613				
Limburg	0.156(0.256)	-0.349	0.661	0.5420				
E_Fland	0.131(0.241)	-0.344	0.606	0.5874				
W_Fland	-0.107(0.298)	-0.694	0.481	0.7209				
x-coordinate					0.003(0.002)	-0.002	0.007	0.2428
y-coordinate					-0.001(0.004)	-0.009	0.007	0.7839
Gender	0.034(0.128)	-0.219	0.287	0.7927	0.128(0.128)	-0.124	0.380	0.3182
Age	-0.009(0.163)	-0.330	0.312	- 0.9566	0.070(0.162)	-0.249	0.389	0.6653
1	-1.308(0.746)	-2.781	0.164	0.0813	-0.359(0.747)	-1.834	1.116	0.6314
2	-0.451(0.762)	-1.954	1.052	0.5548	0.504(0.737)	-0.950	1.958	0.4946
3	-0.850(0.771)	-2.371	0.672	0.2719	0.152(0.742)	-1.313	1.616	0.8384
4	-1.071(0.753)	-2.557	0.415	0.1567	-0.132(0.742)	-1.598	1.333	0.8587
E 5	-1.089(0.778)	-2.624	0.447	0.1635	-0.063(0.774)	-1.591	1.464	0.9348
X 6	-0.928(0.762)	-2.433	0.576	0.2249	0.196(0.723)	-1.230	1.623	0.7861
A 7	-1.520(0.800)	-3.098	0.058	0.0590	-0.380(0.755)	-1.870	1.111	0.6160
M 8	-0.982(0.786)	-2.533	0.569	0.2131	0.180(0.752)	-1.304	1.663	0.8113
I 9	-0.534(0.759)	-2.032	0.963	0.4824	0.392(0.747)	-1.082	1.865	0.6005
N 10	-1.143(0.773)	-2.668	0.382	0.1408	-0.130(0.759)	-1.628	1.369	0.8647
E 11	-1.858(0.802)	-3.440	-0.275	0.0217	-0.658(0.754)	-2.147	0.831	0.3844
R 12	-0.738(0.839)	-2.393	0.918	0.3803	0.254(0.796)	-1.317	1.824	0.7503
S 13	-1.482(0.785)	-3.032	0.068	0.0607	-0.463(0.757)	-1.958	1.032	0.5417
14	-0.742(0.792)	-2.304	0.821	0.3500	0.368(0.747)	-1.106	1.842	0.6225
15	-1.274(0.732)	-2.718	0.170	0.0835	-0.394(0.716)	-1.808	1.020	0.5828
16	-0.577(0.733)	-2.023	0.869	0.4322	0.342(0.732)	-1.102	1.787	0.6404

† The random school effects variance component σ^2, was estimated to be nearly zero

Table 3. The classification matrices and the conditional classification probabilities of the examiners 12, 13, and 16 versus the gold standard when scoring caries experience.

Examiner	M_j	conditional classification probabilities (γ_{jrh}) estimate	
		sample ($m_{jab}/\sum_d m_{jad}$)	posterior
j=12	$\begin{pmatrix} 16 & 1 & 0 & 0 \\ 1 & 2 & 2 & 0 \\ 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}$	$\begin{pmatrix} .941 & .333 & .000 & .000 \\ .059 & .667 & .222 & .000 \\ .000 & .000 & .778 & .000 \\ .000 & .000 & .000 & 1.000 \end{pmatrix}$	$\begin{pmatrix} .846 & .115 & .044 & .067 \\ .078 & .771 & .132 & .069 \\ .039 & .057 & .781 & .069 \\ .038 & .057 & .043 & .796 \end{pmatrix}$
j=13	$\begin{pmatrix} 18 & 0 & 1 & 0 \\ 2 & 3 & 1 & 0 \\ 0 & 1 & 8 & 0 \\ 2 & 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} .818 & .000 & .100 & .000 \\ .091 & .750 & .100 & .000 \\ .000 & .250 & .800 & .000 \\ .091 & .000 & .000 & 1.000 \end{pmatrix}$	$\begin{pmatrix} .775 & .066 & .10 & .086 \\ .097 & .733 & .10 & .086 \\ .032 & .134 & .75 & .090 \\ .096 & .067 & .05 & .738 \end{pmatrix}$
j=16	$\begin{pmatrix} 17 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}$	$\begin{pmatrix} 1.000 & .000 & .000 & .000 \\ .000 & 1.000 & .000 & .000 \\ .000 & .000 & 1.000 & .000 \\ .000 & .000 & .000 & 1.000 \end{pmatrix}$	$\begin{pmatrix} .953 & .018 & .017 & .018 \\ .016 & .945 & .017 & .018 \\ .016 & .018 & .949 & .018 \\ .016 & .018 & .017 & .945 \end{pmatrix}$

Table 4. Weighted kappa(κ_w) measuring agreement between the gold standard and each of the 16 dental examiners when scoring caries experience.

		95% Conf. Limits				95% Conf. Limits	
	Estimate(ASE)	Lower	Upper		Estimate(ASE)	Lower	Upper
κ_{w_1}	0.7460(0.094)	0.562	0.931	κ_{w_9}	0.8163(0.114)	0.593	1.040
κ_{w_2}	0.8930(0.050)	0.794	0.992	$\kappa_{w_{10}}$	1.0000(0.000)	1.000	1.000
κ_{w_3}	0.7587(0.120)	0.525	0.993	$\kappa_{w_{11}}$	0.9208(0.055)	0.813	1.029
κ_{w_4}	0.8094(0.076)	0.661	0.958	$\kappa_{w_{12}}$	0.8885(0.055)	0.782	0.995
κ_{w_5}	0.8074(0.079)	0.653	0.962	$\kappa_{w_{13}}$	0.6851(0.116)	0.458	0.912
κ_{w_6}	0.9265(0.052)	0.825	1.028	$\kappa_{w_{14}}$	0.8242(0.081)	0.665	0.984
κ_{w_7}	0.9627(0.037)	0.890	1.035	$\kappa_{w_{15}}$	0.8636(0.099)	0.670	1.057
κ_{w_8}	0.8855(0.064)	0.760	1.011	$\kappa_{w_{16}}$	1.0000(0.000)	1.000	1.000

Table 5. Parameter estimates from a corrected random effects multinomial logit model (6) predicting the degree of caries experience, controlling for geographical effects and externally controlling for examiner's effects (frequentist approach), using the SAS procedure NLMIXED.

	Provinces				x-y coordinates			
	Estimate	Conf. Limits			Estimate	Conf. Limits		
Parameter	(SE)	Lower	Upper	P-value	(SE)	Lower	Upper	P-value
λ_1	-0.804(1.204)	-3.180	1.572	0.5054	-0.890(1.302)	-3.460	1.680	0.4952
λ_2	0.362(0.053)	0.257	0.468	$< .0001$	0.360(0.053)	0.256	0.465	$< .0001$
λ_3	1.586(0.095)	1.400	1.773	$< .0001$	1.581(0.094)	1.395	1.767	$< .0001$
Brabant	0.192(0.220)	-0.242	0.625	0.3839				
Limburg	0.365(0.210)	-0.049	0.780	0.0839				
E_Fland	0.156(0.197)	-0.232	0.545	0.4274				
W_Fland	-0.352(0.215)	-0.776	0.073	0.1038				
x-coordinate					0.004(0.001)	0.001	0.007	0.0031
y-coordinate					-0.003(0.004)	-0.010	0.005	0.4901
Gender	0.030(0.134)	-0.233	0.294	0.8199	0.031(0.133)	-0.232	0.294	0.8162
Age	0.129(0.168)	-0.202	0.460	0.4415	0.143(0.166)	-0.186	0.471	0.3921
σ^2 †	0.046(0.082)	-0.116	0.208	0.5743	0.044(0.081)	-0.116	0.204	0.5871

† Confidence limits are omitted because negative values for the 95% confidence lower limits are reported by SAS

Table 6. Parameter estimates from a corrected random effects multinomial logit model (6) predicting the degree of caries experience, controlling for geographical effects and externally controlling for examiners' effect (Bayesian approach), using WINBUGS when scoring caries experience.

	Provinces				x-y coordinates			
		Posterior				Posterior		
	Posterior	Quantiles		Bayesian	Posterior	Quantiles		Bayesian
Parameter	Mean(SD)	2.5%	97.5%	P-value	Mean(SD)	2.5%	97.5%	P-value
λ_1	-0.871(1.376)	-3.604	1.820	0.2585	-0.937(1.576)	-4.065	2.189	0.2769
λ_2	0.347(0.058)	0.235	0.464	0.0000	0.354(0.060)	0.239	0.472	0.0000
λ_3	1.519(0.104)	1.319	1.730	0.0000	1.550(0.107)	1.345	1.764	0.0000
Brabant	0.227(0.247)	-0.263	0.717	0.1682				
Limburg	0.474(0.240)	0.014	0.965	0.0230				
E_Fland	0.155(0.228)	-0.291	0.612	0.2241				
W_Fland	-0.364(0.242)	-0.848	0.108	0.0818				
x-coordinate					0.005(0.002)	0.002	0.008	0.0015
y-coordinate					-0.003(0.005)	-0.013	0.006	0.2184
Gender	0.030(0.152)	-0.266	0.337	0.4156	0.034(0.156)	-0.274	0.331	0.4085
Age	0.124(0.192)	-0.251	0.507	0.2620	0.140(0.195)	-0.251	0.520	0.2339
σ^2	0.083(0.080)	0.001	0.291	0.0000	0.057(0.023)	0.023	0.112	0.0000

Table 7. The posterior summary statistics of the w's, which characterize the precision of examiner j in reference to gold standard, based on WINBUGS.

		Posterior	
	Posterior	Quantiles	
	Mean(SD)	2.5%	97.5%
w_1	0.797(0.068)	0.621	0.888
w_2	0.848(0.045)	0.748	0.931
w_3	0.823(0.058)	0.679	0.913
w_4	0.818(0.056)	0.678	0.905
w_5	0.810(0.058)	0.667	0.895
w_6	0.870(0.046)	0.781	0.957
w_7	0.874(0.049)	0.781	0.968
w_8	0.848(0.048)	0.740	0.935
w_9	0.795(0.080)	0.588	0.902
w_{10}	0.881(0.054)	0.784	0.984
w_{11}	0.862(0.051)	0.756	0.956
w_{12}	0.830(0.051)	0.708	0.915
w_{13}	0.819(0.052)	0.690	0.899
w_{14}	0.809(0.059)	0.665	0.896
w_{15}	0.846(0.057)	0.713	0.945
w_{16}	0.885(0.052)	0.791	0.984