Online supplement to Modelling segregation as unevenness and as clustering

All the models are calculated using the MLwiN software which has both likelihood and MCMC capabilities (Charlton e*t al*, 2017). More details on the procedures used are given in the manuals of Jones and Subramanian (2013, 2017) and Browne (2017). Here we provide MCMC trajectories and estimated posterior distributions for two models, the first being strictly hierarchical and the second being a cross-classified multiple membership specification. MLwiN syntax, a Stata-do file and an R command are also provided, and these can be used to specify and estimate the models in the MLwiN software which provides for fast estimation of complex models.

Strictly hierarchical model

The model is specified as follows

$$Y_i \sim Binomial(n_i, \pi_i)$$

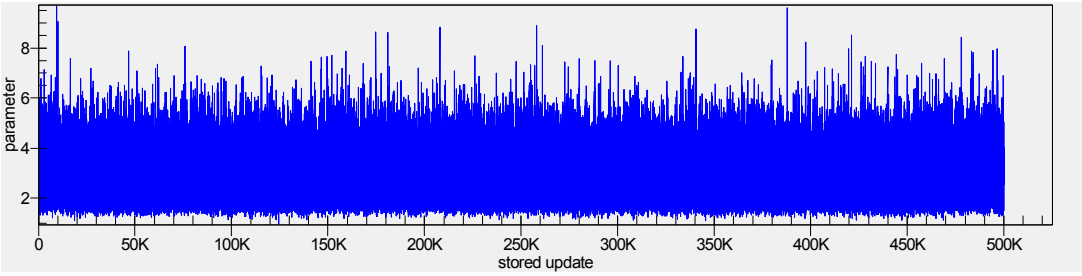
$$E\left(log_e\left(\frac{\pi_i}{1-\pi_i}\right)\right) = \beta_0 + \mu^{(3)}_{Zone(i)} + \mu^{(2)}_{OA(i)}$$

$$\mu^{(3)}_{Zone(i)} \sim N(0, \sigma^2_{\mu^3})$$
$$\mu^{(2)}_{OA(i)} \sim N(0, \sigma^2_{\mu^2})$$
$$Var\ (Y_i|\pi_i) = \frac{\pi_i(1-\pi_i)}{n_i}$$

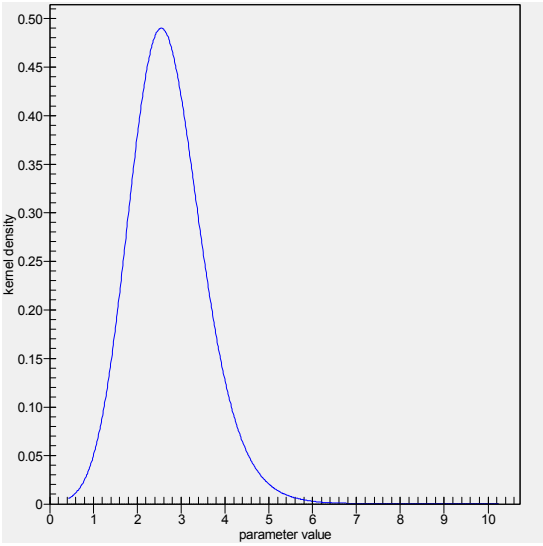
with two variances measuring segregation at the Zonal and Output Area level in addition to the level 1 binomial stochastic variation.

Between Zone variance: σ^2_3

The model was calibrated with 500,000 monitoring estimates after a burnin of 5,000 draws and the estimates are shown on the plot below. The effective sample size is equivalent to 386,291 independent draws and there is no sign of trending so that the sampler has been run sufficiently long to produce high-quality estimates. It could have comfortably been estimated with a much shorter run.

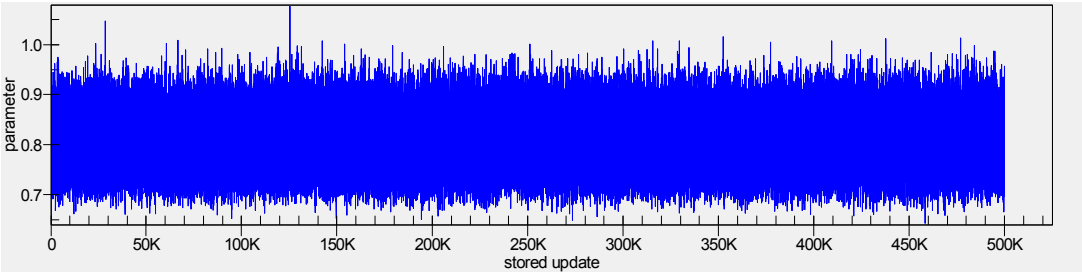


These draws are then turned into a kernel density plot to give the posterior distribution of the variance. The positive skewness for this higher-level is marked (there are only 39 Zones) so the Bayesian credible intervals of the variance are found to be asymmetric.

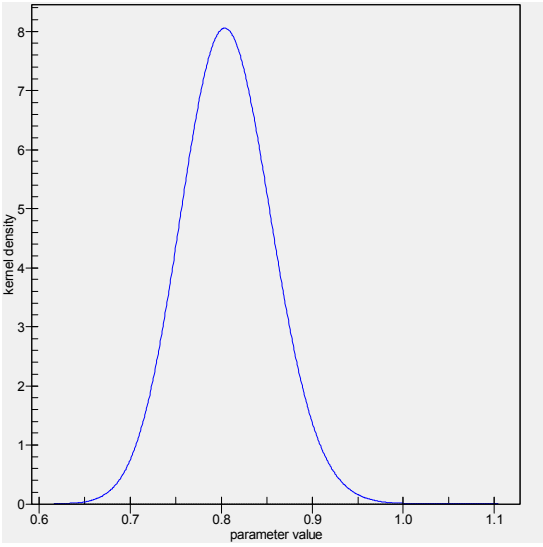


Within Zone between Output area variance: σ_2^2

The between OA variance is obtained in the same way and the 500,000 monitoring draws, plotted below, behave as 198,938 independent draws which is more than sufficient to characterise the posterior distribution.



The kernel density plot gives the posterior distribution of the variance and this time (there are 969 OAs) the plot shows a much more normal symmetric distribution.



Cross-classified Multiple Membership model MCMC trajectories

The model is specified as follows:

$$E\left(log_e\left(\frac{\pi_i}{1-\pi_i}\right)\right) = \beta_0 + \sum_{j \in 40NHood(i)} w_{i,j}^{(4)} \mu_j^{(4)} + \sum_{j \in 3NHood(i)} w_{i,j}^{(3)} \mu_j^{(3)} + + \mu_{OA(i)}^{(2)}$$

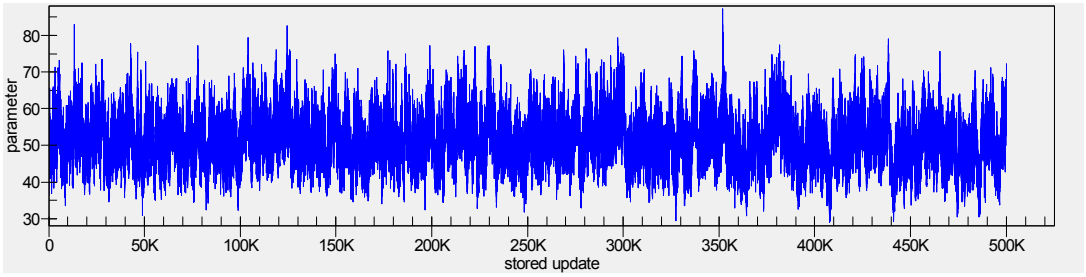
$$\mu_{40NHood(i)}^{(4)} \sim N(0, \sigma_{\mu^4}^2)$$
$$\mu_{3NHood(i)}^{(3)} \sim N(0, \sigma_{\mu^3}^2)$$
$$\mu_{OA(i)}^{(2)} \sim N(0, \sigma_{\mu^2}^2)$$

$$Var\ (Y_i|\pi_i) = \frac{\pi_i(1-\pi_i)}{n_i}$$

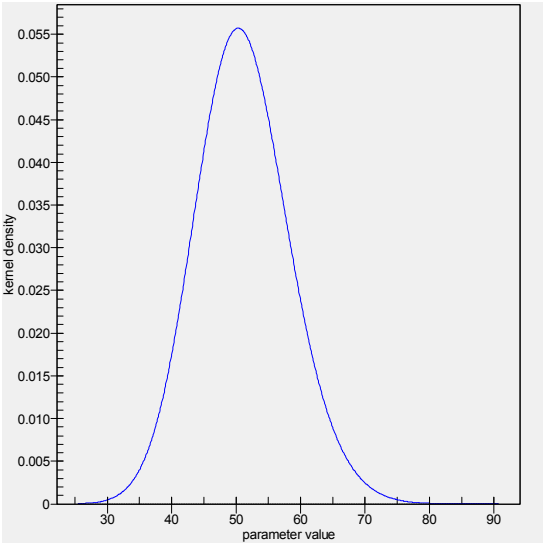
There are 3 estimated variances terms in addition to stochastic variance and (as an illustration) the large neighbourhood consists of 40 members, the smaller neighbourhood consists of 3 members, and there is additional unstructured variance between OAs. Again, a burnin of 5,000 draws was followed by a monitoring phase based on 500,000 simulations for each parameter. Each of the parameter chains in these more complex models is more correlated and the long monitoring period is definitely needed to characterise the posterior distribution.

Between Neighbourhood variance with 40 members: σ_4^2

The trajectories for this variance clearly show less white noise than is the case for the strict hierarchy but, while there is slower mixing, the parameter space is still being explored and there is no overall trend indicating that the sampler has achieved the equilibrium distribution. There are now the equivalent of 772 independent draws which are used to characterise the posterior distribution.

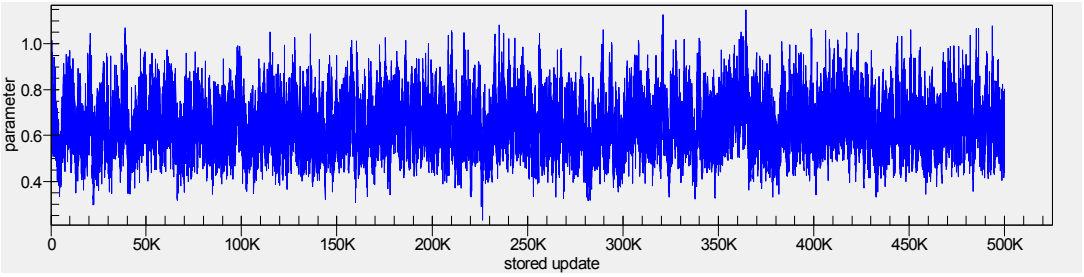


The kernel density smooth shows an approximate normal distribution as may be expected given that there are 969 ‘patches’.

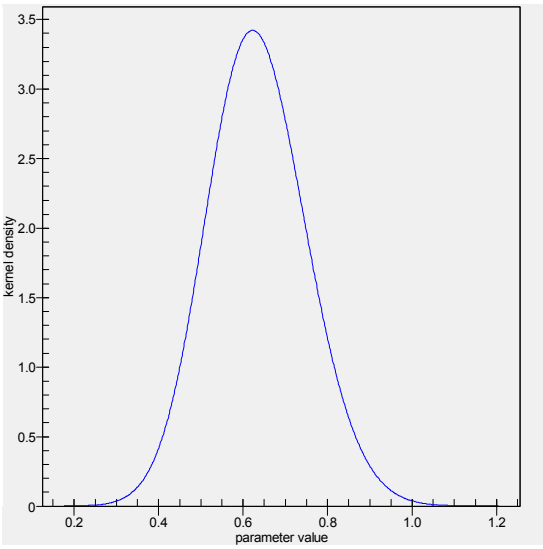


Between Neighbourhood variance with 3 members: σ_3^2

The trajectory for this variance again appears to be behave well in that there is no overall trend and the parameter space is being explored. The 500,000 draws are quite autocorrelated, however, and they are estimated to have the information content equivalent to 949 independent draws. But this is sufficient to characterise the posterior distribution.

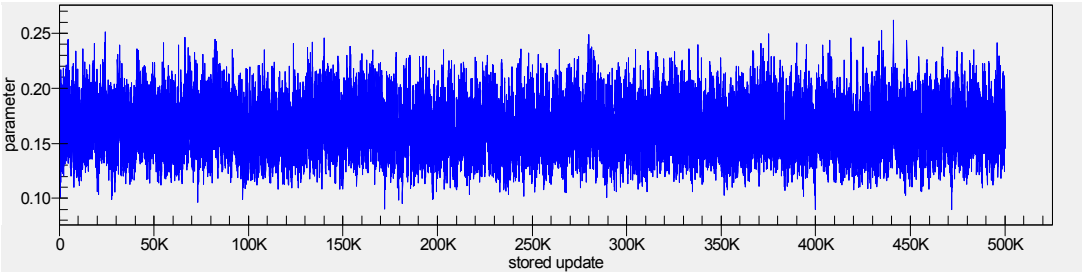


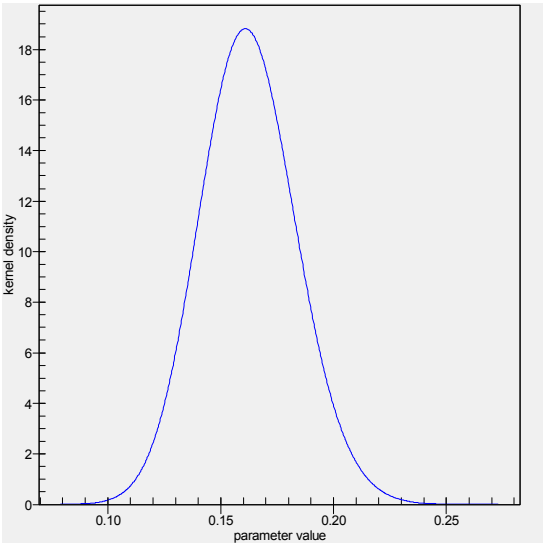
The posterior distribution is clearly approximately normal and symmetric.



Unstructured variance between OAs: σ_2^2

This variance also shows that the equilibrium distribution had been reached during the 5000 draw burnin as there is no trend in the estimates. The 500,000 monitoring draws have an information content of 2,164 independent estimates and this is sufficient to portray the posterior distribution which is normally distributed.





MLwiN macro commands to specify and estimate the cross-classified multiple membership model

MLwiN can operate through commands and syntax and this can be formed into an executable macro; details on commands and macros are given in Rasbash *et al* (2003). The following commands can be used to set up a four-classification model with the larger patch consisting of 50 Output Areas and the smaller patch consisting of 7 OAs and there being unevenness between OAs and stochastic variation according to a binomial distribution.

```
Note Clear any current model specification
CLEAR

Note set up the response variable and the 4-classification structure
RESP 'IndRate'
IDEN 4 '1sto50'
IDEN 3 'Seven.1'
IDEN 2 'UniqueOA'
IDEN 1 'UniqueOA'
```

```
Note 'IndRate' is the observed proportion of Indians in each OA
Note '1sto50' must be the first of 50 consecutive columns giving the
Note multiple membership OA identifiers for classification 4
Note 'Seven.1' must be the first of 7 consecutive columns giving the
Note multiple membership OA identifiers for classification 3
Note 'UniqueOA' gives unique identifier for each Output Area
Note this is specified for classification 2 (unevenness)
Note and again at classification 1 (pure stochastic variation)

Note specify as a binomial model with 'Total' (ie Indian plus Non-
Note  Indian) as the denominator of the proportion
RDISt 1 0
LFUN 0
DOFFs 1 'Total'

Note add a fixed constant (all 1s) and allow associated variance at
Note each of the three higher levels
Note display this model which is strictly hierarchical
ADDT 'cons'
SETV 4 'cons'
SETV 3 'cons'
SETV 2 'cons'
WSET

Note choose settings for less-correlated MCMC chains
ORTH 1
HCEN 1 4

Note set up the cross classification
Note for classification 4 with weights for the 50-membership
Note zonation; there must be 50 consecutive columns starting at
Note  column 'wt50.1'
Note For classification 3 with weights for the 7-membership zonation
Note there must be 7 consecutive columns starting at column '7wt1'
MULM 4 50 'wt50.1'
MULM 3 7 '7wt1'
XCLA 1

Note choose general notation and display the model
EXPA 2
NOTA 0
INDE 1
WSET

Note Initially use IGLS maximum likelihood to get starting values
Note and store estimates; these are not to be interpreted
METH 1
BATCH 1
START
MWIPE
MSTORE 'IGLS Estimates'

Note switch to MCMC estimation
EMODE 3
Note burnin for 5k, scale factor 5.8, 50% acceptance for Metropolis
```

```
Note Hastings as logit model does not have closed form
MCMC 0 5000 1 5.8 50 10 2 2 2 1 1 2

Note iterate for 5k stored iterations (500,000 total – storing
Note every 100 to get less corelated chain;
Note store estimated chain in c1090, deviance in c1090
Note parameter means in c1003, parameter s.d. in c1004
MCMC 1 5000 100 c1090 c1091 c1003 c1004 1 2
Note copy parameter means and s.d. to c1096 to c1099
PUPN c1003 c1004


Note store model MCMC model estimates and details
MSTORE 'MCMC Estimates'
```

Using runmlwin to use MLwiN from within Stata

runmlwin is a Stata command which allows Stata users to run the faster MLwiN software from within Stata on more complex multilevel models than allowed by the standard Stata commands (Leckie and Charlton, 2013). Both Stata and MLwiN must be available to the user. The following do-file sets up the CCMM model with a 50 OA and a 7 OA patch and initially estimates a likelihood model to obtain starting values and then uses MCMC estimation. That is the code produces exactly the same model as the native MLwiN code given above.

```
use "leicsmm.dta", clear

* Run IGLS model for starting values
* (nosort option is required as ID columns do not match the expected
hierarchy)
quietly runmlwin indrate cons, ///
         level4(id50_1: cons) ///
         level3(id7_1: cons) ///
         level2(uniqueoa: cons) ///
         level1(uniqueoa:) ///
         discrete(distribution(binomial) link(logit) denom(total))
///
         nosort nopause
estimates store igls

* Now fit the model with MCMC
runmlwin indrate cons, ///
         level4(id50_1: cons, mmids(id50_1-id50_50)
mmweights(wt50_1-wt50_50)) ///
         level3(id7_1: cons, mmids(id7_1-id7_7) mmweights(wt7_1-
wt7_7)) ///
         level2(uniqueoa: cons) ///
         level1(uniqueoa:) ///
discrete(distribution(binomial) link(logit) denom(total)) ///
         mcmc(burnin(5000) chain(500000) thinning(100) orth hcen(4))
initsmodel(igls) nopause
estimates store mcmc

* Display the results
estimates table mcmc
```

Using R2MLwiN use MLwiN from within R

R2MLwiN is an R command interface to the MLwiN multilevel modelling software (Zhang et al, 2016) Both R and MLwiN must be available to the user. The following R commands sets up the CCMM model with a 50 OA and a 7 OA patch and initially estimates a likelihood model to obtain starting values and then uses MCMC estimation. That is the code produces exactly the same model as the native MLwiN code given above.

```
library(R2MLwiN)
library(foreign)

leicsmm <- read.dta("leicsmm.dta")

(mcmc <- runMLwiN(logit(indrate, total) ~ 1 + (1 | id50_1) + (1 |
id7_1) + (1 | uniqueoa),
          D = "Binomial",
          estoptions = list(
              EstM = 1,
              mcmcMeth = list(burnin = 5000, iterations = 500000,
thinning = 100),
              mcmcOptions = list(orth = 1, hcen = 4),
              mm = list(
                  list(mmvar = paste0("id50_", 1:50), weights =
paste0("wt50_", 1:50)),
                  list(mmvar = paste0("id7_", 1:7), weights =
paste0("wt7_", 1:7)),
                  NA,
                  NA
              )
          ),
          data = leicsmm))
```

References in the online material

Browne WJ (2017) *MCMC Estimation in MLwiN 3.0.* University of Bristol: Centre for Multilevel Modelling. http://www.bristol.ac.uk/cmm/media/software/mlwin/downloads/manuals/3-01/mcmc-web.pdf

Charlton CJ, Rasbash J. Browne WJ Healy M and Cameron B (2017) Software: *MLwiN Version 3.00*. University of Bristol: Centre for Multilevel Modelling.

Jones K and Subramanian SV (2017) *Developing multilevel models for analysing contextuality, heterogeneity and change using MLwiN 3.0, Volume 1*. University of Bristol: Centre for Multilevel Modelling. https://www.researchgate.net/publication/260771330

Jones K and Subramanian SV (2013) *Developing multilevel models for analysing contextuality, heterogeneity and change using MLwiN Volume 2*, University of Bristol: Centre for Multilevel Modelling. https://www.researchgate.net/publication/260772180

Leckie, G. and Charlton, C. (2013). runmlwin - A Program to Run the MLwiN Multilevel Modelling Software from within Stata. *Journal of Statistical Software*, 52 (11),1-40.

Rasbash, J., Browne, W.J. and Goldstein, H. (2003) *MLwiN command manual*, University of Bristol: Centre for Multilevel Modelling.
http://www.bristol.ac.uk/cmm/media/software/mlwin/downloads/manuals/2-00/commman20.pdf

Zhang, Z, Parker, RMA, Charlton, CMJ, Leckie, G and Browne WJ. (2016) R2MLwiN: A Package to Run MLwiN from within R. *Journal of Statistical Software*, 72(10), 1-43.