



Spatial design model and estimation

In many multilevel models, the focus is on the possible relationship between an outcome variable of interest and some characteristics of respondents nested within some form of hierarchy. This may be students nested within schools or respondents nested within geographical units as is often the case with census data or data from sample designs where final sampling units are chosen within geographically defined areas. Spatial dependence becomes relevant if, apart from the relationship between the outcome and predictors of interest at level-1 of the hierarchy, the actual geographical proximity of level-2 units may impact this relationship as well, essentially adding a third dimension Z to the relationship between the predictor(s) X and the outcome Y .

If we want to make provision for the possibility of spatial heterogeneity, we have to consider that respondents' characteristics of interest may vary by location, in which case the assumption of constant variance may no longer be realistic. The closer the respondents are geographically, the more likely that one observation may influence another; the larger the distance between them, the smaller the potential influence may be. This type of spatial connectivity can be modelled by assigning a spatial weight to each observation, with the weight assigned a reflection of how much the value of a nearby observation may impact on another.

In the standard hierarchical linear model, it is assumed that the covariance between level-1 and level-2 random effects is equal to zero. It is also assumed that the neighborhood random effects are independent. When dealing with data from contiguous sites, this assumption may no longer be realistic and it may become necessary to take the spatial dependence of neighborhood social processes into account as well.

A model incorporating an autocorrelation term is a good way to model such a situation, as autocorrelation will allow the impact of proximity to fade with geographical distance and is appropriate when the assumption of independence between neighboring values is no longer tenable.

In matrix notation, we can write this model as

$$\mathbf{Y} = \gamma(\mathbf{1}_N) + \mathbf{Vb} + \boldsymbol{\varepsilon}$$

where

- γ is a fixed (scalar) effect,
- \mathbf{b} is a $J \times 1$ vector of level-2 random spatially autoregressive effects,
- \mathbf{V} a $N \times j$ block diagonal design matrix that assigns to each level-2 unit j the appropriate element b_j of the random effects vector \mathbf{b} , and
- $\boldsymbol{\varepsilon}$ is the $N \times 1$ vector of level-1 errors, $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_N)$.

The vector \mathbf{b} can, in turn, be expressed as

$$\mathbf{b} = \rho \mathbf{W} \mathbf{b} + \mathbf{u},$$

where

- ρ is the spatial dependence parameter,
- \mathbf{W} is a non-symmetric spatial weight matrix defined in such a way that the element $w_{jj'}$ is negatively related to the distance between neighborhoods j and j' , and
- \mathbf{u} a $J \times 1$ vector of level-2 errors, $\mathbf{u} \sim N(0, \tau \mathbf{I}_J)$.

It is further assumed that \mathbf{u} is independent of $\boldsymbol{\varepsilon}$, and that σ^2 and τ are scalar level-1 and level-2 variances respectively.

The contiguity matrix \mathbf{W} may be a binary contiguity matrix indicating that sites are contiguous to each other. It could, in other circumstances, be more complex and incorporate the distance between sites along with characteristics of the areas such as relative area, the proportion of common boundary, etc. The matrix \mathbf{W} implemented in the HLM program is a binary contiguity matrix where $w_{jj'} = 1$ if neighborhoods j and j' are contiguous, and 0 if not. In addition, the rows of the weight matrix are standardized so that the total distance between neighborhood j and all other neighborhoods is unity. This row-standardization is performed so that the sum of the contiguous neighbor contributions has an upper limit of 1:

$$w_{jj'}^* = w_{jj'} / \sum_{h=1}^J w_{jh}$$

so that

$$\sum_{j=1}^J w_{jj'}^* = 1.$$

This helps to ensure that the spatial dependence parameter ρ will have an absolute value less than 1. The spatial dependence parameter may be interpreted as follows in the following cases:

- $\rho \neq 0$: spatial dependence between sites is present.
- $\rho = 0$: no spatial dependence is present, and the model reduces to a standard one-way random-effects ANCOVA model $\mathbf{Y} = \gamma(\mathbf{1}_N) + \mathbf{V}\mathbf{u} + \boldsymbol{\varepsilon}$.
- $\rho > 0$: Indicates that a site is typically surrounded by other sites with similar values on the outcome of interest.
- $\rho < 0$: High-value sites are typically surrounded by low-value sites, and vice versa.

If the equations above are rewritten in terms of \mathbf{b} , we get

$$\mathbf{Y} = \gamma(\mathbf{1}_N) + \mathbf{V}(\mathbf{I}_J - \rho\mathbf{W})^{-1} \mathbf{u} + \boldsymbol{\varepsilon}.$$

This is a special case of HLM where the design matrix for the random effects \mathbf{u} is $\mathbf{V}(\mathbf{I}_J - \rho\mathbf{W})^{-1}$. As this expression involves the inversion of the matrix $\mathbf{I}_J - \rho\mathbf{W}$, it is necessary for this matrix to be non-singular. Non-singularity is a requirement for the existence of a unique inverse of a matrix. No site can appear in the data as its own neighbor.

In HLM, this model is estimated by using maximum likelihood via the EM algorithm. The spatial dependence parameter ρ is re-estimated on each iteration.

Example

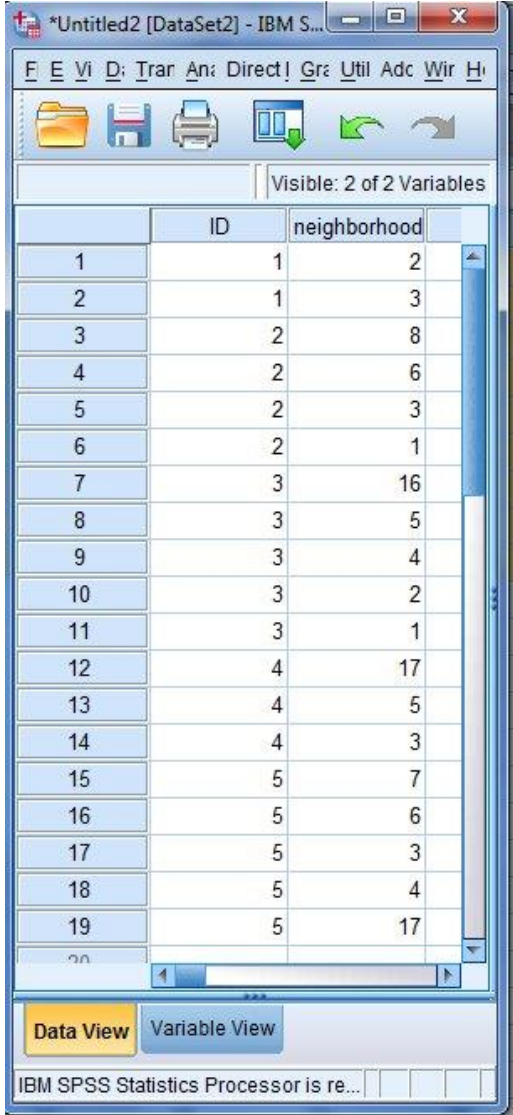
We consider the example provided with the program. Data on 7 720 residents from 342 Chicago neighborhoods collected by the Project of Human Development in Chicago Neighborhoods are used. The outcome variable of interest is collective efficacy, which is defined as social cohesion among neighbors combined with their willingness to intervene on behalf of the common good. This variable is measured on a scale consisting of ten items, indicating whether people in the neighborhood know each other, trust each other, share common values, and can be relied on in various ways to maintain public order. The individual scores serving as outcome in this example was based on the responses to these items.

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Collective Efficacy	7729	1.00	5.00	3.4175	.71014
Valid N (listwise)	7729				

The number of respondents from the neighborhoods varied between 7 and 60.

Creating a spatial dependence matrix for use in HLM

The data set described above is now used to illustrate the creation of the spatial dependence matrix for use in creating an MDM in HLM. Information on the geographical proximity of neighborhoods are typical given in the form shown below. In this example, the ID represents the neighborhood ID, and the variable NEIGHBORHOOD any neighborhood adjacent to the neighborhood given in column 1.



The screenshot shows the IBM SPSS Statistics Processor interface. The main window displays a data table with two columns: 'ID' and 'neighborhood'. The table contains 19 rows of data. The 'ID' column lists neighborhood identifiers from 1 to 19, and the 'neighborhood' column lists the adjacent neighborhood IDs for each. The interface includes a menu bar (File, Edit, View, Data, Transform, Analyze, Direct, Graph, Utilities, Advanced, Window, Help), a toolbar with icons for file operations, and a status bar at the bottom indicating 'IBM SPSS Statistics Processor is re...'. The 'Data View' tab is selected.

ID	neighborhood
1	2
2	3
3	8
4	6
5	3
6	1
7	16
8	5
9	4
10	2
11	1
12	17
13	5
14	3
15	7
16	6
17	3
18	4
19	17

The first neighborhood shares borders with 2 others, namely neighborhoods 2 and 3. Neighborhood 3 shares borders with all four other neighborhoods shown, that is 1,2,4 and 5. Neighborhoods 4 and 5 also share borders with neighborhood 17. While neighborhood 1 shares borders with only two neighborhoods, neighborhoods 4 and 5 have 5 neighbors each.

To use this information as a spatial design matrix in HLM, the data must be presented in a different form. Instead of having multiple rows of information for each neighborhood ID, a single row for each is needed. Additional columns are introduced into the data set to represent the neighborhoods adjacent to each. In the example below, the data shown above has been reformatted.

The screenshot shows the IBM SPSS Statistics Data Editor window with a dataset named '*test.sav [DataSet1]'. The data is displayed in a table with 7 columns: 'id', 'n1', 'n2', 'n3', 'n4', 'n5', and 'count'. The 'id' column lists neighborhoods 1 through 6. The 'n1' through 'n5' columns represent the IDs of adjacent neighborhoods. The 'count' column shows the number of adjacent neighborhoods for each ID. The status bar at the bottom indicates 'IBM SPSS Statistics Processor is ready'.

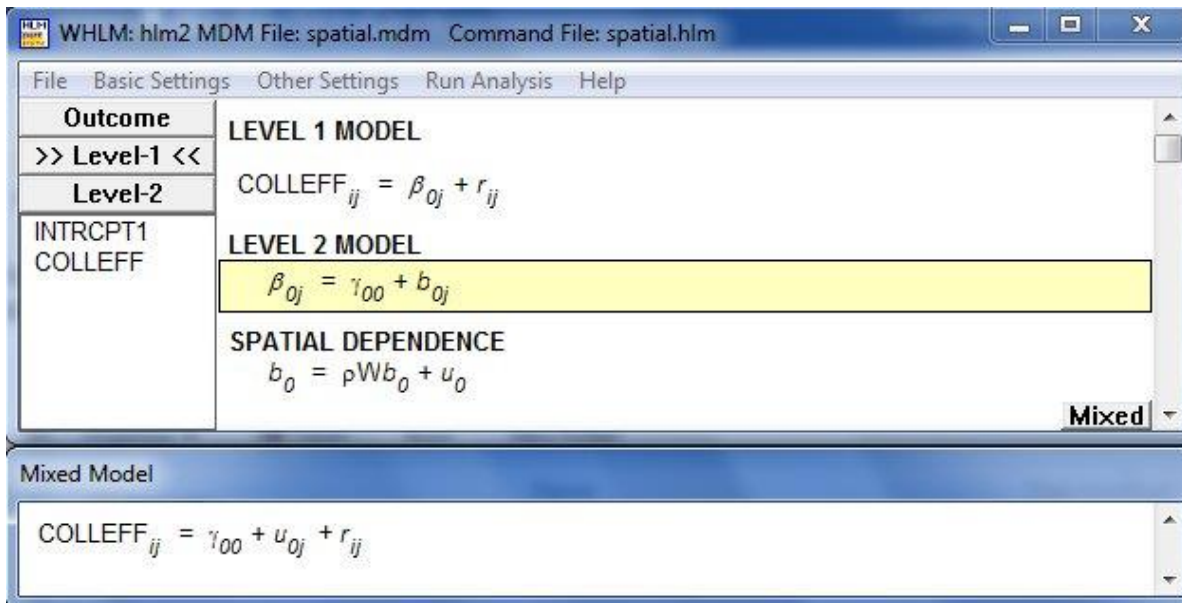
	id	n1	n2	n3	n4	n5	count
1	1	2	3	.	.	.	2
2	2	8	6	3	1	.	4
3	3	16	5	4	2	1	5
4	4	17	5	3	.	.	3
5	5	7	6	3	4	17	5
6							

The variables n1 to n5 represent the IDs of neighborhoods adjacent to each of the 5 neighborhoods shown above. As at most 5 neighborhoods were adjacent to each of the 5 neighborhoods in the data shown above, only 5 columns (n1 to n5) are needed. The number of columns required is equal to the maximum number of neighbors for any neighborhood in the data set of interest.

The count variable records the actual number of adjoining neighborhoods: neighborhood 3, for example, had 5 neighbors are recorded by n1 to n5 and thus the value of count for this neighborhood is 5. Neighborhood 4 with three neighbors have values on 3 of the columns n1 to n5, and subsequently a count of 3. The maximum value of the variable count is thus equal to the number of columns used to represent adjoining neighborhoods. No neighborhood can appear in the reformatted data as its own neighbor.

Interpreting the results of the model

The following model is fitted to PHDCN data. In this model b_0 represents the level-2 random spatially autoregressive effects (corresponding to the vector \mathbf{b} in the model discussed in the introduction), W the spatial weight matrix and ρ the spatial dependence parameter. Keep in mind that in this model σ^2 and τ are scalar level-1 and level-2 variances respectively.



A spatial dependence analysis produces two sets of results: one for the standard 2-level model with no autoregressive component, and one for the model with potential spatial dependence. The first set of results is for the standard 2-level model:

Final Results - Iteration 6

Iterations stopped due to small change in likelihood function

$$\sigma^2 = 0.42136$$

Standard error of $\sigma^2 = 0.00693$

τ

INTRCPT1, β_0 0.08904

Standard error of τ

INTRCPT1, β_0 0.00850

Note that for the standard model, τ is not assumed to be scalar, and thus both estimate and standard error appears in the first section of output and in the final table of variance components (see below).

Approximate confidence intervals of tau variances
 INTRCPT1: (0.074,0.107)

Random level-1 coefficient	Reliability estimate
INTRCPT1, β_0	0.799

The value of the log-likelihood function at iteration 6 = -7.911855E+03

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. d.f.	p-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	3.433243	0.018056	190.142	341	<0.001

**Final estimation of fixed effects
 (with robust standard errors)**

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. d.f.	p-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	3.433243	0.018056	190.144	341	<0.001

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	d.f.	χ^2	p-value
INTRCPT1, u_0	0.29839	0.08904	341	1870.37148	<0.001
level-1, r	0.64913	0.42136			

Statistics for the current model

Deviance = 15823.710765
 Number of estimated parameters = 3

The results for the spatial dependence model are given next:

Iterations stopped due to small change in likelihood function

$$\sigma^2 = 0.42149$$

τ
INTRCPT1, β 0.03477

ρ
INTRCPT1, β 0.81701

The estimated ρ of 0.81701 indicates that a site is typically surrounded by other sites with similar values on the outcome of interest and that spatial dependence is present.

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. d.f.	p-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	3.404181	0.056443	60.312	341	<0.001

Statistics for the current model

Deviance = 15671.980461
Number of estimated parameters = 4

The fixed effects results are not all that different between the two models ($\hat{\gamma}_{00} = 3.4332$ versus $\hat{\gamma}_{00} = 3.4042$). When the standard errors of $\hat{\gamma}_{00}$ (0.018 versus 0.056) is compared over the models, it indicates that, given that $\hat{\rho} = 0.8170$, there is an underestimation of the standard errors when spatial dependence is ignored. Note that there is no table of results for variance components: in the spatial dependence model τ is assumed to be a scalar. Comparison of the two models using deviance statistics is included in the output. From the result (given below) it can be concluded that the model with spatial dependence provides a better fit.

Regular HLM vs. HLM with spatial dependence model comparison test

χ^2 statistic = 151.73030

Degrees of freedom = 1

p -value = <0.001

The final results given provide information on the neighborhood-specific variance and the average covariance between pairs of adjacent neighborhoods. Note that these depend on τ , the magnitude of the spatial dependence correlation ρ , and the configuration of neighborhoods near that neighborhood.

A comparison of the standard errors for $\hat{\gamma}_{00}$ the regular HLM and HLM with spatial dependence (.018 vs .056) suggests that, given $\hat{\rho}$ is equal to .8, that there is an underestimation of the standard errors when spatial dependence is ignored.

Average Level-2 Variance = 0.088502

Average Level-2 Covariance = 0.005961

It is also possible to obtain spatial empirical Bayes estimates of the neighborhood collective efficacy measures by requesting the level-2 residual file via the **Basic Settings** dialog box. The contents of this file are shown below for the first 20 neighborhoods. The variable u_intrcp represents empirical Bayes estimates for the standard 2-level model, while b_intrcp represents the empirical Bayes estimates for the model with spatial dependence.

resfil2.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketi Graph: Utilitie: Add-on: Window Help

Visible: 5 of 5 Variables

	l2id	nj	u_intrcp	b_intrcp	f_intrcp	
1	1	39	-.213	-.420	3.404	
2	2	43	-.168	-.292	3.404	
3	3	15	.021	-.214	3.404	
4	4	41	-.128	-.203	3.404	
5	5	15	.101	-.021	3.404	
6	6	13	-.099	-.155	3.404	
7	7	39	-.113	-.138	3.404	
8	8	14	.192	.178	3.404	
9	9	45	.287	.328	3.404	
10	10	17	.281	.376	3.404	
11	11	11	-.112	-.002	3.404	
12	12	14	.076	.250	3.404	
13	13	17	.135	.270	3.404	
14	14	16	.096	.206	3.404	
15	15	11	.041	.066	3.404	
16	16	13	-.359	-.502	3.404	
17	17	15	.071	-.041	3.404	
18	18	21	-.014	-.024	3.404	
19	19	35	-.076	-.270	3.404	
20	20	11	-.063	-.185	3.404	

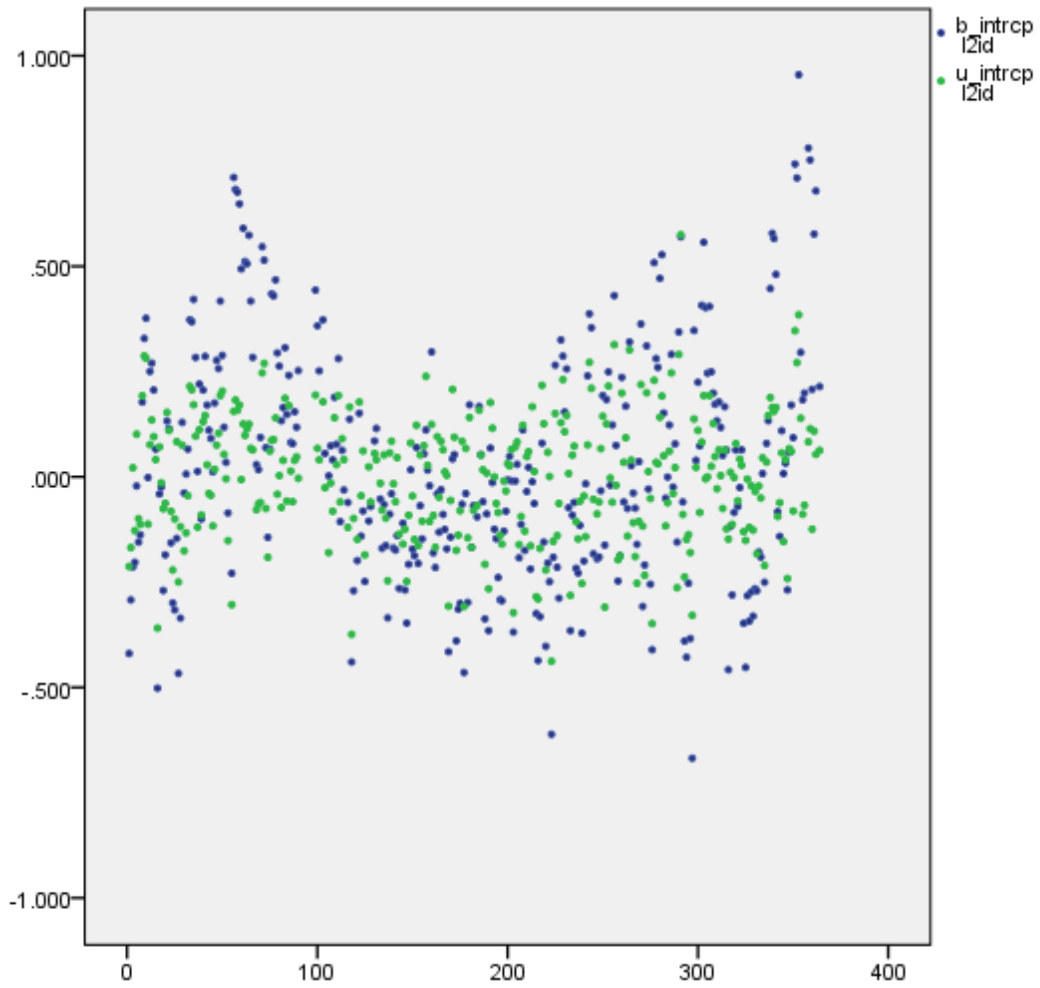
Data View Variable View

IBM SPSS Statistics Processor is ready

Descriptive statistics for the two sets of empirical Bayes estimates are given below. It is clear that there is more variation in *b_intrcp*, that is in the empirical Bayes estimates for the model with spatial dependence. This is also clear from a scatterplot of the estimates.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
u_intrcp	342	-.438	.575	-.00057	.147520
b_intrcp	342	-.668	.955	.03218	.277098
Valid N (listwise)	342				



Spatial dependence models can be fitted to continuous as well as discrete outcomes such as binary outcomes, counted data, ordinal and multinomial outcomes.