VERGLEICHENDE SOZIALFORSCHUNG MIT MEHREBENENMODELLEN IN R

Forschungspraktikum I und II
Dr. Christian Czymara
Advanced multi-level structures

AGENDA

- Beyond two levels
  - Nesting in nesting in nesting
  - Cross-classified models
- Applications
  - Spatial nesting
  - Temporal (and spatial) nesting
- Tutorial: Analyzing relationships of local wealth and feeling a sense of direction in one's life

EVALUTION

- http://r.sd.uni-frankfurt.de/2a01c310

SPATIAL CLUSTERING

BEYOND COUNTRY CLUSTERING

- So far, we analyzed individuals in countries, with a special interest in the impact of country characteristics
- However, for many research questions, a more local impact is reasonable
- For example: Is it really *national* crime levels that shape my perceptions? Or rather crime that happens in my vicinity (neighborhood, city, state, …)?
- Crucial question: Are data available on this level?
- The European Union has a geocoded system for comparison of local statistics: The NUTS regions

NOMENCLATURE DES UNITÉS TERRITORIALES STATISTIQUES

NOMENCLATURE OF TERRITORIAL UNITS FOR STATISTICS (NUTS)

- Within each country, there are three increasingly fine-grained NUTS levels
- Currently, there are 92 NUTS1 regions, 244 NUTS2 regions, and 1215 NUTS3 regions
- Example: Germany
  - NUTS1: 16 states (*Bundesländer*, e. g.: Hesse)
  - NUTS2: 39 government regions (*Regierungbezirke*, e. g.: Darmstadt)
  - NUTS3: 429 districts (*Landkreise* and *kreisfreie Städte*, e. g.: Frankfurt am Main)

GERMANY: NUTS1

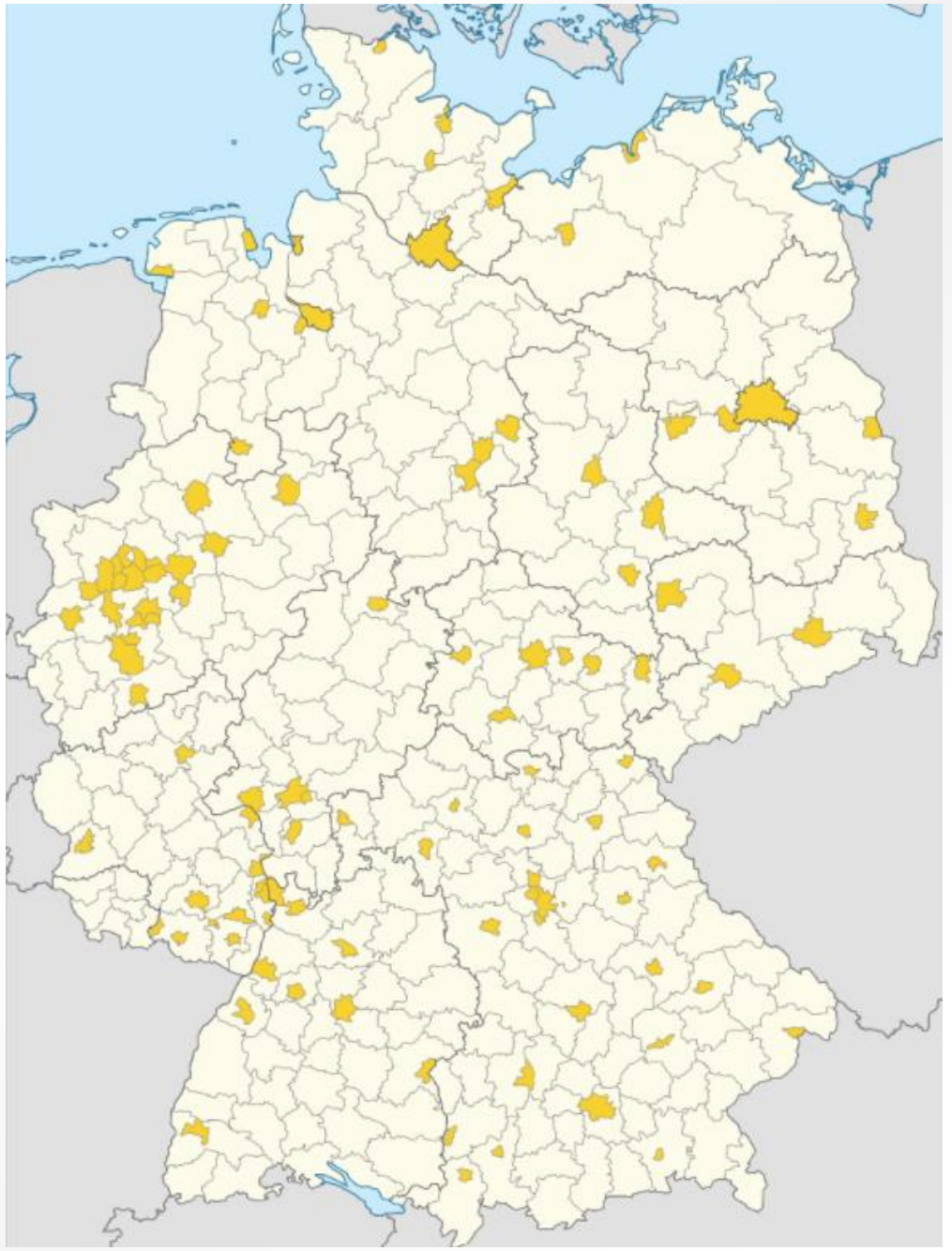
- Source: https://en.wikipedia.org/wiki/States_of_Germany

GERMANY: NUTS2

- Example: Hesse
- Source: https://en.wikipedia.org/wiki/Darmstadt_(region)

GERMANY: NUTS3

- Yellow: Kreisfreie Städte
- Source: https://en.wikipedia.org/wiki/Districts_of_Germany



EXAMPLE: DOES ETHNIC DIVERSITY MAKE NATIVES WELCOMING OR HOSTILE?

- There is a large body of literature on the impact of out-group size on public attitudes toward immigration and immigrants
- However, different mechanisms plausible
  - More immigrants in a *country* → More threat (e.g. because of political discourses or selective media reporting) → More prejudice
  - More immigrants in ones local area → More contact → Less prejudice
- So-called modifiable areal unit problem
- Weber (2015) analyzes this question based on the European Values Study on country and NUTS3 levels

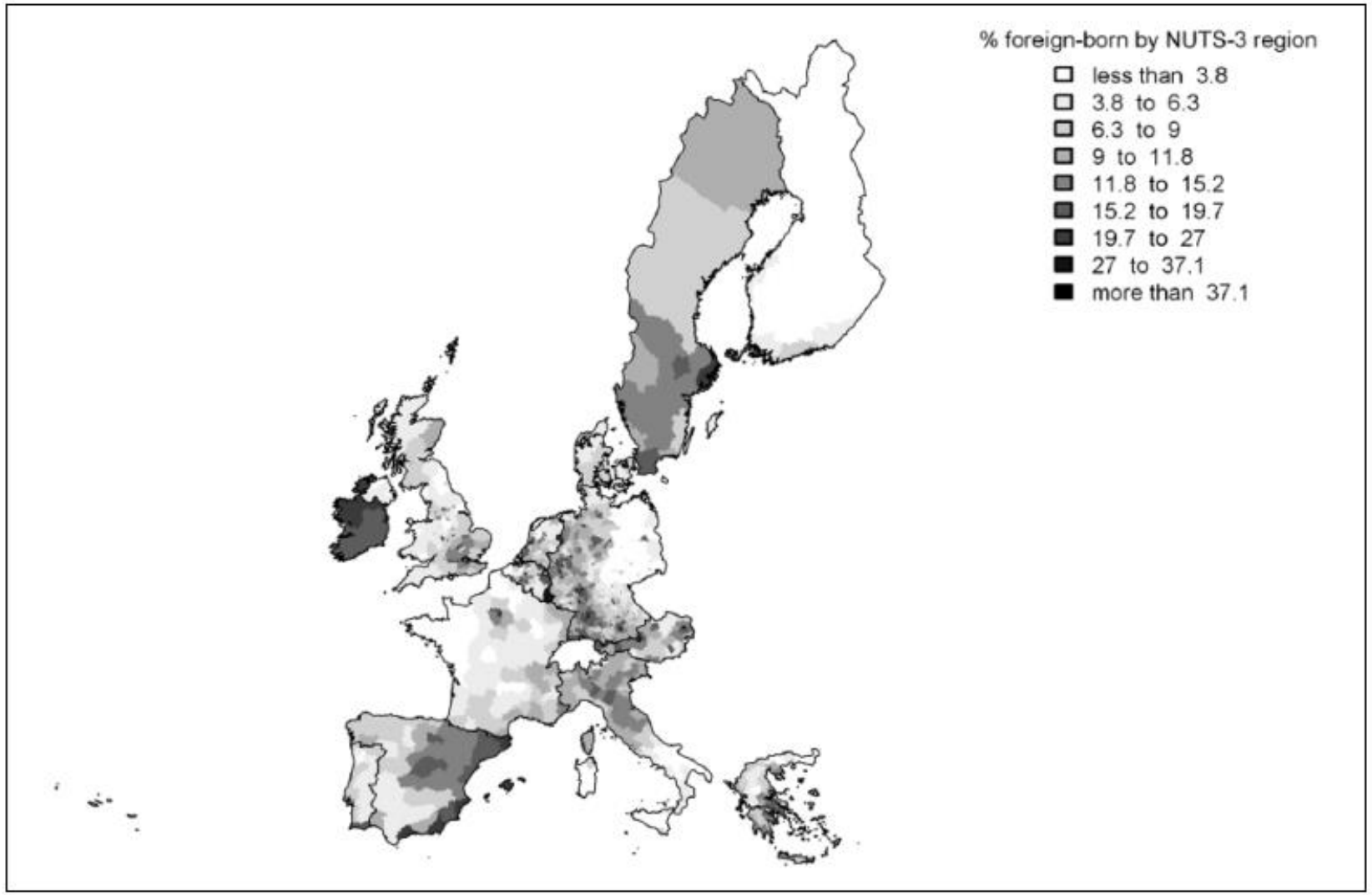


Figure 1. Percentage of foreign-born residents by NUTS-3 region in the EU-15.
Source: Eurostat, national statistical offices, see Note 18 and Appendix 1.

Table 1. Predictors of perceived group threat (three-level analyses).

	Model 1/2: NUTS-3 regions		Model 3/4: NUTS-2 regions		Model 5/6: NUTS-1 regions	
	With LUX	Without LUX	With LUX	Without LUX	With LUX	Without LUX
	B (SE)	*B* (SE)	*B* (SE)	*B* (SE)	*B* (SE)	*B* (SE)
Intercept	.482 (.680)	−.003 (.702)	.916 (1.027)	.338 (1.031)	.236 (1.259)	−.467 (1.258)
Individual level						
Age	.002*** (.000)	.002*** (.000)	.002*** (.000)	.002*** (.000)	.002*** (.000)	.002*** (.000)
Gender (female)	−.038** (.013)	−.038** (.013)	−.040** (.013)	−.043** (.013)	−.041** (.013)	−.043** (.013)
Education	−.142*** (.005)	−.145*** (.005)	−.143*** (.005)	−.147*** (.005)	−.145*** (.005)	−.148*** (.005)
Immigrant background	−.407*** (.019)	−.397*** (.021)	−.410*** (.019)	−.401*** (.021)	−.417*** (.019)	−.409*** (.021)
Class (*reference group: middle class*)						
Working class	.013 (.014)	.011 (.015)	.015 (.015)	.013 (.015)	.020 (.015)	.018 (.015)
Upper class	−.120*** (.019)	−.131*** (.019)	−.125*** (.019)	−.136*** (.020)	−.120*** (.019)	−.144*** (.020)
Moving experience (*reference group: non-movers*)						
Has moved to more diverse region	−.093*** (.027)	−.092*** (.027)	−.113*** (.028)	−.112*** (.027)	−.130*** (.027)	−.128*** (.027)
Has moved to less diverse region	−.103*** (.030)	−.102*** (.030)	−.098*** (.030)	−.097*** (.030)	−.086** (.030)	−.085** (.029)
Contextual levels						
Percentage of immigrants in region	−.007* (.003)	−.008* (.003)	−.004 (.004)	−.004 (.004)	−.009* (.004)	−.010* (.004)
Log(GDP per capita) in region	−.021 (.155)	.010 (.154)	−.108 (.238)	−.070 (.233)	.061 (.292)	.111 (.285)
Percentage of immigrants in country	.022 (.014)	.055* (.019)	.015 (.013)	.055* (.019)	.015 (.013)	.060* (.021)
Pseudo-R^2	.10	.11	.09	.11	.09	.11

GDP: gross domestic product; NUTS: *Nomenclature des Unités Territoriales Statistiques*; SE: standard error.
Source: European Values Study (2008), national statistical offices, own calculations.
Multi-level linear regression analyses (maximum likelihood). Dependent variable: Perceived threat of immigration. For coding of variables see text. *N* (individuals) = 22,683; *N* (NUTS-1) = 70; *N* (NUTS-2) = 207; *N* (NUTS-3) = 624; *N* (countries) = 15. For models without Luxembourg: *N* (individuals) = 21,073; *N* (NUTS-3) = 623; *N* (countries) = 14.
Pseudo-$R^2 = ((\sigma_0^2+\tau_{m0}+\tau_{\beta 0})-(\sigma^2+\tau_m+\tau_\beta))/(\sigma_0^2+\tau_{m0}+\tau_{\beta 0})$, where σ^2 = Level-1 variance; τ_m = Level-2 variance; τ_β = Level-3 variance; and σ_0^2, τ_{m0}, and $\tau_{\beta 0}$ refer to the respective empty model (not shown).
* $p < .05$; ** $p < .01$; *** $p < .001$.

THREE-LEVEL MODELS

MULTIPLE RANDOM INTERCEPTS

- NUTS regions have a clear hierarchy: County > NUTS1 > NUTS2 > NUTS3
- Put differently, individuals are nested in NUTS3, which are nested in NUTS2, which are nested in NUTS1, which are nested in countries
- Three-level example: individual i in region j of country k
- Example: Erika Mustermann (level 1) is from Detmold (level 2) in Germany (level 3)
- Different situation could be a so-called cross-classified model, where individuals are nested in two independent higher-level groups simultaneously (more later)

MULTIPLE RANDOM INTERCEPTS

- $y_{ijk} = \beta_0 + u_{0j} + u_{0jk} + e_{ijk}$
- β_0: Mean value across all countries and regions
- u_{0j}: Random intercept on country level (between country variance)
- u_{0jk}: Random intercept on regional level (within country / between regional variance)
- e_{ijk}: Ideosynchratic error (within region within country / between individual variance)
- We can now explain variance on each of these three levels

RANDOM EFFECTS SPECIFICATION WITH LME4

Formula	Alternative	Meaning
`(1 | g)`	`1 + (1 | g)`	Random intercept with fixed mean.
`0 + offset(o) + (1 | g)`	`-1 + offset(o) + (1 | g)`	Random intercept with *a priori* means.
`(1 | g1/g2)`	`(1 | g1)+(1 | g1:g2)`	Intercept varying among `g1` and `g2` within `g1`.
`(1 | g1) + (1 | g2)`	`1 + (1 | g1) + (1 | g2).`	Intercept varying among `g1` and `g2`.
`x + (x | g)`	`1 + x + (1 + x | g)`	Correlated random intercept and slope.
`x + (x || g)`	`1 + x + (1 | g) + (0 + x | g)`	Uncorrelated random intercept and slope.

Table 2: Examples of the right-hand-sides of mixed-effects model formulas. The names of grouping factors are denoted `g`, `g1`, and `g2`, and covariates and *a priori* known offsets as `x` and `o`.

TEMPORAL SPATIAL CLUSTERING: REPEATED CROSS-NATIONAL DATA

REPEATED CROSS-NATIONAL DATA

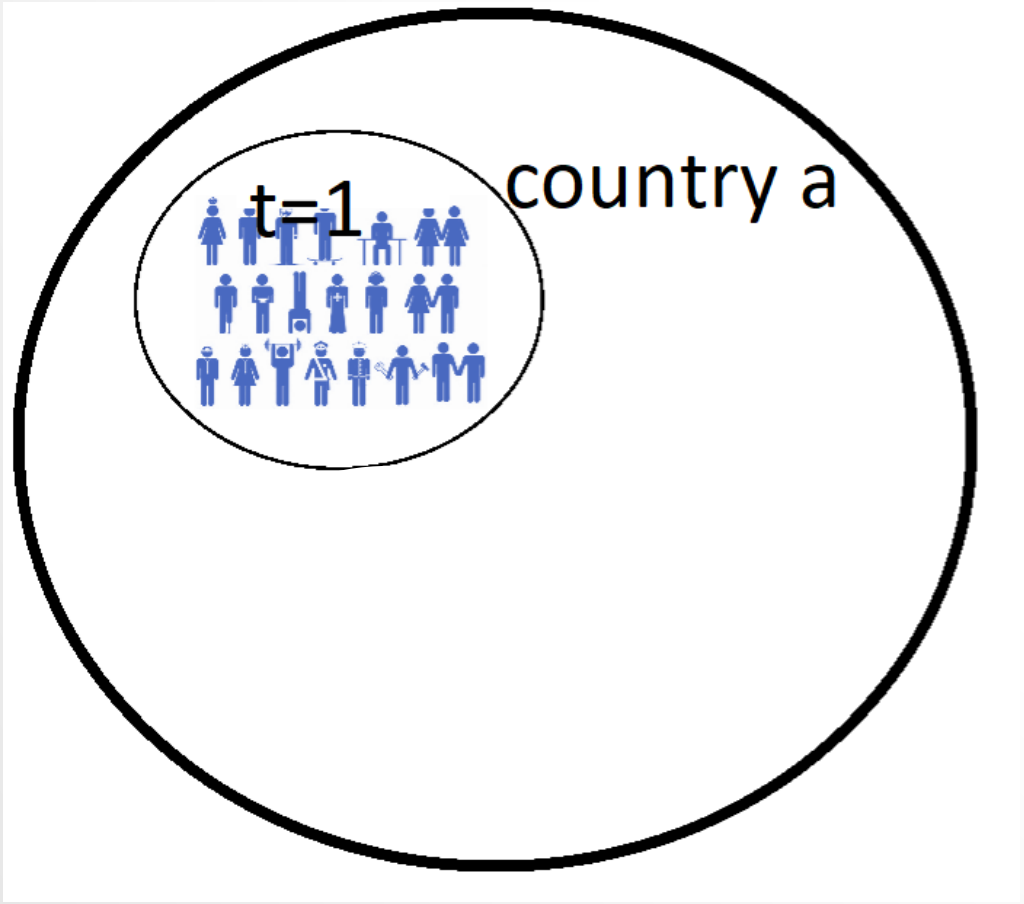
- Many cross-national survey programs are repeated every few years
  - Called Repeated Cross-Sectional Data
  - Or macro-panel / pseudo-panel data
  - I will call it *Repeated Cross-National Data*

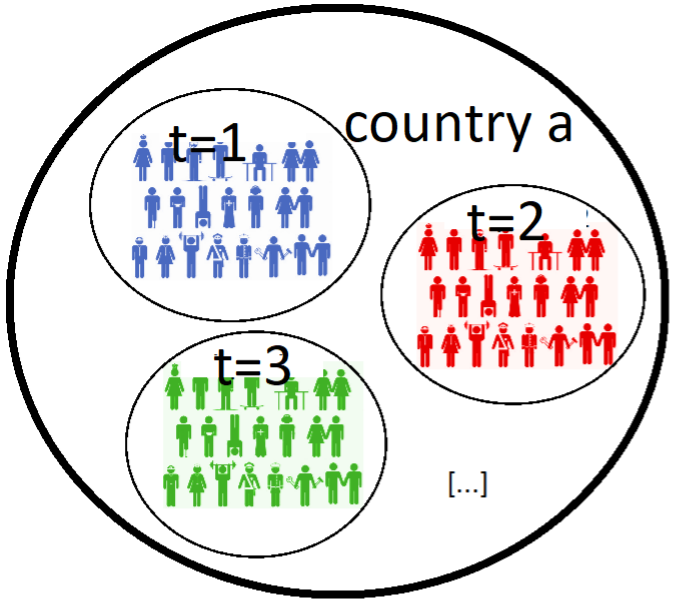
→Different individuals in each wave

→But: panel data on the country-level

- However, this means that the data structure is quite complex
  - Individuals nested in countries (see cross-sectional case)
  - Time (country-years) nested in countries (see panel data)

TIME SERIES OF COUNTRY A




country a
t=1
t=2
t=3
[...]

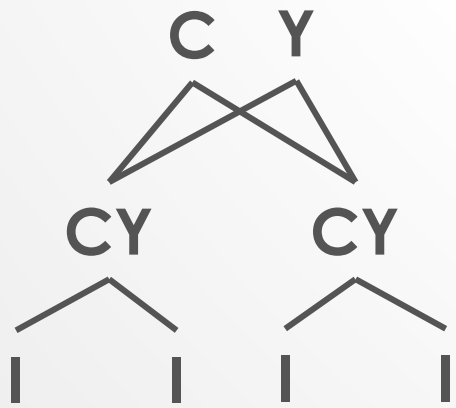
RANDOM EFFECTS STRUCTURE FOR REPEATED CROSS-NATIONAL DATA

DATA STRUCTURE

- The data we analyzed so far were hierarchical with two levels: individuals (level 1) nested in countries (level 2)
- Now, we have four potential levels:
  - Countries (Norway, Estonia, Sweden etc.) → C
  - Years (2002, 2004, 2006 etc.) → Y
  - Country-years (Norway 2002, Norway 2004, Estonia 2002 etc.) → CY
  - Individuals → I
- Thus, possible statistical dependencies are more complex
  - I nested in C (people in Sweden share something that separates them from people in France)
  - I nested in Y (people in 2010 share something that separates them from people in 2018)
  - CY nested in C (France in 2002 shares something with France in 2008 that separates it from UK in 2002)
  - CY nested in Y (France in 2002 shares something with UK in 2002 that separates it from France in 2008)
  - …
- Ignoring each of the possible statistical dependencies within clusters will lead to an underestimation of standard errors (too significant effects, see first session on HLMs)

RANDOM EFFECTS STRUCTURE

- Various ways to account for the different kinds of clustering
- The most complex approach would model individuals (level 1) nested in country-years (level 2) and country-years as *cross-classified* in both countries (level 3) and years (also level 3)
- With random effects for countries, years and country-years (see Schmidt-Catran & Fairbrother 2015: 25)

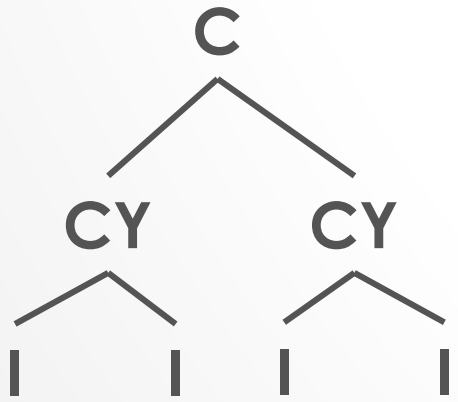


RANDOM EFFECTS STRUCTURE

- Such a RE structure is very complex and might face problems of convergence
- Research interest are often in variables on the C and the CY level, less on the Y level
- Modeling CY nested in Y accounts for similarities of respondents interviewed at the same time point, independent of country (individuals interviewed in 2010 more similar those interviewed in 2004 and 2010)
- However, often, there are no variables that only vary on this level (over time but not across countries)
- That must be something that happens over time and affects all countries similarly (COVID-19? A meteor?)

RANDOM EFFECTS STRUCTURE

- Hence, Schmidt-Catran & Fairbrother (2015) suggest to ignore the clustering on the Y level (but not on the CY level!)
- Without Y, the RE structure simplifies to



- Which is a three level structure: *Individuals are nested in country-years, which are nested in countries*
- This accounts for all relevant dependencies of such a data structure

RANDOM EFFECTS STRUCTURE

- For cross-sectional comparative data, we had two levels: I in C
- … and one random intercept, accounting for clustering of I in C
- Now we have three levels: I in CY in C
- Thus, need two random intercepts:
  - One for clustering of I in CY
  - One for clustering of CY in C

A THREE-LEVEL RE MODEL FOR RCND

- $y_{ijt} = \beta_0 + \beta x_{ijt} + \gamma z_{jt} + \delta z_j + u_j + u_{tj} + e_{ijt}$
- Indexes
  - i: individual-level
  - j: country-level
  - t: time-level
- x_{ijt}: individual-level variable (effect: β)
- z_{jt}: time-varying country-level variable (effect: γ)
- z_j: time-constant country-level variable (effect: δ)

RANDOM EFFECTS PARAMETERS

- $y_{ijt} = \beta_0 + \beta x_{ijt} + \gamma z_{jt} + \delta z_j + u_j + u_{jt} + e_{ijt}$
- u_j: Random effect on country level (level 3 error)

→Variance between countries

- u_{jt}: Random effect on country-year level (level 2 error)

→Variance within countries / between country-years

- e_{ijt}: Idiosyncratic error (level 1 error)

→Variance within country-years / between individuals

LITERATURE

- Schmidt-Catran & Fairbrother (2015). The random effects in multilevel models: Getting them wrong and getting them right. European Sociological Review, 32 (1), 23-38.