Two Composites Exercise

Jim Grace

U.S. Department of the Interior
U.S. Geological Survey

1

This module deals with a slightly more complex situation that can occur with working with composites, that of multiple effects.

An appropriate citation for this material is

Grace, J.B., and Bollen, K.A. (2006) The interface between theory and data in structural equation models: U. S. Geological Survey Open-File Report 2006-1363, 33 p.

Note, see especially the example in Figure 12 in the above publication.

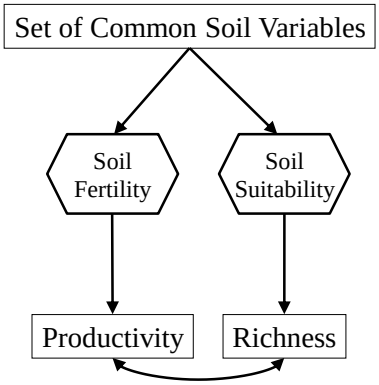
https://profile.usgs.gov/myscience/upload_folder/ci2012Nov2316254439968Grace%20and%20Bollen2006_USGS_OFR.pdf

Notes: IP-056512; Support provided by the USGS Land Use R&D and Ecosystems Programs. I would like to acknowledge formal review of this material by Jesse Miller and Phil Hahn, University of Wisconsin. Many helpful informal comments have contributed to the final version of this presentation. The use of trade names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Last revised 18.08.26.

Source: https://www.usgs.gov/centers/wetland-and-aquatic-research-center/science/quantitative-analysis-using-structural-equation

In this exercise, the goal is to estimate composites for two different response variables of interest in a model. The example involves data from grasslands and the effects of soil influences on productivity and species richness.



USGS

2

Previous modules on composites have dealt with the case where there is one response for which effects of multiple causes are being combined. However, it is quite possible there may be multiple responses to be modelled. Here there will be some challenges. In the process of addressing these challenges, you will get more familiarity with practical solutions for modeling with composites.

There are typically several steps in this process.

Step 1: Run model without composites using all soil indicators.

Step 2: Prune contributing indicators and select final uncomposited model.

Step 3: Create composite variables within lavaan, estimate model, and evaluate.

Note: You may want to consult the module "Modeling with Composite Variables" to refresh yourself with the options before attempting this exercise.

USGS

3

The setup in R.

```
### TWO-COMPOSITES EXERCISE
# data extracted from Grace et al 2016 Nature paper

### Load libraries
library(lavaan)
library(AICcmodavg)
source("D:/TalksAndTrips/FY2017/Germany/Workshop/Part
1/lavaan.modavg.R")


### Read Data and Rename Variables
# Set working directory
setwd("D:/ppt_files/_education/SEM.10-Modeling with
Composite Variables/NutnetExample")

# read data
dat <- read.csv("TwoCompositesExercise_2016.csv")
names(dat)
```

USGS

4

lavaan.mod.avg.R can be obtained from "http://jarrettbyrnes.info/ubc_sem/lavaan_materials/lavaan.modavg.R" if need be.

The setup in R continued*.

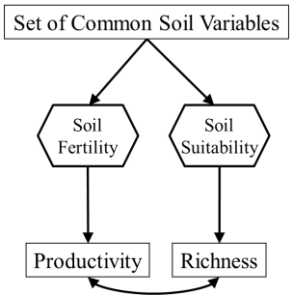
```
# rename and adjust scale of variables for
convenience
dat2 <- with(dat, data.frame(site))
dat2$Rich  <- dat$ln.rich
dat2$Prod  <- dat$ln.prod
dat2$Sand  <- dat$sand.prop
dat2$Silt  <- dat$silt.prop/100
dat2$PH    <- dat$ph/100
dat2$P     <- dat$ln.p/10
dat2$C     <- dat$ln.c
dat2$N     <- dat$ln.n
dat2$K     <- dat$ln.k
```

*Note that I am giving you enough code to get started without consulting the code provided with this exercise.

5

It is typical that we need to code the variables. In this case we are trying to get the variance roughly equal for the lavaan analysis.

Lavaan code for initial model.




```
##### Forming Composites within lavaan
### Step 1: Run model without composites using all soil
### indicators
mod1 <- 'Rich ~ Sand +Silt +PH +P +C +N +K
         Prod ~ Sand +Silt +PH +P +C +N +K'
mod1.fit <- sem(mod1, data=dat2, meanstructure = TRUE)
summary(mod1.fit)
```

USGS

Here is the code for the initial model. Accompanying this module are data and code files.

[When you have finished with your work, go to the next slides to compare with those anticipated for this exercise. You may also wish to consult the code file provided with this exercise, which sometimes has additional details.]

USGS

7

A solution is provided in the following slides.

Step 2: Prune contributing indicators and select final uncomposited model.

```
##### Forming Composites within lavaan
### Step 1: Run model without composites using all soil
### indicators
mod1 <- 'Rich ~ Sand +Silt +PH +P +C +N +K
         Prod ~ Sand +Silt +PH +P +C +N +K'
mod1.fit <- sem(mod1, data=dat2, meanstructure = TRUE)
summary(mod1.fit)
```

8

USGS

As is typical with building models containing composites, we first test the inclusion of indicators in a model omitting the composites.

Output from model 1.

```
Regressions:
                    Estimate  Std.Err  Z-value  P(>|z|)
  Rich ~
    Sand              -0.741    0.331   -2.235    0.025
    Silt            -104.632   44.552   -2.349    0.019
    PH                 4.599    3.884    1.184    0.236
    P                 -0.597    0.375   -1.591    0.112
    C                 -0.035    0.122   -0.284    0.776
    N                  0.331    0.445    0.744    0.457
    K                 -0.057    0.045   -1.253    0.210
  Prod ~
    Sand               0.152    0.197    0.774    0.439
    Silt              52.216   26.486    1.971    0.049
    PH                -5.064    2.309   -2.193    0.028
    P                  0.820    0.223    3.675    0.000
    C                  0.327    0.072    4.516    0.000
    N                 -1.156    0.265   -4.365    0.000
    K                 -0.005    0.027   -0.200    0.842
```

9

There may be better ways to identify a suitable set of predictors, but here I simply eliminated one indicator from each potential composite at a time, working from the ones with the largest p-value at a time. I take this approach because we do not have an adequate theory for specifying the set of indicators for the composites in this case.

Step 2: Prune contributing indicators and select final uncomposited model.

```
summary(mod1.fit) # look at initial model output
# revised model
mod1a <- 'Rich ~ Sand +Silt +PH +P +N +K
          Prod ~ Sand +Silt +PH +P +C +N'
mod1a.fit <- sem(mod1a, data=dat2, meanstructure = TRUE)
summary(mod1a.fit)
```

Continue with this process until you have a set of models to compare.

```
### Compare all mods using AICc criterion
aictab.lavaan(list
(mod1.fit, mod1a.fit, mod1b.fit, mod1c.fit, mod1d.fit,
mod1e.fit),
c("Model1", "Model1a", "Model1b", "Model1c", "Model1d",
"Model1e"))
```

USGS

10

My approach to model building often involves choosing a set of various models, then comparing them using a AICc table. This may or may not be ideal approach to the problem.

Output from model comparison.

	K	AICc	Delta_AICc	AICcWt	Cum.Wt	LL
Model1e	12	-945.03	0.00	0.54	0.54	486.92
Model1d	13	-944.14	0.89	0.34	0.88	488.17
Model1c	14	-941.74	3.29	0.10	0.98	488.80
Model1b	15	-938.31	6.72	0.02	1.00	489.04
Model1a	17	-869.47	75.56	0.00	1.00	459.02
Model1	19	-859.28	85.75	0.00	1.00	459.08

Results from selected model.

Regressions:

	Estimate	Std.Err	Z-value	P(>|z|)
Rich ~				
Sand	-0.656	0.254	-2.581	0.010
Silt	-95.098	36.678	-2.593	0.010
Prod ~				
Silt	32.581	10.606	3.072	0.002
PH	-5.764	1.888	-3.053	0.002
P	0.877	0.209	4.201	0.000
C	0.294	0.063	4.655	0.000
N	-1.027	0.225	-4.557	0.000

USGS

11

When creating the composites within lavaan, I use the discovered parameters as initial values in the composite building process (next slide).

(Note, however, I usually end up computing composite scores outside of lavaan and bringing them in to the lavaan modeling process as a new variable. This is shown in later slides.)

Step 3: Create composite variables within lavaan.

```
# create model with composites
#(set loading using uncomposited parameter estimates)

mod2 <- 'SoilSuitability <~ -0.656*Sand +Silt
         SoilFertility   <~ 32.581*Silt +PH +P +C +N
         Rich ~ SoilSuitability
         Prod ~ SoilFertility'

mod2.fit <- sem(mod2, data=dat2, meanstructure = TRUE)
```

USGS

12

The coefficients in the red boxes are brought over from the previous slide.

Result from attempting to run model.

```
Warning messages:
1: In lav_partable_check(lavpartable, categorical =
categorical, warn = TRUE) :
  lavaan WARNING: missing intercepts are set to zero:
[SoilSuitability SoilFertility]
2: In lav_model_vcov(lavmodel = lavmodel, lavsamplestats =
lavsamplestats,  :
  lavaan WARNING: could not compute standard errors!
  lavaan NOTE: this may be a symptom that the model is not
identified.

3: In lav_object_post_check(lavobject) :
  lavaan WARNING: observed variable error term matrix
(theta) is not positive definite; use inspect(fit,"theta")
to investigate.
```

USGS

13

lavaan is not able to estimate more than one composite in a model, which is why we should expect this error message for this situation.

Try creating only one composite.

```
# try model with only 1 composite
mod2a <- 'SoilSuitability <~ -0.656*Sand +Silt
          Rich ~ SoilSuitability
          Prod ~ Silt +PH +P +C +N'
mod2a.fit <- sem(mod2a, data=dat2, meanstructure = TRUE)
```

```
> mod2a.fit <- sem(mod2a, data=dat2, meanstructure = TRUE)
Warning message:
In lav_partable_check(lavpartable, categorical =
categorical, warn = TRUE) :
  lavaan WARNING: missing intercepts are set to zero:
[SoilSuitability]
```

Note that we get a warning, but only because we have asked for meanstructures, and the output seems fine. Generally, in this case we would remove the "meanstructure = TRUE" statement and rerun to be safe.

USGS

14

Just another "tricky bit" that pops up in covariance modeling.

Output from model 2a.

```
> mod2a.fit <- sem(mod2a, data=dat2)
summary(mod2a.fit)

Composites:
                      Estimate  Std.Err  Z-value  P(>|z|)
  SoilSuitability <~
    Sand                -0.656
    Silt               -95.155   19.559   -4.865    0.000

Regressions:
                    Estimate  Std.Err  Z-value  P(>|z|)
  Rich ~
    SoilSuitabilty     0.999    0.387    2.581    0.010
  Prod ~
    Silt              32.581   10.606    3.072    0.002
    PH                -5.764    1.888   -3.053    0.002
    P                  0.877    0.209    4.201    0.000
    C                  0.294    0.063    4.655    0.000
    N                 -1.027    0.225   -4.557    0.000
```

USGS

15

Here are some results.

Bottomline, lavaan can only estimate 1 composite variable at a time.

16

text here

Alternative Approach: Forming Composites by hand

```
### Redo Step 3, computing composites by hand
# get output from selected model to extract parameters for
# specification purposes
summary(mod1e.fit)

# calculate composites for model and bring into data set
dat2$SoilSuitability  <- with(dat2,
1.039 -0.656*Sand -95.098*Silt)

dat2$SoilFertility    <- with(dat2,
0.019 +32.581*Silt -5.764*PH +0.877*P +0.294*C -1.027*N)


### Step 4: Run model and evaluate output
mod3 <- 'Rich ~ SoilSuitability
         Prod ~ SoilFertility'
mod3.fit <- sem(mod3, data=dat2, meanstructure = TRUE)
```

USGS

text here

Output from model 3

```
Regressions:
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.all
  Rich ~
    SoilSuitabilty    1.000    0.372    2.684    0.007    0.395
  Prod ~
    SoilFertility     0.999    0.131    7.628    0.000    0.774

Covariances:
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.all
  Rich ~~
    Prod             -0.000    0.002   -0.182    0.856   -0.029
```

USGS

18

Note that when we bring in composite scores, we only get partial information in each of the separate modeling steps.

Compare results from uncomposited and composited models.

```
## From composited model (mod 3)
lavaan (0.5-20) converged normally after  35 iterations

  Number of observations                            39

  Estimator                                         ML
  Minimum Function Test Statistic                1.306
  Degrees of freedom                                 2
  P-value (Chi-square)                           0.520
```

```
## From uncomposited model (mod 1e)
lavaan (0.5-20) converged normally after  76 iterations

  Number of observations                            39

  Estimator                                         ML
  Minimum Function Test Statistic                5.106
  Degrees of freedom                                 5
  P-value (Chi-square)                           0.403
```

USGS Model testing/selection based on uncomposited model. [19]

In my experience, when a stage-one uncomposited model fits, then the composited one will also, though the fit measure estimates and degrees of freedom are different.

Compare results from uncomposited and composited models.

```
## From composited model (mod 3)

R-Square:
                    Estimate
    Rich               0.156
    Prod               0.599
```

```
## From uncomposited model (mod 1e)

R-Square:
                    Estimate
    Rich               0.156
    Prod               0.599
```

Validation of methodology from a variance explanation view.

USGS

20

And, the raw parameter estimates should be the same using the procedures presented here.

You will generally want the correlation between composites.

```
> cor.test(dat2$SoilSuitability, dat2$SoilFertility)

      Pearson's product-moment correlation

data:  dat2$SoilSuitability and dat2$SoilFertility
t = -0.88152, df = 37, p-value = 0.3837

alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4390709  0.1802510
sample estimates:
       cor
-0.1434229
```

21

I find it quite interesting to discover what the correlation is between composites (when there is more than one in a model).

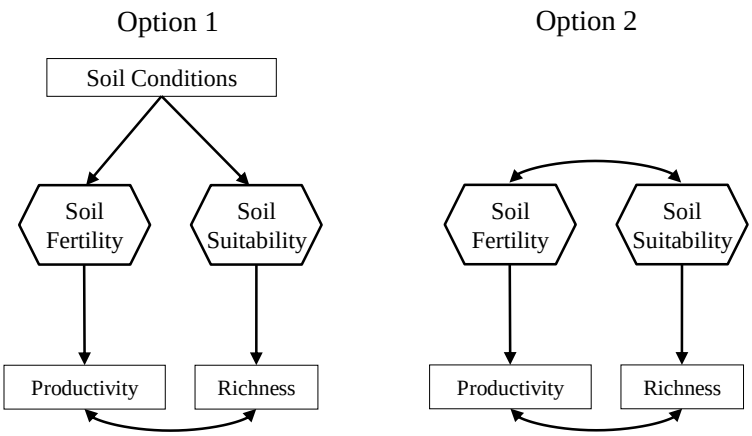
You also need the error correlation between Prod and Rich in this case.

```
Covariances:
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.all
  Rich ~~
    Prod             -0.000    0.002   -0.180    0.857    -0.029
```

22

The error correlation tells us something about the “other forces” that are operating.

There are various ways you can present this visually. Best to have all the details in a table somewhere.

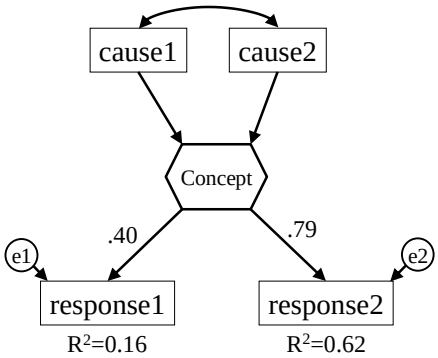


USGS

23

I might decide between Option 1 and Option 2 based on the degree to which showing the soil conditions explicity help interpret the results or just cloud the picture.

It is sometimes more conceptually representative to omit some of the complexities from the diagram when reporting results from multiple composite effects.



On rare occasions you could simplify results from a model to avoid having the machinery distract from the message. I would omit path coefficients on the diagram for the links from causes to the concept/composite, but show the coefficients for links from concept to responses. Report all results in a table for full disclosure.

USGS

We are, as always, allowed some creative license when presenting the results in picture form because we wish to convey our results to the reader as simply as possible. ALWAYS present the full results, un abridged in a table or appendix.