USGS
science for a changing world

Latent Variables: Confirmatory Factor Analysis

Jim Grace

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

1

This material is covered in

Grace, J.B. 2006. Structural Equation Modeling and Natural Systems. Cambridge University Press

The example presented is adapted from

Matteson, K.C., Grace, J.B., and Minor, E.S. 2012. Direct and indirect effects of land use on floral resources and flower-visiting insects across an urban landscape. *Oikos* 122:682-694.

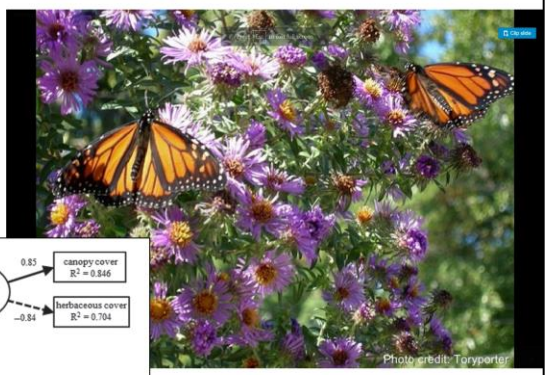
Notes: IP-056512; Support provided by the USGS Climate & Land Use R&D and Ecosystems Programs. I would like to acknowledge formal review of this material by Jesse Miller and Phil Hahn, University of Wisconsin. Many helpful informal comments have contributed to the final version of this presentation.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

Last revised 17.02.08.

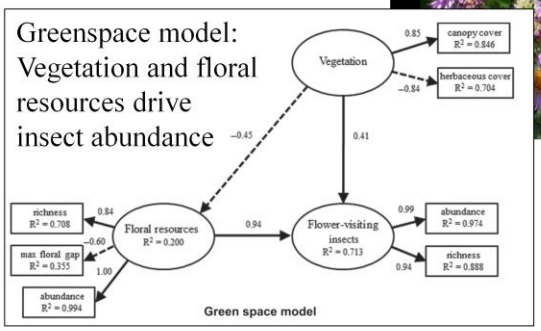
Source: https://www.usgs.gov/centers/wetland-and-aquatic-research-center/science/quantitative-analysis-using-structural-equation

Example CFA: Floral Resources and Flower-Visiting Insects

Matteson, Grace, and Minor (2012) Direct and indirect effects of land use on floral resources and flower-visiting insects across an urban landscape. *Oikos* 122:682-694.



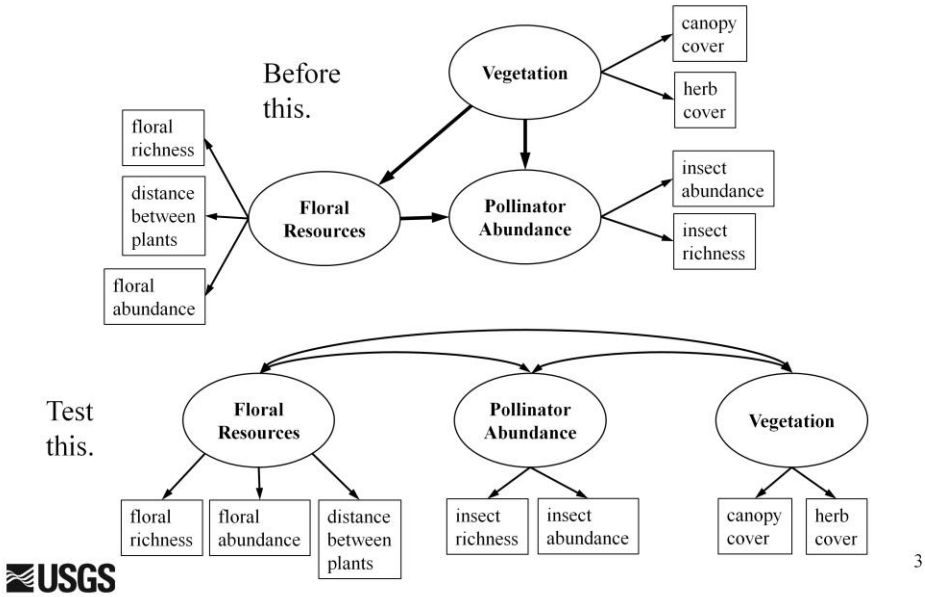
Greenspace model: Vegetation and floral resources drive insect abundance



USGS

The example is from a study of urban pollinating insects and their dependence on floral resources.

Before we analyze a full, multi-indicator LV model, we must first evaluate the LV – Indicator relationships in a CFA.



USGS

3

The standard 2-step method for evaluating SE models with LVs having multiple indicators is done in 2 steps. First, you must evaluate the measurement part of the hypothesis. To do this, you allow all LVs to freely intercorrelate (they are exogenous in a CFA usually, so this happens automatically in lavaan).

Procedural steps:

1. Develop, fit, and test CFA model (i.e., test measurement validity).

2. Allow for directed relationships among LVs and test full model.

USGS

4

Here is a further description of the usual 2-step modeling evaluation.

Example Set Up for Greenspace Model

```
### Read data
setwd("")
dat <- read.csv("SEM.9.2-CFA_data.csv")

### Rename and create variables
# Create sem.dat
sem.dat <-  with(dat,data.frame(Sample))
sem.dat$InsectRich <- with(dat, LogPollRich)
sem.dat$InsectAbund <- with(dat, LogPollAb)
sem.dat$FloralRich <- with(dat, LogGenRich)
sem.dat$FloralAbund <- with(dat, LogFloralAb)
sem.dat$FloralGap <- with(dat, LogMaxGap)
sem.dat$HerbCover <- with(dat, Herb30m)
sem.dat$CanopyCover <- with(dat, Canopy30m)

# Recoding to make positive relation to other indicators
sem.dat$FloralDensity <- with(dat, max(LogMaxGap)-LogMaxGap)
sem.dat$CanopyOpenness <- with(dat, max(Canopy30m)-Canopy30m)
```

USGS

5

I went ahead and reverse-coded two indicators that are negatively correlated, which is generally good practice for multi-indicator latent variable models. I don't actually end up using the reflected measures in this case, but generally it is a good idea because it makes convergence easier for the algorithms.

Specify Greenspace CFA

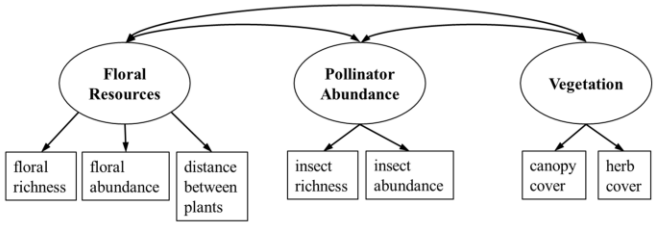
```
### Load needed libraries
library(lavaan)
library(AICcmodavg)
source(". . ./lavaan.modavg.R")

# Specify model
cfa1 <-
'Vegetation           =~ CanopyCover +HerbCover
 FloralResources      =~ FloralRich +FloralAbund +FloralGap
 PollinatorAbundance =~ InsectRich +InsectAbund'
```


USGS

Model specification is straightforward (assuming you have already looked at the module on Modeling with Latent Variables.

Before going further, check the correlations among indicators!




```
# Inspect indicator correlation patterns
indicators <- with(sem.dat,
data.frame(CanopyCover, HerbCover, FloralRich, FloralAbund,
           FloralGap, InsectRich, InsectAbund))

print(cor(indicators), digits=2)
```

USGS

7

Very important to look at the patterns of correlations in the data to see if they match conceptual expectations (on next slide).

Correlations among indicators – "convergent validity"

	CanopyCover	HerbCover	FloralRich	FloralAbund	FloralGap	InsectRich	InsectAbund
CanopyCover	1.000						
HerbCover	-0.793	1.000					
FloralRich	-0.434	0.215	1.00				
FloralAbund	-0.453	0.289	0.84	1.00			
FloralGap	0.367	-0.308	-0.55	-0.59	1.00		
InsectRich	-0.020	-0.058	0.57	0.68	-0.27	1.000	
InsectAbund	-0.041	-0.027	0.62	0.76	-0.32	0.933	1.000

Correlations between indicators – "discriminant validity"

	CanopyCover	HerbCover	FloralRich	FloralAbund	FloralGap	InsectRich	InsectAbund
CanopyCover	1.000						
HerbCover	-0.793	1.000					
FloralRich	-0.434	0.215	1.00				
FloralAbund	-0.453	0.289	0.84	1.00			
FloralGap	0.367	-0.308	-0.55	-0.59	1.00		
InsectRich	-0.020	-0.058	0.57	0.68	-0.27	1.000	
InsectAbund	-0.041	-0.027	0.62	0.76	-0.32	0.933	1.000

We expect to see solid, consistent correlations among indicators of an LV, and weaker correlations among indicators of different LVs.

USGS

8

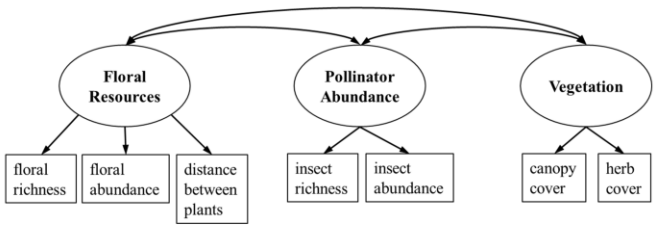
I see some indications of complexities in the measurement hypothesis. For example, floral gap stands out among indicators for floral pollinators. Generally, discriminant validity looks pretty good (except for the problems created by floral gap).

If we go ahead with original model - initial results for "cfa1"

```
# Fit model using lavaan's "cfa" function.
cfa1.fit <- cfa(cfa1, data=sem.dat)

Warning message: In lav_object_post_check(lavobject) :
  lavaan WARNING: some estimated ov variances are negative
```

There are several things that could cause this warning message:
(1) slightly negative error estimates (not a real problem),
(2) need to code indicators for an LV to be positively related,
(3) local non-identification,
(4) general misspecification.



USGS

This slide presents some general thoughts about negative indicator variances, which are known in the SEM literature as "Haywood Cases".

Inspect initial results: Model Fit

Convergence looks OK.

```
> summary(cfa1.fit)
lavaan (0.5-22) converged normally after  93 iterations

  Number of observations                           450

  Estimator                                         ML
  Minimum Function Test Statistic               86.244
  Degrees of freedom                                11
  P-value (Chi-square)                           0.000
```

Measurement hypothesis fails initial test.

However, keep in mind power is very high since n = 450, thus we will want to examine other fit indices that are not dependent on sample size.

USGS

Of course we always look for missing linkages first.

Inspect initial results: LV to indicator relations

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  Vegetation =~
    CanopyCover       1.000
    HerbCover        -0.534    0.033  -16.087    0.000
  FloralResources =~
    FloralRich        1.000
    FloralAbund       1.587    0.054   29.653    0.000
    FloralGap        -0.458    0.032  -14.149    0.000
  PollinatorAbundance =~
    InsectRich        1.000
    InsectAbund       1.891    0.043   44.129    0.000
```

Note that loadings are fixed from first-mentioned indicators to LVs. This is necessary to identify the parameters associated with the LVs.

11

USGS

Since at least one indicator for each LV is set by lavaan to a loading of 1.0, those values are fixed and not estimated.

Inspect initial results: error variances for indicators

Variances: (these are error variances)

	Estimate	Std.Err	z-value	P(>|z|)
.CanopyCover	-0.007	0.003	-1.988	0.047
.HerbCover	0.015	0.001	10.964	0.000
.FloralRich	0.026	0.002	13.228	0.000
.FloralAbund	0.001	0.002	0.395	0.693
.FloralGap	0.024	0.002	14.861	0.000
.InsectRich	0.020	0.002	10.440	0.000
.InsectAbund	-0.011	0.005	-2.205	0.027
Vegetation	0.071	0.005	13.095	0.000
FloralResourcs	0.063	0.006	10.961	0.000
PollintrAbndnc	0.112	0.009	12.672	0.000

The negative error variances are near zero
(keep in mind n = 450, so power is very high).

Also, it is too early to worry about this, because we may decide to include additional links in the model.

12

USGS

Standard practice in SEM software output is to simply state "Variances" when referring to the "Error Variances".

Decision Time:
Do we keep original hypothesis and search for solutions to resolve low correlation between FloralGap and other indicators of floral resources?

Or,
Do we drop FloralGap as an indicator?

USGS

13

Let's try to resolve the model so we can retain this indicator.

Moving forward with model cfa1 – closer look at model fit

```
# Deeper look at model fit
fitMeasures(cfa1.fit, c("gfi", "agfi", "cfi", "rni"))
```

```
  gfi   cfi   rni
0.950 0.973 0.973
```

These indexes have generally been used with a conventional cutoff in which values larger than .95 are considered good fitting models.

So, these fit indices suggest most of the total possible discrepancy is explained by the model.

USGS

For additional information in fit measures, refer to the tutorial on Model Evaluation.

Still, let's dig a little deeper: Error correlations.

```
# Request modification indices - error correlations only
subset(modindices(cfa1.fit), mi > 3.0 & op == "~~")
```

```
> subset(modindices(cfa1.fit), mi > 4.0 & op == "~~")

           lhs op         rhs     mi    epc sepc.lv sepc.all sepc.nox
36 CanopyCover ~~  FloralRich 27.160 -0.006  -0.006   -0.081   -0.081
37 CanopyCover ~~ FloralAbund 15.129  0.007   0.007    0.067    0.067
38 CanopyCover ~~   FloralGap  4.821 -0.002  -0.002   -0.045   -0.045
41   HerbCover ~~  FloralRich 20.952 -0.004  -0.004   -0.072   -0.072
42   HerbCover ~~ FloralAbund  8.977  0.004   0.004    0.050    0.050
43   HerbCover ~~   FloralGap  8.655 -0.002  -0.002   -0.068   -0.068
46  FloralRich ~~ FloralAbund 18.692 -0.017  -0.017   -0.142   -0.142
47  FloralRich ~~   FloralGap  7.449 -0.003  -0.003   -0.060   -0.060
52 FloralAbund ~~ InsectAbund  7.648  0.007   0.007    0.028    0.028
54   FloralGap ~~ InsectAbund  5.849  0.003   0.003    0.029    0.029
```

Error correlations among indicators of the same LV are easy to understand.

Error correlations across LVs seem more likely to represent causal effects at the LV level. For example, effects of Vegetation indicators on Floral Resources indicators is expected.

USGS

15

Interpreting modification indices for CFA models is like reading tea leaves. One needs to let theoretical thinking carry a lot of weight. It looks like there are many missing connections. As will be shown in the rest of the tutorial, once we include the logically obvious error correlation between FloralRich and FloralAbund, nearly all the others disappear! The lesson is you must let theory guide your thinking here.

Still, let's dig a little deeper: Cross-loadings.

```
# Request modification indices - cross-loadings only
subset(modindices(cfa1.fit), mi > 4.0 & op == "=~")
```

```
>subset(modindices(cfa1.fit), mi > 4.0 & op == "=~")

                lhs op             rhs     mi    epc
21          Vegetation =~  FloralRich  9.091 -0.094
22          Vegetation =~ FloralAbund 13.775  0.169
23          Vegetation =~   FloralGap  4.349  0.060
33 PollinatorAbundance =~ FloralAbund 22.914  0.300
34 PollinatorAbundance =~   FloralGap 32.545  0.189
```

None of these are conceptually compelling to me as indicator relationships.

USGS

16

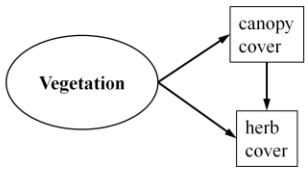
A cross loading is where we choose to interpret the indicator from another factor (LV) to become part of the measurement instrument. Some of these MIs are very large, but it is best to ignore them because there is no strong conceptual basis for saying, for example, that Floral Gap is a measure of Pollinator Abundance.

Still, let's dig a little deeper: Between-indicator effects.

```
# Request modification indices - between-indicator effects
subset(modindices(cfa1.fit), mi > 4.0 & op == "~")
```

```
> subset(modindices(cfa1.fit), mi > 4.0 & op == "~")
[1] lhs      op       rhs      mi       epc
<0 rows> (or 0-length row.names)
```

This would be an example of a between-indicator effect.



No between-indicator effects are indicated.

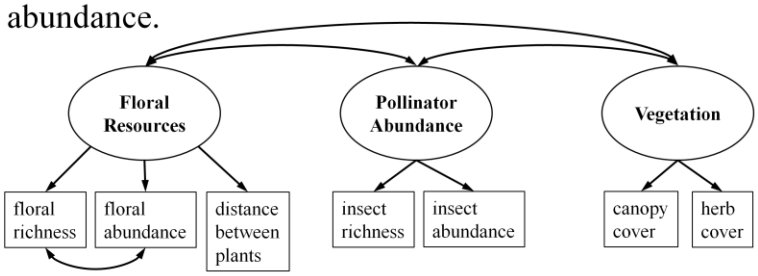
(keep in mind this would be a residual relationship once the common variance caused by the vegetation effect is removed).

17

USGS

Causal modeling requires one to stay flexible in model specification and true to causal modeling logic.

cfa2 is model with error correlation between floral richness and abundance.




```
# Specify new model, "cfa2"
cfa2 <-
'Vegetation          =~ CanopyCover +HerbCover
 FloralResources     =~ FloralRich +FloralAbund +FloralGap
 PollinatorAbundance =~ InsectRich +InsectAbund
 FloralRich ~~ FloralAbund'
```

```
> anova(cfa1.fit, cfa2.fit)
Chi Square Difference Test
         Df     AIC     BIC  Chisq Chisq diff Df diff Pr(>Chisq)
cfa2.fit 10 -1274.9 -1200.9 63.936
cfa1.fit 11 -1254.6 -1184.7 86.244     22.308       1  2.323e-06
```

cfa2 superior

18

It is well known that the number of species found in a small plot or sample is strongly dependent on the number of plants counted in that plot. Species accumulate as you count individuals. So, there is a very strong basis for thinking these two indicators of Floral Resources will be tightly linked, while distance between plant patches is somewhat less tied to the others.

Additional error correlations.

```
# error correlations only
subset(modindices(cfa1.fit), mi > 3.0 & op == "~~")
```

```
> subset(modindices(cfa2.fit), mi > 4.0 & op == "~~")

           lhs op          rhs    mi    epc sepc.lv sepc.all sepc.nox
44   HerbCover ~~    FloralGap 6.659 -0.002  -0.002   -0.069   -0.069
52 FloralAbund ~~ InsectAbund 4.403  0.023   0.023    0.091    0.091
```

All the large error correlations are now gone. Remaining ones are small and illogical. Therefore, I would be inclined to accept the new model "cfa2".

19

This radical reduction in MI values resulting from the one link added is common when working with CFA models.

Results from the revised CFA model "cfa2"

```
> summary(cfa2.fit, rsq=TRUE, standardized=TRUE)
lavaan (0.5-22) converged normally after  90 iterations

  Number of observations                           450

  Estimator                                         ML
  Minimum Function Test Statistic               63.936
  Degrees of freedom                                10
  P-value (Chi-square)                           0.000

> fitMeasures(cfa2.fit, c("gfi", "cfi", "rni"))
  gfi   cfi   rni
0.962 0.981 0.981
```

The absolute fit measures are all well above 0.95 now, indicating very close fit.

USGS

This level of absolute fit is excellent for such models and a sample size this big (i.e., this much statistical power).

Results from the revised CFA model "cfa2" (cont.)

```
Latent Variables:
                          Estimate  Std.Err  z-value  P(>|z|)   Std.all
  Vegetation =~
    CanopyCover              1.000                              1.080
    HerbCover               -0.506    0.034  -14.942    0.000   -0.734
  FloralResources =~
    FloralRich               1.000                              0.947
    FloralAbund              1.535    0.050   30.566    0.000    1.085
    FloralGap               -0.377    0.034  -10.994    0.000   -0.551
  PollinatorAbundance =~
    InsectRich               1.000                              0.922
    InsectAbund              1.883    0.042   44.848    0.000    1.012
```

Our model would be better behaved if we used the following process for the 2-indicator LVs:

(1) standardized the variables and

(2) estimated one loading for both indicators to achieve local identification.

21

USGS

Demonstration of this procedure is planned for another module.

Results from the revised CFA model "cfa2" (cont.)

Variances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.CanopyCover	-0.011	0.004	-2.753	0.006	-0.011	-0.167
.HerbCover	0.016	0.001	11.182	0.000	0.016	0.462
.FloralRich	0.009	0.004	2.115	0.034	0.009	0.104
.FloralAbund	-0.028	0.008	-3.381	0.001	-0.028	-0.177
.FloralGap	0.026	0.002	14.433	0.000	0.026	0.696
.InsectRich	0.020	0.002	10.495	0.000	0.020	0.151
.InsectAbund	-0.009	0.005	-1.937	0.053	-0.009	-0.024
Vegetation	0.075	0.006	13.144	0.000	1.000	1.000
FloralResourcs	0.079	0.007	10.896	0.000	1.000	1.000
PollintrAbndnc	0.112	0.009	12.733	0.000	1.000	1.000

All of the negative error variances are close enough to zero to be tolerable.

USGS

Here are some more results.

Results from the revised CFA model "cfa2" (cont.)

```
R-Square:
                  Estimate
    CanopyCover         NA
    HerbCover        0.538
    FloralRich       0.896
    FloralAbund         NA
    FloralGap        0.304
    InsectRich       0.849
    InsectAbund         NA
```

The R-squares are non-estimable for those indicators whose loadings with the LVs is fixed to 1.0.

USGS

And a bit more results.