Modèles d'équations structuraux

Contents

Introduction

Les modèles d'équations structurelles appartiennent à une famille de modèles qui consistent en un ensemble d'équations mathématiques et d'hypothèses sur un système étudié. Ces hypothèses découlent de nos connaissances préalables ou de nos suppositions sur le fonctionnement du système. En statistique, un système est un ensemble de variables ou de phénomènes qui sont étudiés dans le cadre d'une analyse ou d'une étude. Ces variables peuvent être liées par des relations complexes. Notre objectif avec l'analyse est de comprendre comment ces variables interagissent les unes avec les autres ou comment elles influencent un résultat ou un phénomène en particulier.

Contenu du cours

- Types de variables et relations dans un modèle d'équations structurelles ;
- Du modèle théorique au modèle statistique ;
- Ajustement du modèle dans lavaan.

Types des variables et liens dans un modèle d'équation structurel:

La structure théorique d'un modèle d'équations structurelles englobe plusieurs types de variables, définissant leurs caractéristiques et leurs rôles dans le modèle.

Selon leur **nature**, les variables peuvent être classées en 1) **variables latentes** et 2) **variables observées**. Une **variable latente** est une variable qui n'est pas mesurée directement, mais qui représente des concepts ou des traits qui n'ont pas vraiment d'unité de mesure. Il s'agit d'un type de variable plus souvent utilisé en psychologie (l'intelligence, la satisfaction, etc.). Les **variables observées** sont des variables mesurées ou collectées avec des méthodes établies par la discipline. En écologie, nous travaillons souvent avec des variables observées.

Selon le **rôle** des variables dans le modèle théorique, les variables peuvent être **exogènes** ou **endogènes**. Les **variables exogènes** sont des variables indépendantes qui influencent les autres variables dans les modèles, mais qui ne sont pas influencées par aucune autre variable en retour. Elles représentent les moteurs des changements dans notre système. Les **variables endogènes**, vice-versa, sont des variables qui sont influencées

par les variables exogènes ou d'autres variables du modèle, et représentent normalement le noyau de notre système et les résultats des processus que nous sommes en train de décrire avec notre modèle.

Selon la façon dont les variables sont conceptualisées dans le modèles, les variables peuvent avoir un statut de modérateur et de médiateur. Un **modérateur** est une variable qui influence la force et la direction du lien entre deux variables. Une variable modératrice n'explique pas les "causes" du lien, mais elle intervient seulement dans les aspects quantitatives de la relation étudiée entre deux variables. Un **médiateur** est une variable qui explique le lien entre une variable indépendante et une variable dépendante. Une variable agit comme variable médiatrice lorsque elle représente la cause du lien entre deux variables.

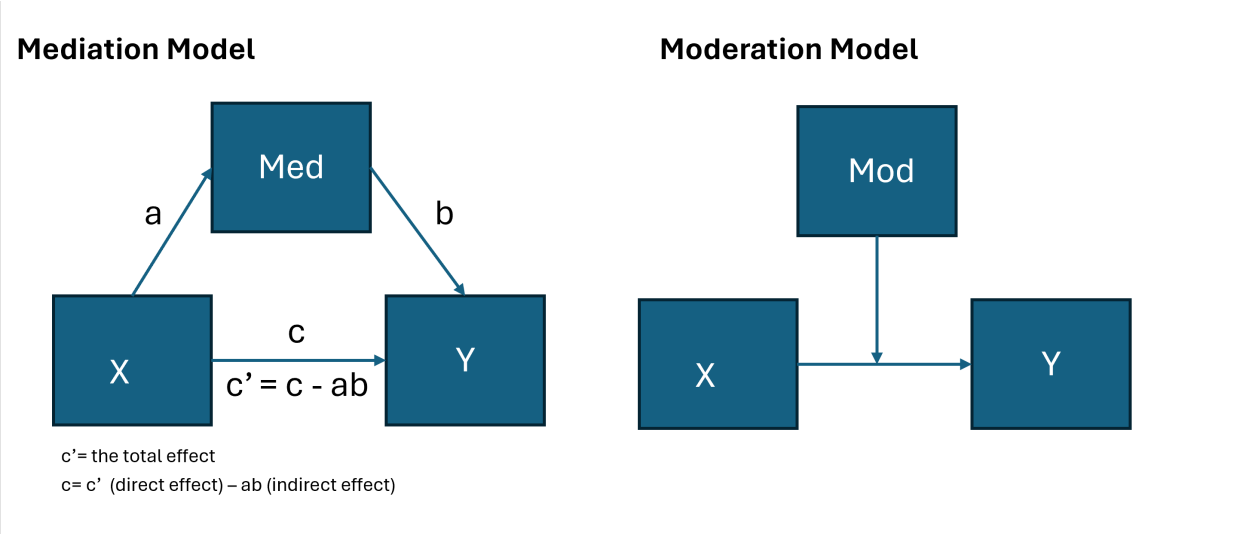


Figure 1: Modèles pour une analyses de Médiation et une analyses des moderateurs

Dans la figure, les fléchés représentent des **chemin**. Dans un modèle d'équations structurelles un chémin répresnete une relation qui relie deux variables, pouvant être directe ou indirecte, et pouvant inclure des effets médiatisés. Comme vous pouvez le voir dans l'image, les effets médiatisés se réfèrent à une situation où l'impact d'une variable indépendante sur une variable dépendante passe par une autre variable. Autrement dit, la variable médiatrice transmet ou médie l'effet de la variable indépendante sur la variable dépendante.

Dans un modèle d'équations structurelles, l'analyse de médiation et e modération sont intégrées pour obtenir une compréhension plus profonde des relations entre nos variables, en tenant compte à la fois de processus médiatiques et des effets des modérateurs. Cependant, on peut réaliser des analyses des modérateurs et de médiateurs si on souhaite tester des les relations entres trois variables avec le package Psych:

La fonction 'mediate' du package 'psych' vous permet de conduire une analyse de médiation avec la fonction 'mediate'. La variable médiatrice doit être mise entre parenthèses afin d'informer la fonction de son rôle dans le modèle. Dans l'exemple que nous allons utiliser pour montrer la fonction, nous utilisons la fonction 'mediate' pour tester les effets directs et indirects de la température sur la largeur des cernes de croissance. En effet, dans l'hémisphère nord et dans des environnements froids où la température est un facteur limitant pour la croissance, la largeur des cernes de croissance augmente avec la température. Cependant, la largeur du cerne de croissance est étroitement liée au nombre de cellules du bois qui composent le cerne, et qui est, à son tour, reliée à la température. Ainsi, une partie de l'effet total de la température sur la largeur du cerne est en effet médiée par le nombre de cellules. Si l'effet direct de la température sur le nombre de cellules est plus grand par rapport à celui sur la largeur du cerne, les changements de température seront plus liés aux changements du nombre de cellules qu'à la largeur du cerne, et donc ces derniers seront plus facilement prédictibles.

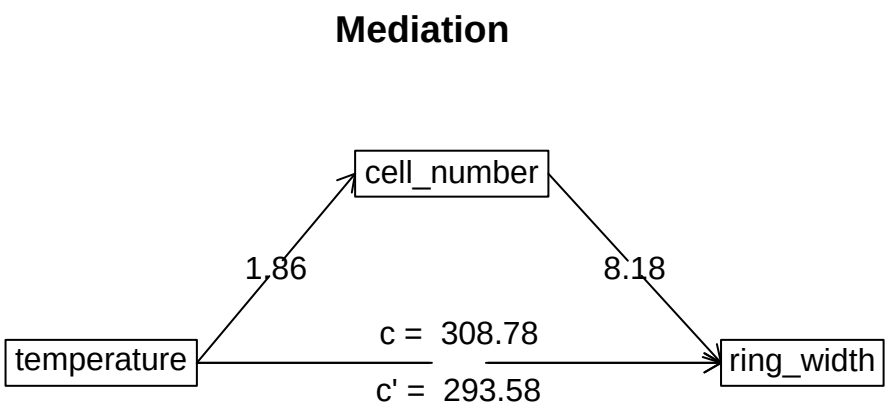
```
require(psych)
```

```
## Le chargement a nécessité le package : psych
```

```
## Warning: le package 'psych' a été compilé avec la version R 4.3.3
```

```
ringdatacell <- read.csv("C:/Users/buttoval/Documents/ECL8202/donnees//ringdatacell.CSV")


medanalysys<-mediate( ring_width ~ temperature + (cell_number) , data=ringdatacell )
```



Dans le schéma on voit très bien que l'effet totale de la température est c = 8.15, mais l'effet direct de la temperature sur le cerne de croissance c'= 0.56, est beaucoup plus petit.

Nota bene la fonction "mediate" nous donne les coefficients standardisées par defaut, ainsi que les moyennes centrées. pour éviter cela il faut utiliser les aurgments "zero=FALSE" et std=FALSE.

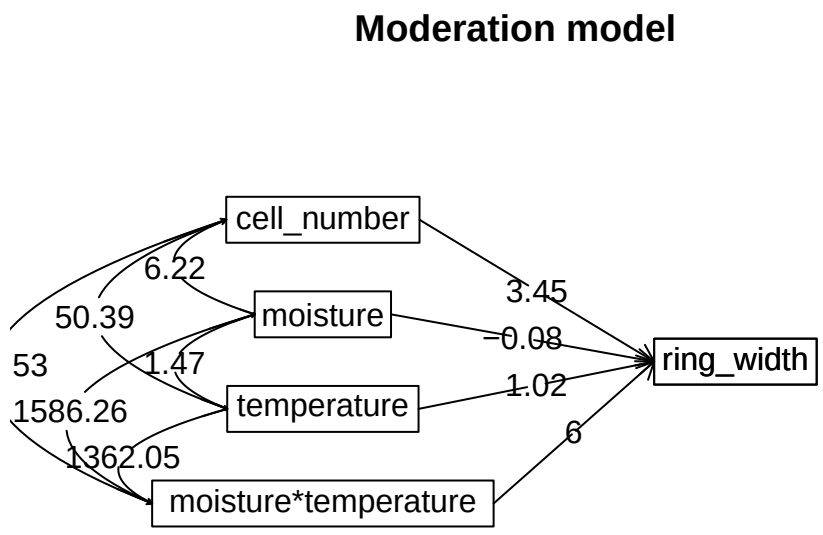
```
summary(medanalysys)
```

```
## Call: mediate(y = ring_width ~ temperature + (cell_number), data = ringdatacell)
##
## Direct effect estimates (traditional regression)    (c') X + M on Y
##             ring_width     se     t df     Prob
## Intercept      -157.63 459.43 -0.34 97 7.32e-01
## temperature     293.58  44.61  6.58 97 2.38e-09
## cell_number       8.18  21.29  0.38 97 7.02e-01
##
## R = 0.84 R2 = 0.7   F = 112.67 on 2 and 97 DF   p-value:  5.08e-26
##
##  Total effect estimates (c) (X on Y)
##             ring_width     se     t df     Prob
## Intercept       -93.82 426.50 -0.22 98 8.26e-01
```

```
## temperature     308.78  20.49 15.07 98 2.88e-27
##
##  'a'  effect estimates (X on M)
##             cell_number   se     t df     Prob
## Intercept          7.80 2.03  3.84 98 2.20e-04
## temperature        1.86 0.10 19.04 98 1.06e-34
##
##  'b'  effect estimates (M on Y controlling for X)
##             ring_width    se    t df  Prob
## cell_number       8.18 21.29 0.38 97 0.702
##
##  'ab'  effect estimates (through all  mediators)
##             ring_width  boot    sd  lower upper
## temperature      15.21 13.39 42.94 -72.05 93.59
```

Dans une analyse de modération, on examine comment l'effet d'une variable indépendante sur une variable dépendante peut être modifié par une autre variable, appelée modérateur. Pour évaluer cette interaction, on inclut généralement un terme d'interaction dans le modèle de régression multiple. Ce terme d'interaction permet de tester si l'effet de la variable indépendante sur la variable dépendante varie en fonction des niveaux du modérateur. En résumé, une analyse de modération est une régression multiple avec une interaction. On peut obtenir un modèle de modération en utilisant la fonction médiate, mais sans spécifier l'effet du modérateur.

```
medanalysys<-mediate( ring_width ~   cell_number  + moisture + moisture*temperature,
                      data=ringdatacell,zero= FALSE,std=FALSE)
```



```r
summary(medanalysys)
```

```
## Call: mediate(y = ring_width ~ cell_number + moisture + moisture *
##     temperature, data = ringdatacell, std = FALSE, zero = FALSE)
##
## No mediator specified leads to traditional regression
##                       ring_width    se      t df      Prob
## Intercept                   9.04 11.16   0.81 95  4.20e-01
## cell_number                 3.45  0.13  27.50 95  4.97e-47
## moisture                   -0.08  0.22  -0.33 95  7.39e-01
## temperature                 1.02  0.66   1.55 95  1.25e-01
## moisture*temperature        6.00  0.01 522.12 95 4.55e-166
##
## R = 1 R2 = 1   F = 2400995 on 4 and 95 DF   p-value:  9.14e-237
```

Du modèle théorique au modèle statistique:

Lorsqu'on décide de réaliser une SEM, une hypothèse bien définie représente le meilleur investissement pour valoriser cette analyse. Nous allons donc classer nos variables selon la typologie définie dans le paragraphe précédent, puis construire notre modèle **a priori**. Ce modèle représente notre compréhension du système basée sur les preuves scientifiques que nous avons recueillies dans notre étude, et contient nos hypothèses sous forme de liens entre les variables.

Dans une SEM, on établit un modèle théorique basé sur des relations postulées entre les variables, puis on teste ce modèle avec des données réelles pour voir s'il correspond bien à ces données. L'objectif est de déterminer si le modèle théorique est statistiquement valide et peut être généralisé aux données réelles. La validation du modèle nécessite donc de ne pas rejeter l'hypothèse nulle selon laquelle les relations entre les variables telles qu'elles sont spécifiées dans le modèle théorique sont également présentes dans les données réelles, et que toute différence observée entre le modèle théorique et les données réelles est due au hasard ou à des erreurs de mesures.

Dans un SEM, on peut utilier des symboles peuvent être utilisés pour représenter les liens et les variables dansun diagramme:

Figure 2: Symboles pour les diagrammes

Il y a plusieurs packages qui peuvent vous permettre de réaliser un modèle SEM sur R. Ici nous allons utiliser Lavaan, qui utilise ça propre syntaxe pour définir les variables du modèle et leur liens.

Formule et définition	operateur	signification
variable latente	=~	obtenue à partir de
Covariable	~~	est correlé avec
intercepte	~1	intercepte

Pour un exemple, nous allons utiliser un jeu de données simulées contenants les informations suivantes:

Température : La température moyenne en degrés Celsius enregistrées pendant la saison de croissance. Humidité : Le pourcentage moyen d'humidité relatif enregistré pendant la saison de croissance. Taille de la tige : La taille moyenne de la tige de la plante, en cm nombre de cellules : le nombre total de cellules observées dans chaque cerne de croissance Largeur des cernes : La largeur moyenne des anneaux de croissance, une mesure de la croissance annuelle des arbres.

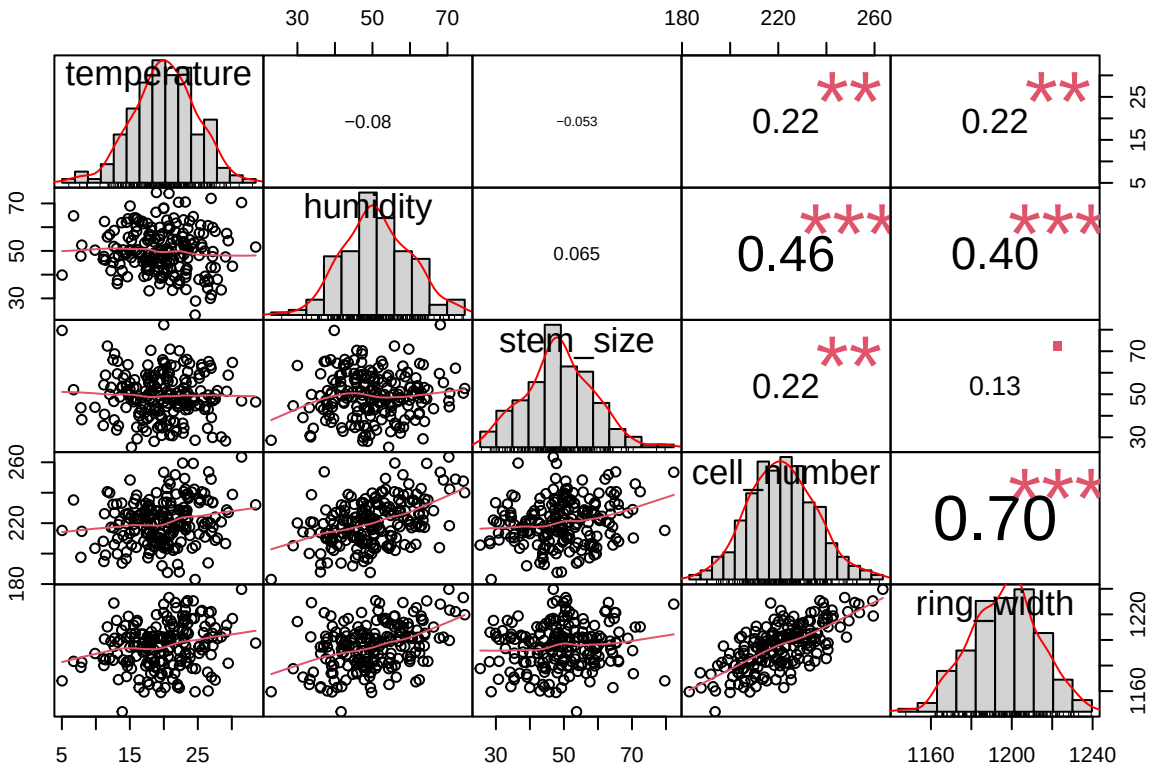
```
simulated_data <- read.csv("C:/Users/buttoval/Documents/ECL8202/donnees/simulatedsemring.csv")
```

Avant d'ajuster un modèle SEM, il est souvent utile de visualiser une matrice de corrélation entre les variables

```
library(PerformanceAnalytics)
```

```
## Le chargement a nécessité le package : xts
## Le chargement a nécessité le package : zoo
##
## Attachement du package : 'zoo'
## Les objets suivants sont masqués depuis 'package:base':
##
##     as.Date, as.Date.numeric
##
## Attachement du package : 'PerformanceAnalytics'
## L'objet suivant est masqué depuis 'package:graphics':
##
##     legend
```

```
chart.Correlation(simulated_data, histogram = TRUE, method = "pearson")
```



La matrice de corrélation montre des corrélations élevées entre la plupart de nos variables, mais elle ne nous dit rien sur les relations entre elles. Un modèle théorique basé sur la littérature est proposé pour expliquer les liens entre nos variables et le processus sous-jacent de la croissance:

Sur lavaan, nous pouvons traduire le modèle avec la syntax suivante:

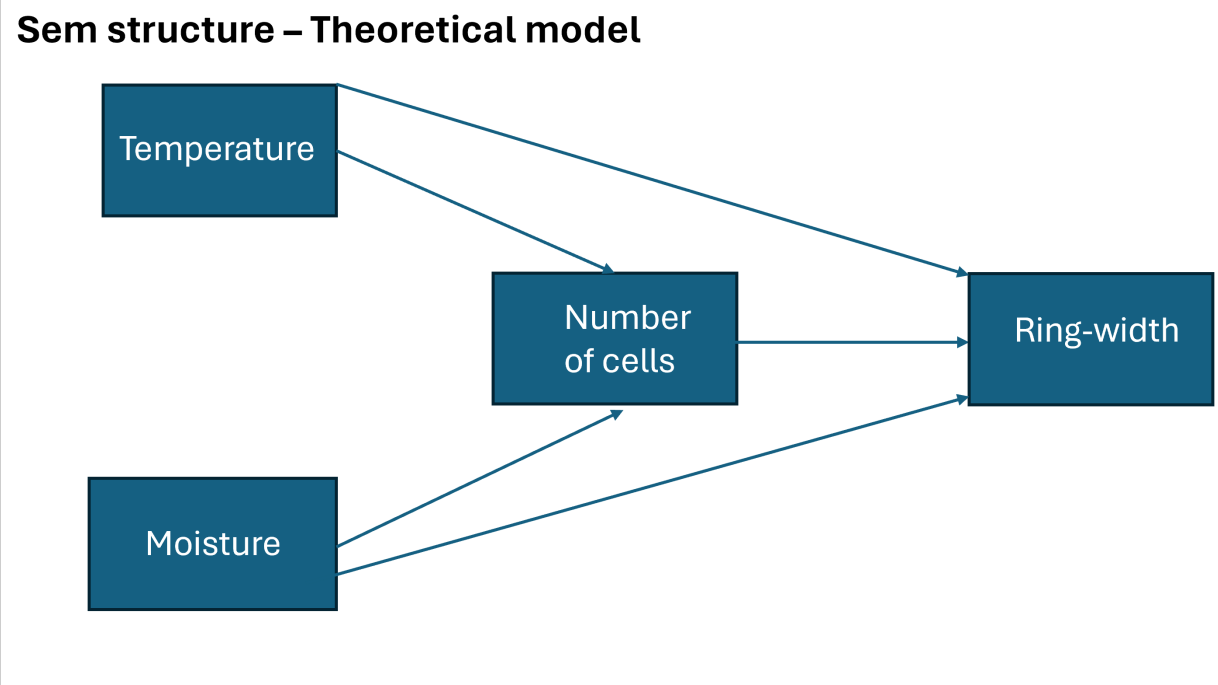


Figure 3: Relations entre les facteurs environnementaux et la croissance: chaque lien est appuyé par des source de litterature

```
require(lavaan)
```

```
## Le chargement a nécessité le package : lavaan
```

```
## This is lavaan 0.6-15
## lavaan is FREE software! Please report any bugs.
```

```
##
## Attachement du package : 'lavaan'
```

```
## L'objet suivant est masqué depuis 'package:psych':
##
##     cor2cov
```

```
myModel <- '

 # regressions
   ring_width ~ temperature + humidity+ cell_number
   cell_number ~  temperature + humidity
   cell_number ~ ~  stem_size

'

fit <- sem(model = myModel,
           data = simulated_data)

summary(fit)
```

```
## lavaan 0.6.15 ended normally after 28 iterations
##
```

```
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                         9
##
##   Number of observations                           200
##
## Model Test User Model:
##
##   Test statistic                                 1.368
##   Degrees of freedom                                 3
##   P-value (Chi-square)                           0.713
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   ring_width ~
##     temperature       0.343    0.187    1.832    0.067
##     humidity          0.235    0.105    2.239    0.025
##     cell_number       0.735    0.069   10.592    0.000
##   cell_number ~
##     temperature       0.803    0.177    4.529    0.000
##     humidity          0.724    0.091    7.934    0.000
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##  .cell_number ~~
##     stem_size        30.682    9.433    3.253    0.001
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .ring_width      150.351   15.035   10.000    0.000
##    .cell_number     156.311   15.631   10.000    0.000
##     stem_size       107.827   10.783   10.000    0.000
```

Ajustement du modèle sur Lavaan

On considère qu'un SEM représente bien notre distribution de données et notre phénomène naturel lorsque le p-value (chi carré) du modèle n'est pas significatif. Cela arrive parce que nous souhaitons observer une correspondance entre les données observées et les prédictions de notre modèle, lesquelles ne doivent pas présenter de différences significatives. Dans ce cas, le p-value est de 0.713, ce qui nous rassure quant à la capacité du modèle à représenter notre phénomène.Il est également utile de vérifier d'autres mesures d'ajustement du modèle, pour obtenir une évaluation plus complète de l'adéquation du modèle. Voici les indicateurs plus courentment utilisés. Voici une synthèse présentée par Joreskog, K., & Sorbom, D. (1993):

Il existe des standards pour nous naviguer dans la présentation de ces indicateurs:

Le résumé de la fonction sem nous fournit ainsi les coefficients de régression qui indiquent la force et la direction de la relation entre nos variables. Les coefficients de covariance mesurent également la force et la direction de la corrélation entre les variables. On peut interpréter ces coefficients comme n'importe quels

Table 1: Fit indices and their acceptable thresholds

Fit Index	Acceptable Threshold Levels	Description
Absolute Fit Indices		
Chi-Square χ^2	Low χ^2 relative to degrees of freedom with an insignificant *p* value ($p > 0.05$)	
Relative χ^2 (χ^2/df)	2:1 (Tabachnik and Fidell, 2007) 3:1 (Kline, 2005)	Adjusts for sample size.
Root Mean Square Error of Approximation (RMSEA)	Values less than 0.07 (Steiger, 2007)	Has a known distribution. Favours parsimony. Values less than 0.03 represent excellent fit.
GFI	Values greater than 0.95	Scaled between 0 and 1, with higher values indicating better model fit. This statistic should be used with caution.
AGFI	Values greater than 0.95	Adjusts the GFI based on the number of parameters in the model. Values can fall outside the 0-1.0 range.
RMR	Good models have small RMR (Tabachnik and Fidell, 2007)	Residual based. The average squared differences between the residuals of the sample covariances and the residuals of the estimated covariances. Unstandardised.
SRMR	SRMR less than 0.08 (Hu and Bentler, 1999)	Standardised version of the RMR. Easier to interpret due to its standardised nature.
Incremental Fit Indices		
NFI	Values greater than 0.95	Assesses fit relative to a baseline model which assumes no covariances between the observed variables. Has a tendency to overestimate fit in small samples.
NNFI (TLI)	Values greater than 0.95	Non-normed, values can fall outside the 0-1 range. Favours parsimony. Performs well in simulation studies (Sharma et al, 2005; McDonald and Marsh, 1990)
CFI	Values greater than 0.95	Normed, 0-1 range.

Figure 4: Ajoustement du modèle

Table 2: Hu and Bentler's Two-Index Presentation Strategy (1999)

Fit Index Combination	Combinational Rules
NNFI (TLI) and SRMR	NNFI of 0.96 or higher and an SRMR of .09 or lower
RMSEA and SRMR	RMSEA of 0.06 or lower and a SRMR of 0.09 or lower
CFI and SRMR	CFI of .96 or higher and a SRMR of 0.09 or lower

Figure 5: Combinaisons d'indicateurs d'ajoustement

coefficients d'une régression linéaire. Cependant, il peut être très utile de standardiser la valeur des coefficients si l'on souhaite comparer l'effet des différentes variables sur notre variable réponse.

```
summary(fit,standardized=TRUE)
```

```
## lavaan 0.6.15 ended normally after 28 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                         9
##
##   Number of observations                           200
##
## Model Test User Model:
##
##   Test statistic                                 1.368
##   Degrees of freedom                                 3
##   P-value (Chi-square)                           0.713
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   ring_width ~
##     temperature       0.343    0.187    1.832    0.067    0.343    0.096
##     humidity          0.235    0.105    2.239    0.025    0.235    0.128
##     cell_number       0.735    0.069   10.592    0.000    0.735    0.618
##   cell_number ~
##     temperature       0.803    0.177    4.529    0.000    0.803    0.267
##     humidity          0.724    0.091    7.934    0.000    0.724    0.467
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##  .cell_number ~~
##     stem_size        30.682    9.433    3.253    0.001   30.682    0.236
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##    .ring_width      150.351   15.035   10.000    0.000  150.351    0.497
##    .cell_number     156.311   15.631   10.000    0.000  156.311    0.730
##     stem_size       107.827   10.783   10.000    0.000  107.827    1.000
```

Références

Revelle, 2024: How to use the psych package for regression and mediation analysis: https://cran.r-project.org/web/packages/psychTools/vignettes/mediation.pdf

Yves Rosseel (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1-36. URL http://www.jstatsoft.org/v48/i02/ https://lavaan.ugent.be/

Joreskog, K., & Sorbom, D. (1993). Structural equation modelling: Guidelines for determining model fit. NY: University Press of America.