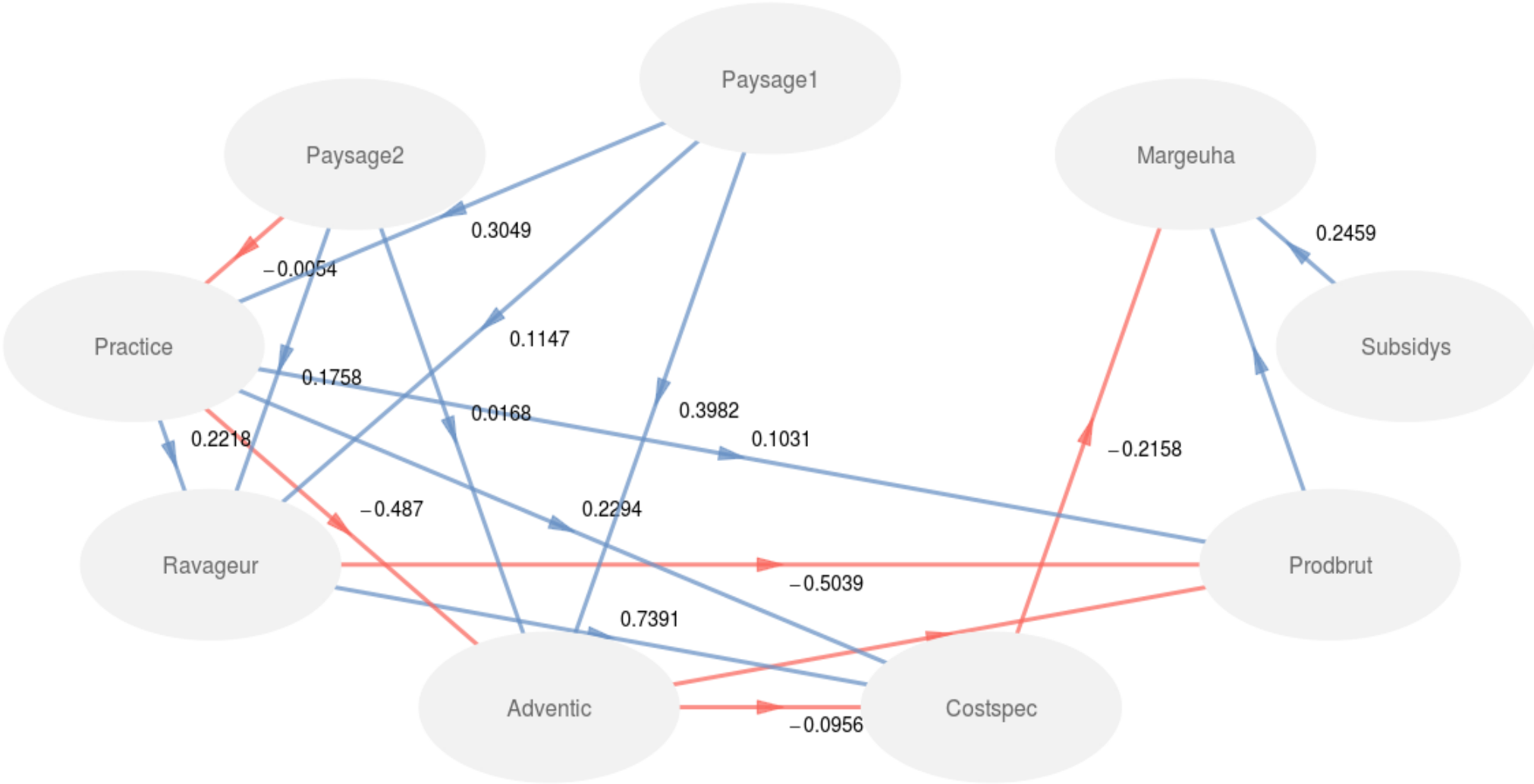
Introduction à l'approche PLS-PM pour la structuration de tableaux



ANF Apprentissage Sémantique, 14 novembre 2018, ENSEEITH Toulouse

Dominique Desbois, UMR Economie publique, INRA-AgroParisTech

Introduction à l'approche **PLS-PM** : **P**artial **L**east **S**quares – **P**ath **M**odelling

- **Problématique** :
  - Définitions ;

- **Méthodologie** :
  - Concepts, modèles et propriétés ;
  - Estimation ;
  - Algorithmes et implantation ;

- **Logiciels** :
  - exemples sous R ;

- **Résultats** :
  - Exemples ;

- **Développements**

- **Bibliographie**

Définition de l'approche PLS-PM

- La modélisation structurelle des moindres carrés partiels (***Partial Least Squares - Path Modelling***, ***PLS-PM***)

est une technique de structuration de tableau de données complexes basée sur des relations linéaires.

- Cette technique est utilisée lorsque des hypothèses de normalité des distributions des observations ne peuvent être testées ou réunies, notamment pour de petits échantillons.

- L'approche PLS-PM est implantée sous **R** par le paquet **plspm** (Sanchez, Trinchera, Russolillo, 2016), disponible sur l'archive CRAN à l'URL:

http://cran.r-project.org/web/packages/plspm/index.html

```
# chargement du paquet pls-pm
library(plspm)
```

Concepts de l'approche PLS-PM

La modélisation structurelle (***P**ath **M**odelling*) est fondée sur le concept de ***variable latente*** qui peut se décliner selon différentes disciplines, par exemple :

- en psychologie, *intelligence* ou *estime de soi*;
- en sociologie, *statut social* ou *structure sociale* ;
- en économie, *utilité* ou *développement économique* ;
- en gestion, *marge* ou *efficacité technique* ;
- en écologie, *fertilité du sol* ou *structure parcellaire* ;

Le concept de variable latente se retrouve dans différents domaines sous les vocables de :

- « *construit* » ou « *variable composite* » ;
- « *complexe de variables* » ;
- « *variable* » « *théorique* » ou « *hypothétique* » ;
- « *facteur* » « *sous-jacent* » ou « *inobservable* ».

Le modèle structurel de l'approche PLS-PM

PLS-PM est une approche de modélisation par équations structurelles (***Path Modelling***), basée sur le critère des moindres carrés partiels (***Partial Least Squares***).

Dans l'approche ***PLS-PM*** , le modèle structurel est un ensemble de construits reliés par des relations causales hypothétiques permettant d'analyser les relations entre plusieurs blocs de variables d'un tableau de données :

- ✓ Chaque bloc de variables joue le rôle d'une **variable latente** (VL) ;
- ✓ L'hypothèse structurelle est celle d'un système de relations linéaires entre blocs (**modèle interne**) supposé rendre compte du fonctionnement théorique du système étudié.

$$VL_j = beta_0 + \sum_{i \to j} beta_{ij} \times VL_i + epsilon_j$$

avec $$cov(VL_j, epsilon_j) = 0$$

- les coefficients **beta** sont les « **liens** » du modèle structurel, pondérations reliant la réponse *VL(j)* à ses déterminants *VL(i)*.

Spécification du modèle structurel PLS-PM

La spécification d'un modèle structurel PLS-PM comprend les étapes suivantes :

- ✓ description du **schéma interne** des variables latentes ;
- ✓ description du **schéma externe** des variables mesurées ;
- ✓ estimation des **paramètres** du modèle.

le modèle structurel du succès d'une équipe de football peut être basé sur l'a priori hypothétique suivant :

la réussite de l'équipe dépend de la qualité de son attaque mais aussi de celle de sa défense

Soit l'équation structurelle :

succès = f(attaque, defense)

Cette relation causale peut être spécifié paramétriquement selon l'équation linéaire suivante :

succes= a*attaque + b*defense

Exemple 1 : championnat de football

Variable	Description
Attaque	
GSH	nombre total de buts à domicile
GSA	nombre total de buts à l'extérieur
SSH	pourcentage de matchs avec buts marqués à domicile
SSA	pourcentage de matchs avec buts marqués à l'extérieur
Défense	
GCH	nombre total de buts concédés à domicile
GCA	nombre total de buts concédés à l'extérieur
CSH	pourcentages de matchs sans buts concédés à domicile
CSA	pourcentages de matchs sans buts concédés à l'extérieur
Succès	
WMH	total de matchs gagnés à domicile
WMA	total de matchs gagnés à l'extérieur
LWR	plus grande séquences de matchs gagnés
LRWL	plus grande séquence de match non perdus
Pénalités	
YC	nombre total de cartons jaunes
RC	nombre total de cartons rouges

Exemple 1 : clubs de football espagnols

```
#       chargement du jeu de données
data(spainfoot)

#       lister les 5 premières lignes de spainfoot
head(spainfoot, n = 5)
##                         GSH GSA SSH SSA GCH  GCA CSH  CSA WMH WMA
LWR LRWL
##      Barcelona    61    44   0.95  0.95  14     21    0.47  0.32  14     13     10     22
##      RealMadrid  49    34   1.00  0.84  29     23    0.37  0.37  14     11     10     18
##      Sevilla           28    26   0.74  0.74  20     19     0.42 0.53  11     10       4
7
##      AtleMadrid   47    33   0.95  0.84  23     34     0.37 0.16  13       7       6      9
##      Villarreal     33    28   0.84  0.68  25     29     0.26 0.16  12       6       5     11

##                         YC     RC
##      Barcelona    76    6
##      RealMadrid  115      9
##      Sevilla              100      8
##      AtleMadrid   116   5
##      Villarreal     102   5
```

Spécification du modèle structurel PLS-PM : **description du schéma interne**

La description du **schéma interne** des variables latentes s'effectue selon les instructions R suivantes :

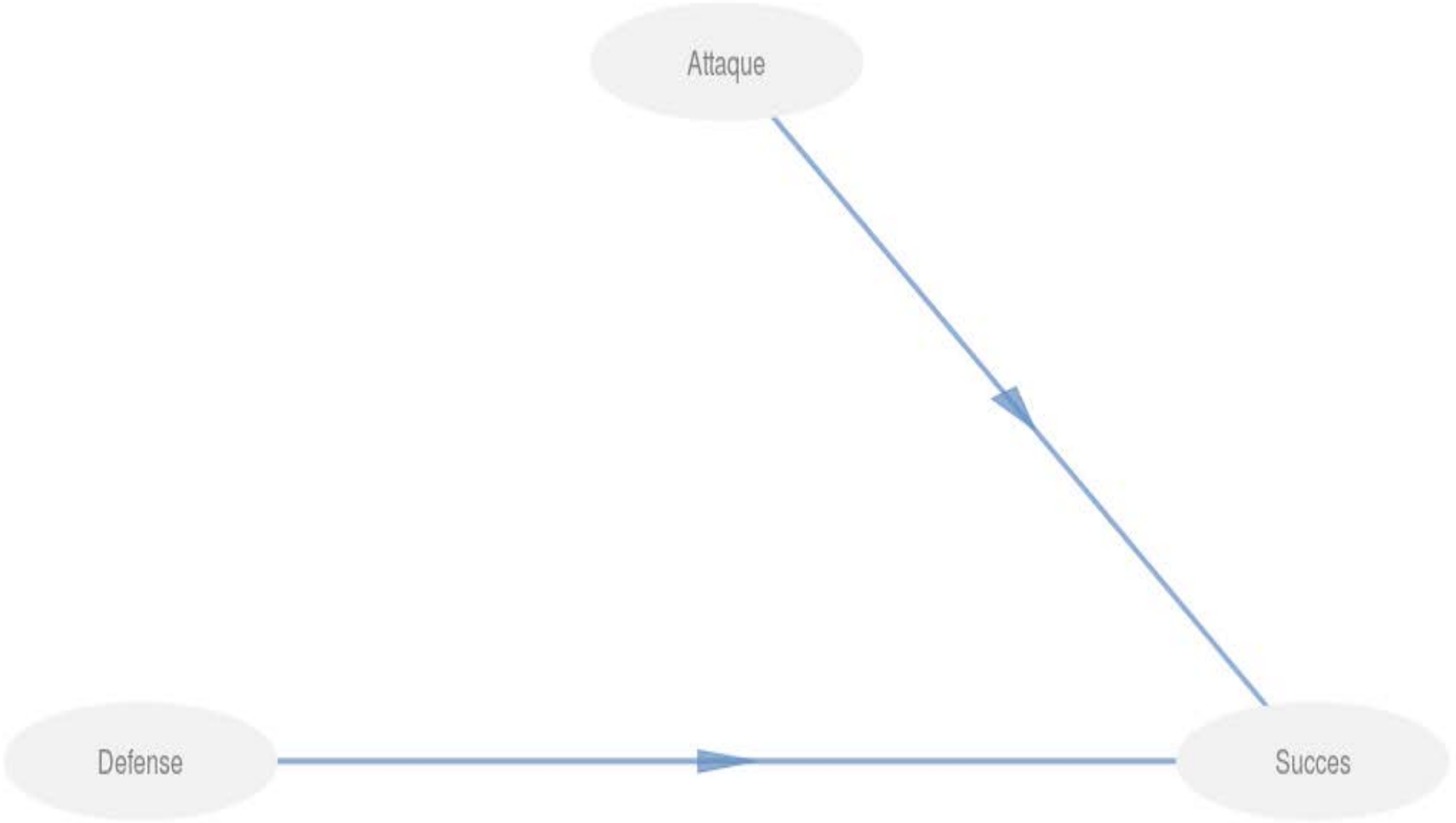
```
# lignes de la matrice de structure
Attaque = c(0, 0, 0)
Defense = c(0, 0, 0)
Succes    = c(1, 1, 0)
# specification  de la matrice de structure
foot_struct = rbind(Attaque, Defense, Succes)
# ajout des noms de colonne
colnames(foot_struct) = rownames(foot_struct)
# matrice de structure
foot_struct
# graphique de structure
innerplot(foot_struct)
```

Spécification du modèle structurel PLS-PM :

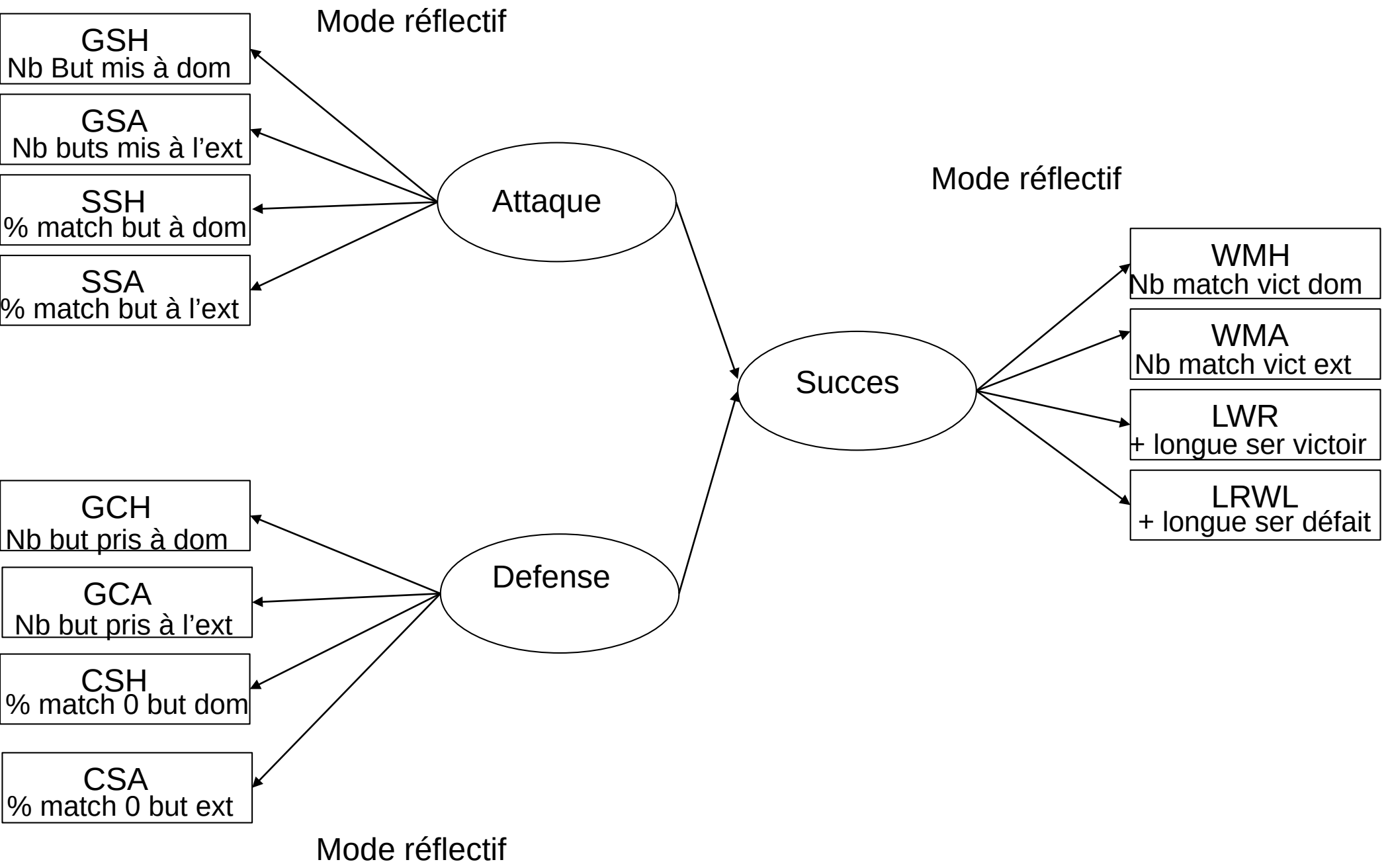
Matrice de structure interne

		Attaque	Defense
Attaque	**0**	**0**	**0**
Defense	**0**	**0**	**0**
Succes	**1**	**1**	**0**

Spécification du modèle structurel PLS-PM : Succès de l'équipe



Spécification du modèle interne PLS-PM : Succès de l'équipe



Modèle structurel PLS-PM : **spécification du schéma externe**

La spécification du **schéma externe** associe les variables mesurées (***variables manifestes***) aux variables latentes (**VL**).

Elle s'effectue selon les instructions R suivantes :

```
# association des variables manifestes (VM)
#  aux variables latentes (VL)
foot_blocs = list(1:4, 5:8, 9:12)
```

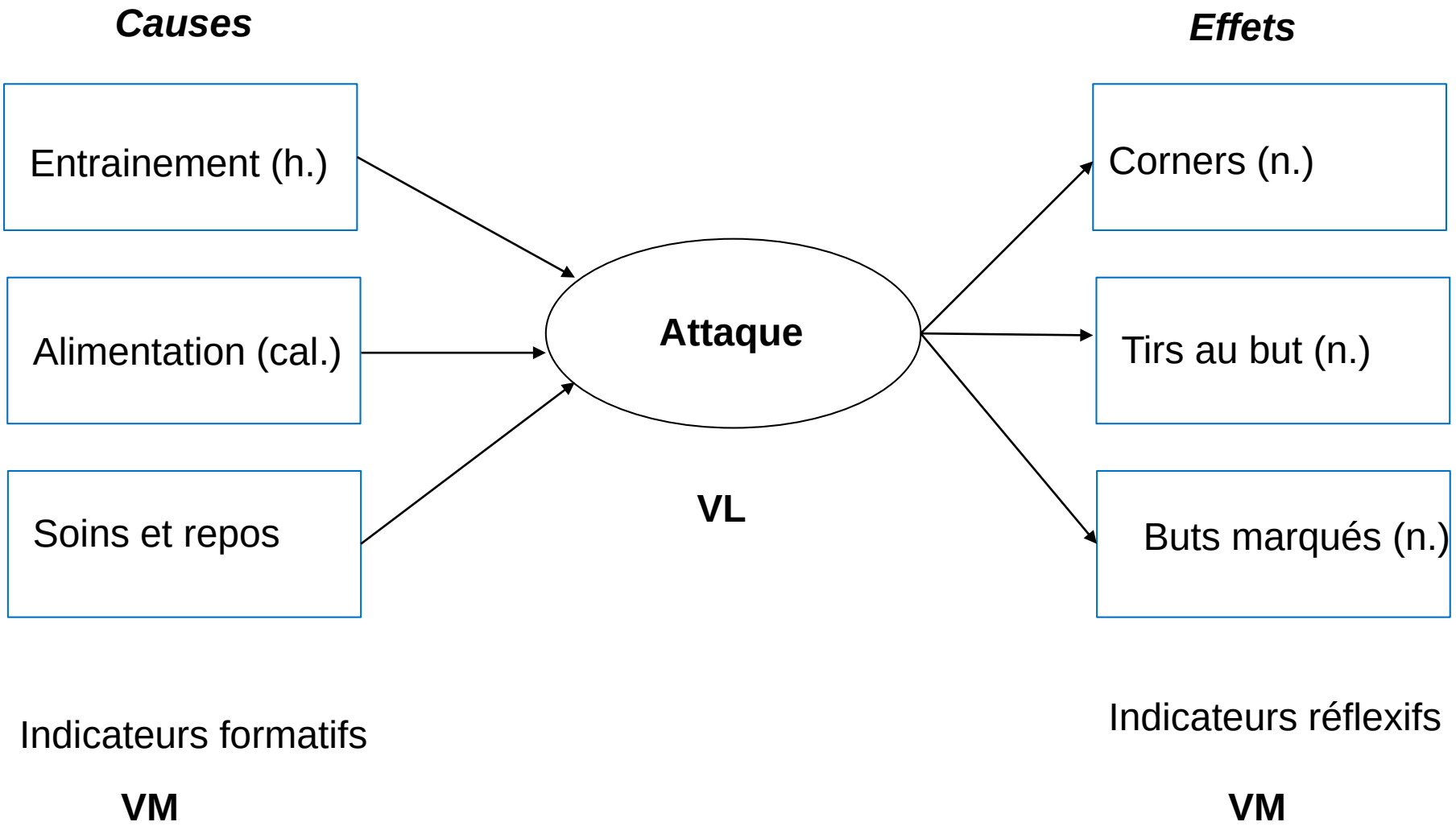
La liste spécifie la composition en variables manifestes (**VM**) de chacun des trois blocs associés aux variables latentes :

- le premier bloc, composé des 4 premières colonnes des données, définit la VL ***Attaque*** ;
- le second bloc, composé des colonnes 5 à 8 des données, définit la VL ***Defense*** ;
- le troisième bloc, composé des colonnes 9 à 12 des données, définit la VL ***Succes*** .

Spécification du modèle structurel PLS-PM : **définition des modes**

Les variables mesurées (***VM***) sont associées aux variables latentes (**VL**) selon un mode qui peut être :

- ✓ soit « ***réflexif*** » (mode ***A***) : $X_{jk} = lambda_{0jk} + lambda_{jk} \times VL_j + epsilon_{jk}$
- soit « ***formatif*** » (mode ***B***) : $VL_j = lambda_{0j} + lambda_{jk} \times X_{jk} + epsilon_j$



Spécification du modèle structurel PLS-PM : **définition des modes**

Les variables mesurées (***VM***) sont associées aux variables latentes (**VL**) selon un mode qui peut être :

- ✓ soit « ***réflectif*** » (mode ***A***) ; $X_{jk} = lambda_{0jk} + lambda_{jk} \times VL_j + epsilon_{jk}$
- soit « ***formatif*** » (mode ***B***). $VL_j = lambda_{0j} + lambda_{jk} \times X_{jk} + epsilon_j$

Le mode le plus courant est le ***mode réflectif*** (A), les variables latentes sont supposées générer les variables mesurées
(les VM « reflètent » les VL) :

 # les trois variables latentes sont mesurées
 # en mode réflectif
✓ ***foot_modes = c("A", "A", "A")***

L'autre mode de mesure, appelé le ***mode formatif*** *(B)*, suppose que les variables manifestes « forment » les variables latentes :

 # Attaque et Defense sont en mode réflectif
 # Succes, en mode formatif
foot_modes = c("A", "A", "B")

Spécification du modèle structurel PLS-PM : **instructions** d'estimation du modèle structurel

L'estimation des paramètres du modèle structurel s'effectue par **exécution de la procédure plspm** selon la syntaxe suivante :

```
# exécution de la procédure plspm
#  en mode réflexif
foot_pls=plspm(spainfoot,foot_struct,foot_blocs,
                    modes = foot_modes)
```

Classe de l'objet ***foot_pls*** :

```
# détermination de la classe de l'objet
class(foot_pls)
```

La **structure** et le **résumé** des résultats peuvent être obtenues avec les instructions suivantes :

```
# structure des résultats
foot_modes = c("A", "A", "A")
#  résumé des résultats
summary(foot_pls)
```

Spécification du modèle structurel PLS-PM : **résultats** d'estimation du modèle structurel

Classe de l'objet ***foot_pls*** :

```
> class(foot_pls)
[1] "plspm"
```

Structure des résultats obtenus :

```
Partial Least Squares Path Modeling (PLS-PM)
---------------------------------------------------------------------
## NAME                                  DESCRIPTION
##       1 $outer_model                  outer model
##       2 $inner_model                  inner model
##       3 $path_coefs                   path coefficients matrix
##       4 $scores                       latent variable scores
##       5 $crossloadings                cross-loadings
##       6 $inner_summary  summary inner model
##       7 $effects                      total effects
##       8 $unidim                       unidimensionality
##       9 $gof                          goodness-of-fit
##       10 $boot                        bootstrap results
##       11 $data                        data matrix
---------------------------------------------------------------------
```

Spécification du modèle structurel PLS-PM : **coefficients** estimés du modèle

Coefficients du modèle interne (entre VL) :

> foot_pls$path_coefs

Résultats obtenus :

	Attaque	***Defense***	***Succes***
Attaque	***0.000000***	***0.0000000***	***0***
Defense	***0.000000***	***0.0000000***	***0***
Succes	***0.757261***	***-0.2836068***	***0***

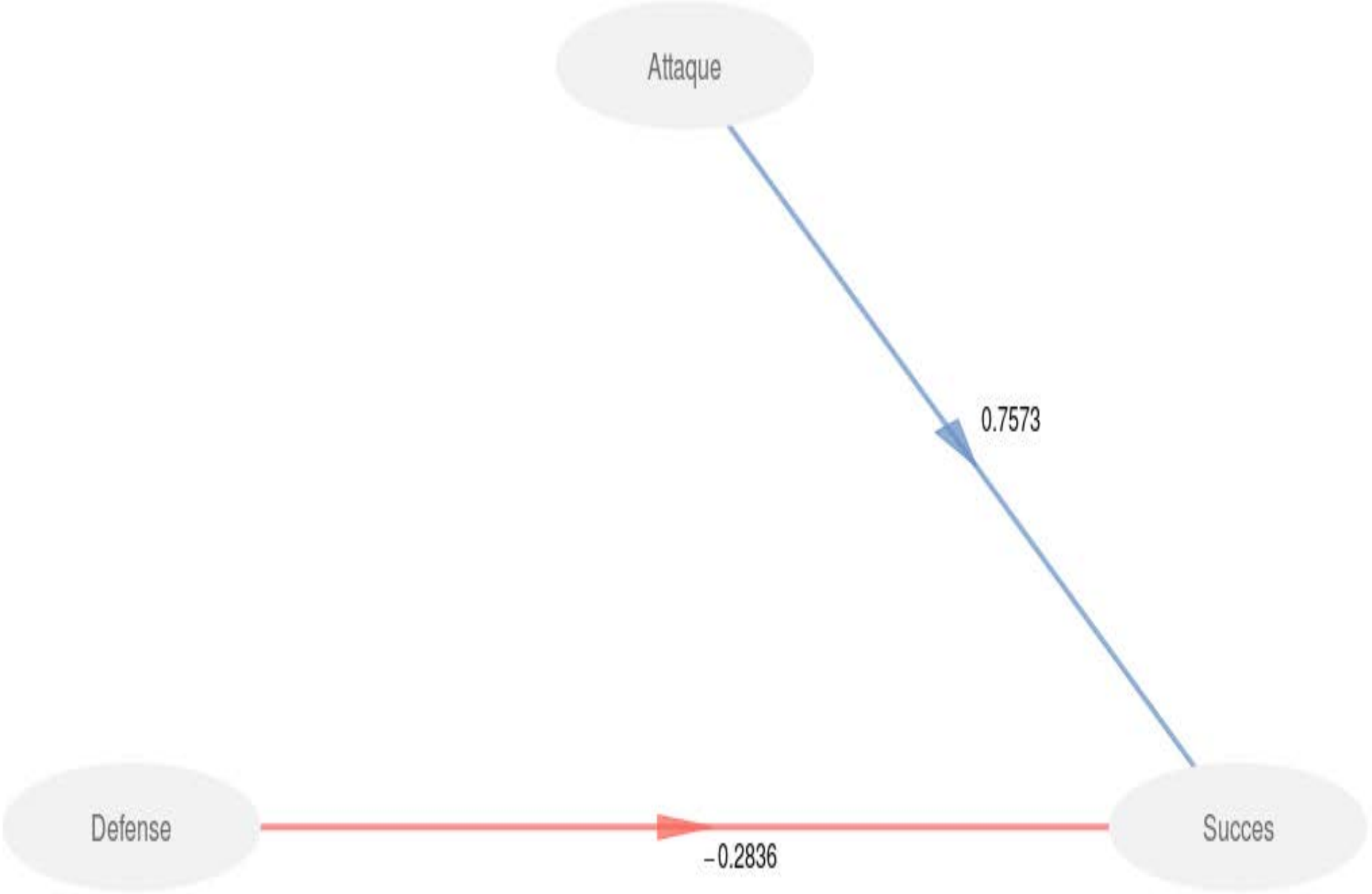
Les estimations des coefficients correspondent au modèle attendu (hypothétique) :

influence positive de la VL Attaque a= 0,757;

influence négative de la VL Défense b= -0,284.

Résultats du modèle PLS-PM : **graphique de structure** du modèle interne

```
> plot(foot_pls)
```



Spécification du modèle structurel PLS-PM : détails de l'estimation du modèle

Modèle interne (entre VL) :

 # modèle interne

foot_pls$inner_model

Résultats obtenus :

$Succes

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	***-2.00e-16***	***0.09***	***-2.17e-15***	***1.00e+00***
Attaque	***7.57e-01***	***0.10***	***7.25e+00***	***1.35e-06***
Defense	***-2.83e-01***	***0.10***	***-2.72e+00***	***1.47e-02***

Les signes des estimations des coefficients correspondent au modèle attendu (hypothétique) :

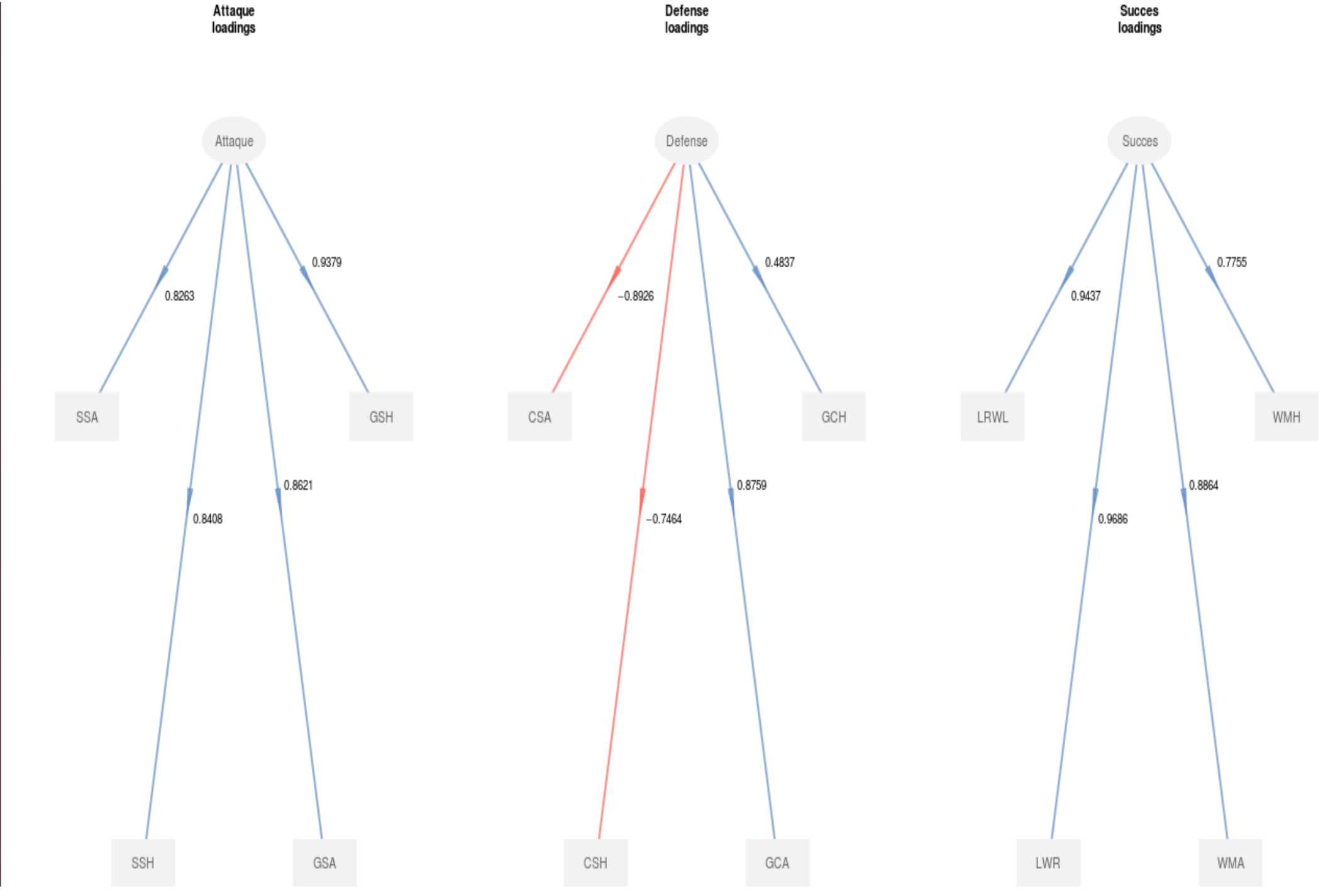
constante nulle

influence positive de la VL Attaque a= 0,757;

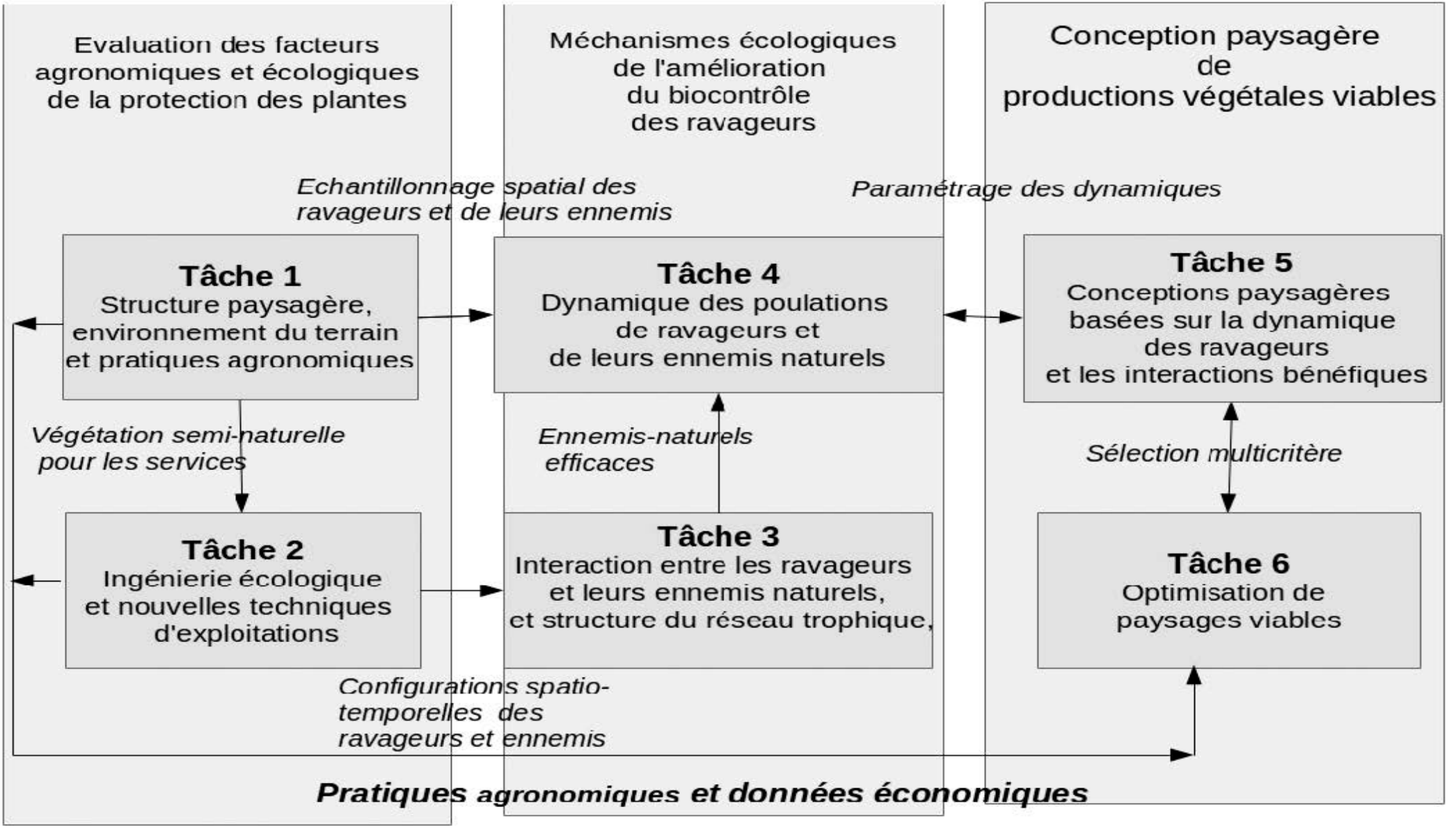
influence négative de la VL Defense b= -0,284.

Résultats du modèle PLS-PM : **graphique de structure** du ***modèle externe***

> plot(foot_pls, what = "loadings", arr.width = 0.1)



Projet Peerless : relations entre **état écologique**, **pratiques agronomiques** et **résultats économiques**



Programme ANR ≪ Agrobiosphere - Viabilité et adaptation des écosystèmes productifs, territoires et ressources face aux changements globaux ≫

Projet Peerless : schéma de description de l'exploitation

Exploitation

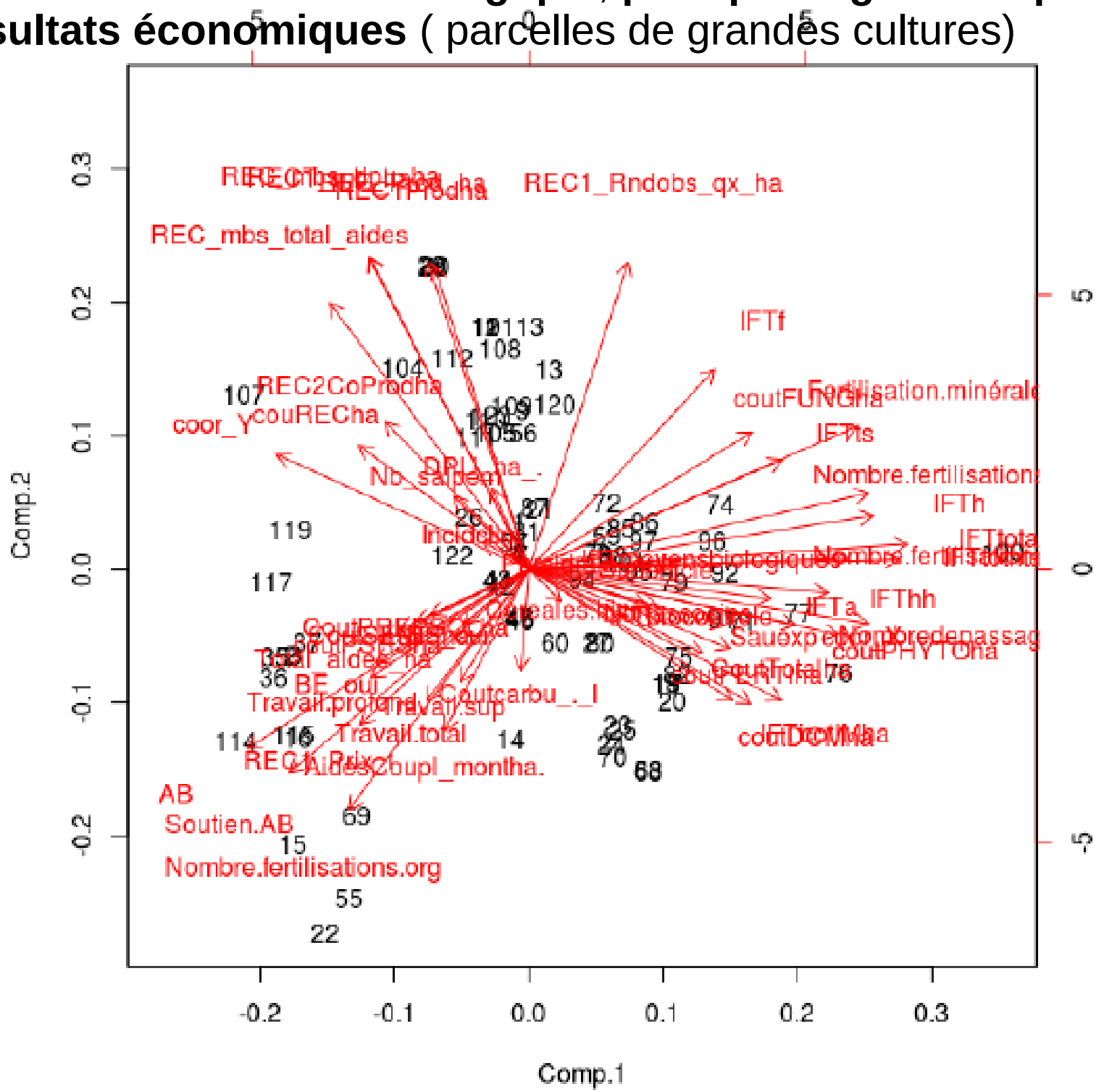
- Identification
- Caractéristiques générales
- Structure de l'assolement
- Structure du cheptel
- Matériel utilisé pour la gestion des bio-agresseurs
- Parc matériel de l'exploitation (SEBIOPAG)

Projet Peerless : schéma de description de la parcelle

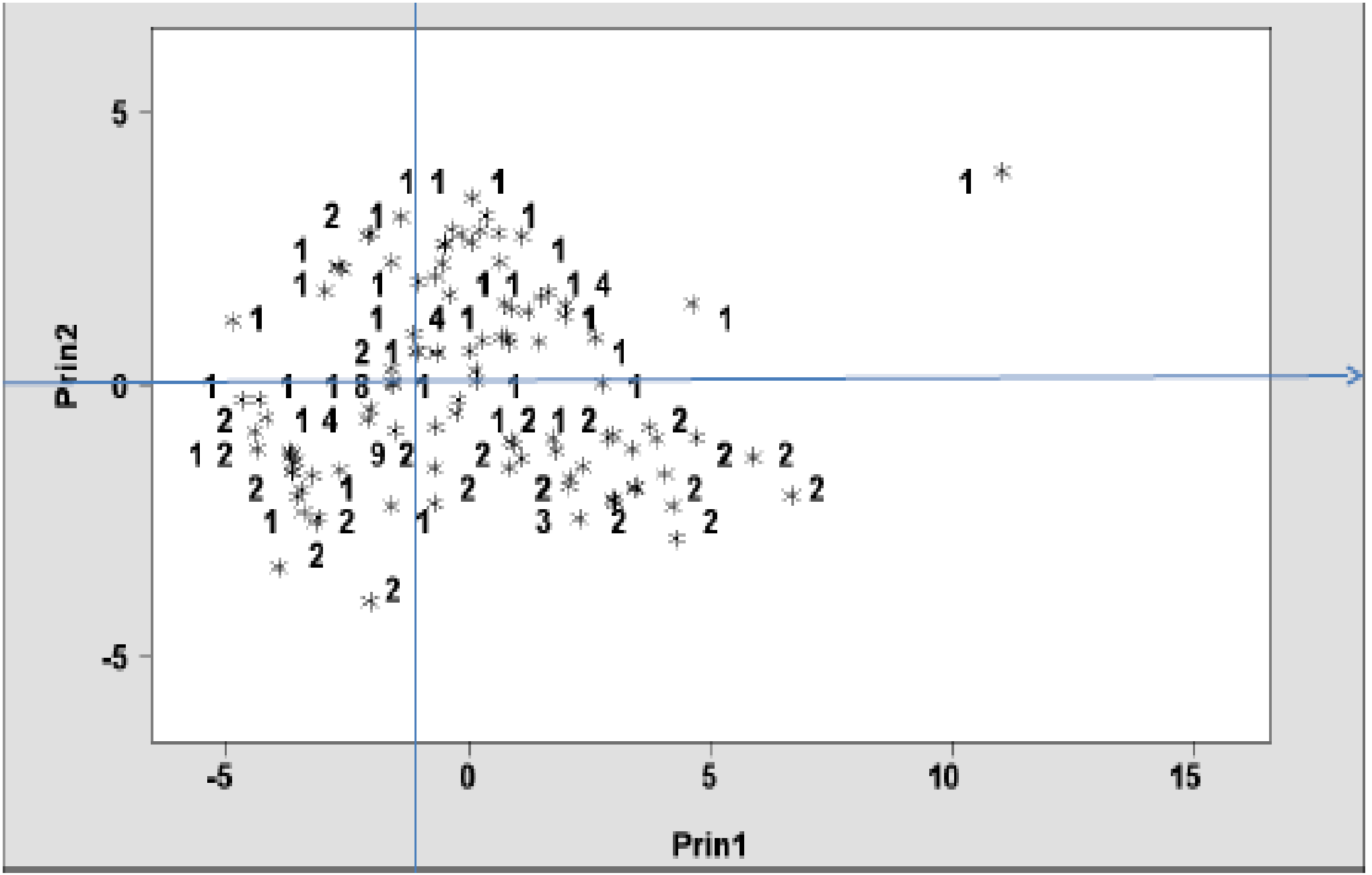
Parcelle

- Identification
- Caractéristiques
- Eléments de bordure
- Préparation du sol et du semis
- Désherbage chimique et mécanique
- Traitements fongicides
- Traitements insecticides et molluscides
- Etapes de la fertilisation
- Etapes de la récolte
- Travaux sous-traités
- Plantation pérenne

Peerless : relations entre **état écologique**, **pratiques agronomiques** et **résultats économiques** (parcelles de grandes cultures)

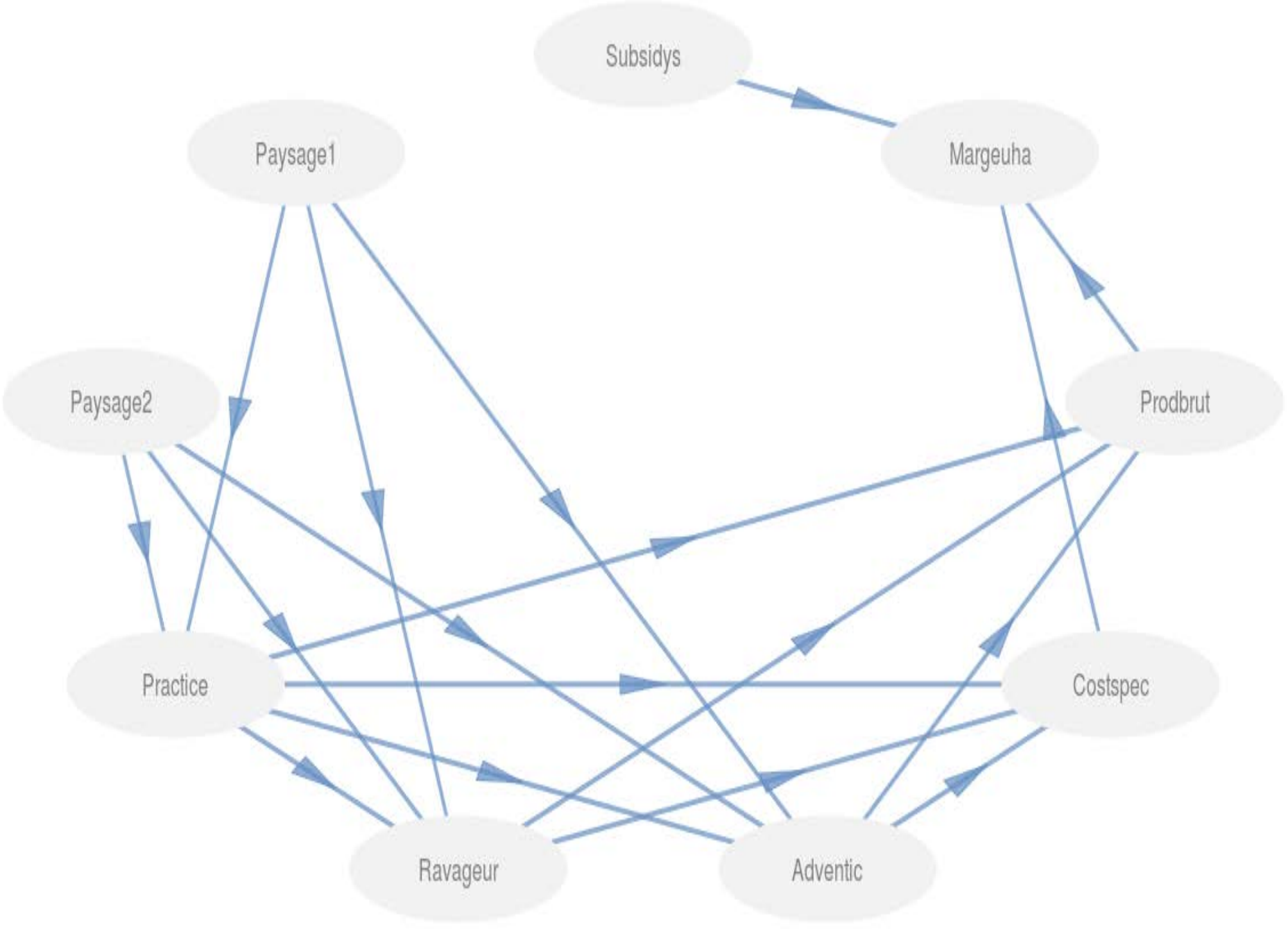


Peerless : relations entre **état écologique**, **pratiques agronomiques** et **résultats économiques** (parcelles de grandes cultures)

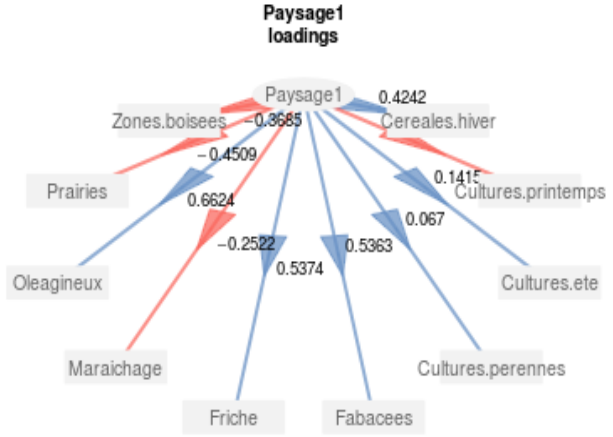


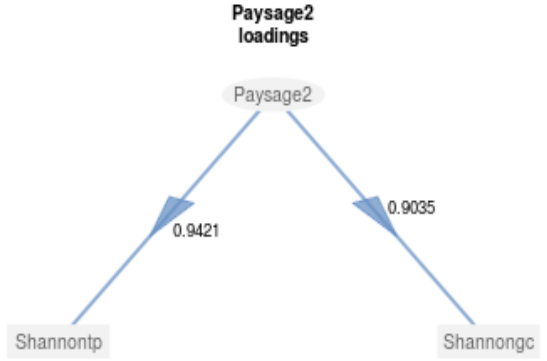
Légende : Type de culture (1=blé, 2=colza, 3=moutarde, 4=orge, 8=triticale,9=céréale+protéagineux)

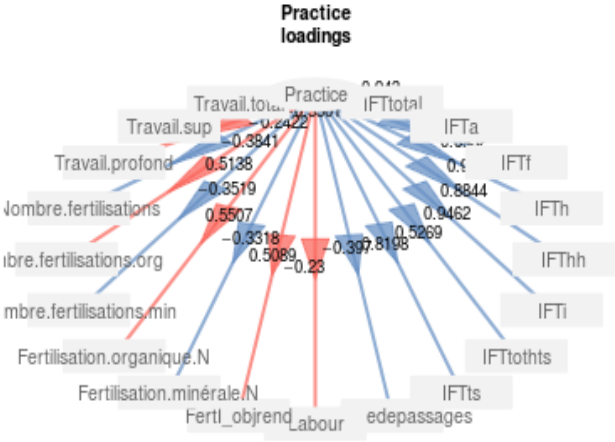
Schéma conceptuel **Peerless** : relations entre **état écologique**, **pratiques agronomiques** et **résultats économiques** (parcelles de grandes cultures)

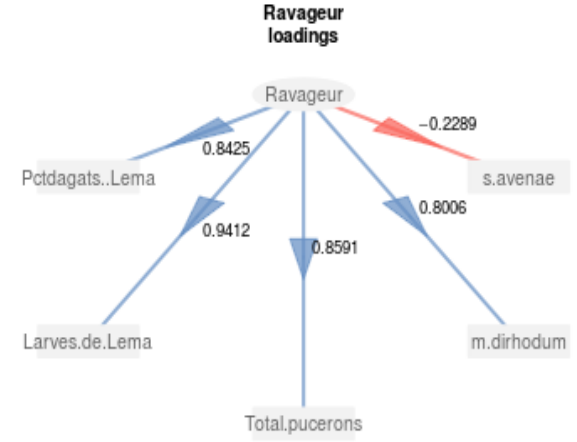


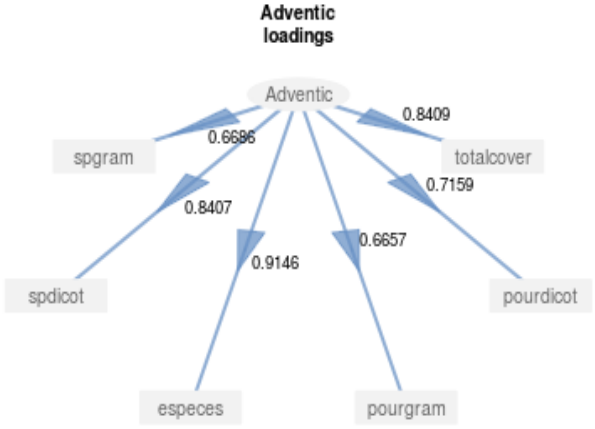
Exemple **Peerless** : relations entre **état écologique**, **pratiques agronomiques** et **résultats économiques** (parcelles de grandes cultures)

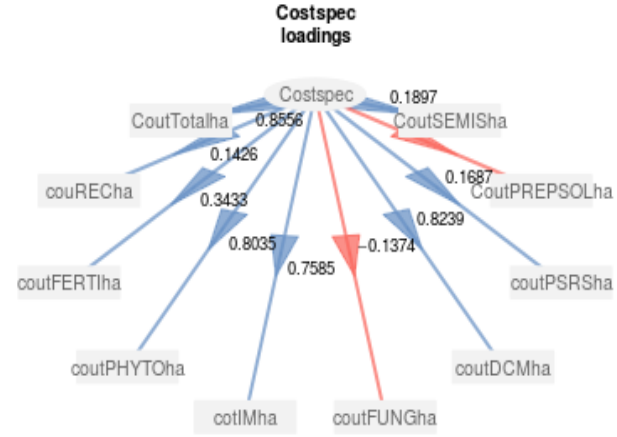


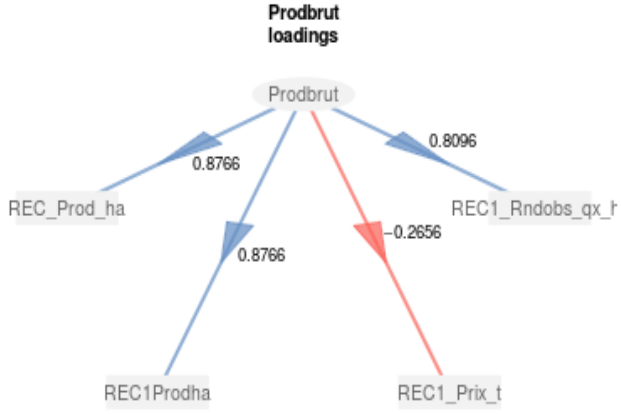


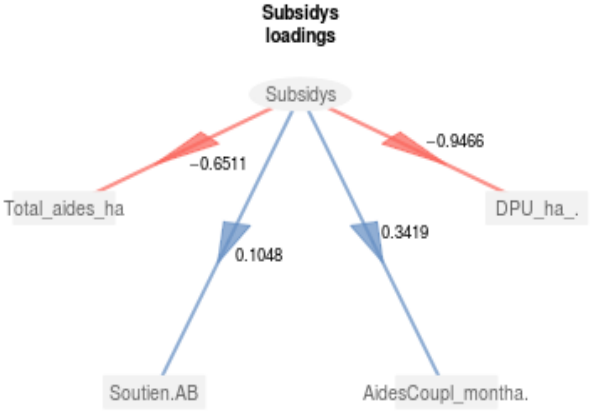


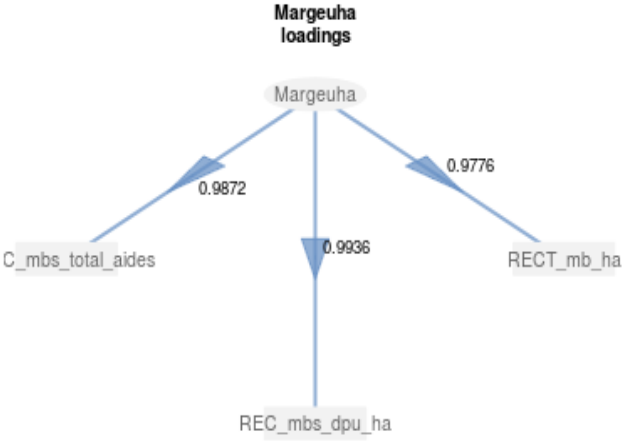












Exemple **Peerless** : relations entre **état écologique**, **pratiques agronomiques** et **résultats économiques** (parcelles de grandes cultures)

Blocs	Mode	Variables observées	Alpha de Cronbach	Rho de Dillon-Goldstein
Paysage1	A	9	0,398	0,073
Paysage2	A	10	0,473	0,642
Practice	A	18	0,943	0,950
Ravageur	A	5	0,807	0,874
Adventic	A	6	0,871	0,904
Costspec	A	10	0,834	0,879
Prodbrut	A	4	0,690	0,805
Subsidys	A	4	0,243	0,125

Les statistiques d'unidimensionnalité (**Alpha de Cronbach** et **Rho de Dillon-Goldstein**) réalisées à partir de ce schéma externe présentent des valeurs trop faibles pour les blocs thématiques :
Subsidys (subventions) ;
Paysage 1 (complexité du parcellaire) ;
Costspec (coûts spécifiques).

Exemple **Peerless** : relations entre **état écologique**, **pratiques agronomiques** et **résultats économiques** (parcelles de grandes cultures)

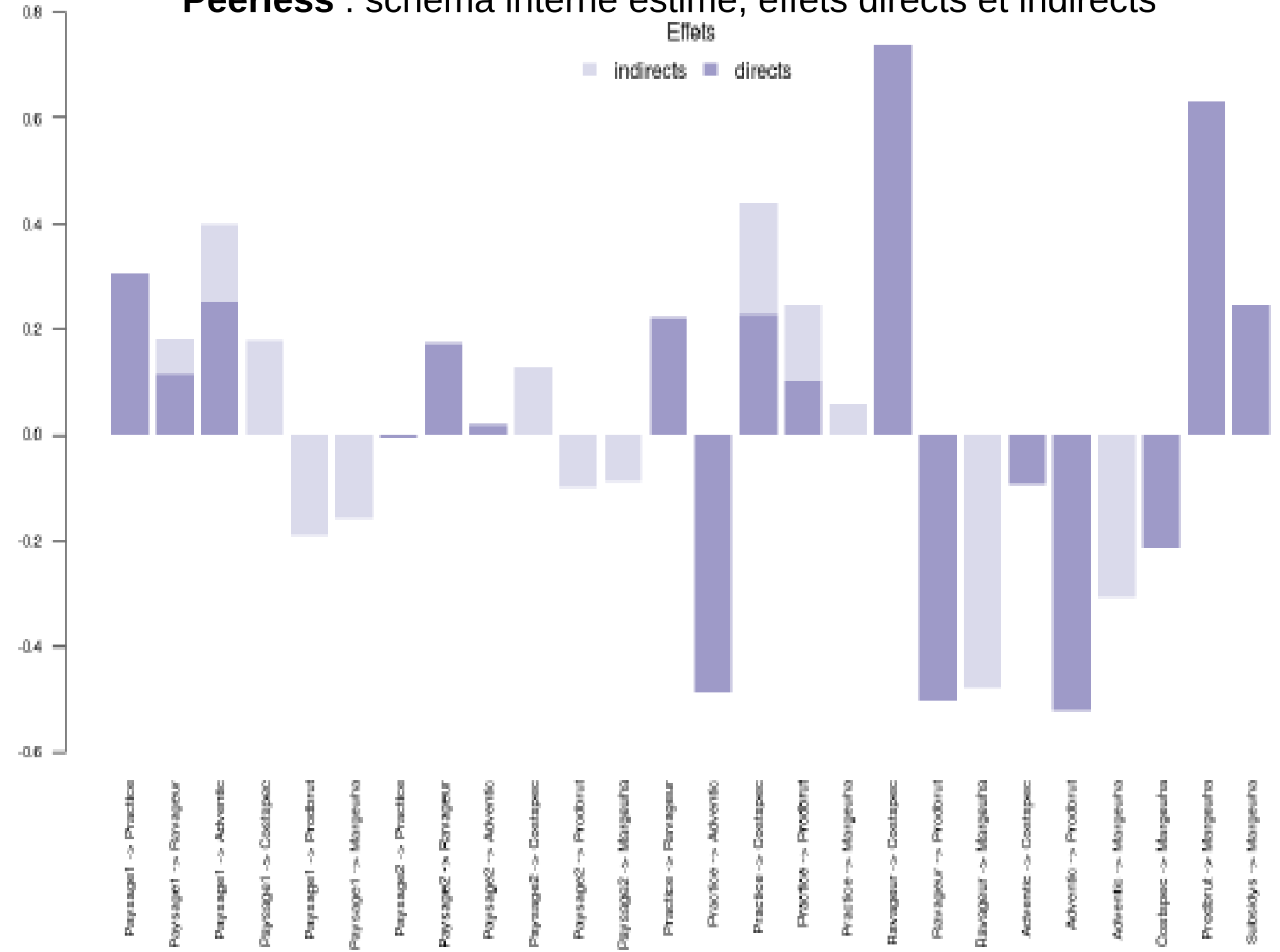
Blocs	Mode	Variables observées	Alpha de Cronbach	Rho de Dillon-Goldstein
Paysage1	A	10	0,000	0,215
Paysage2	A	2	0,829	0,921
Practice	A	19	0,720	0,741
Ravageur	A	5	0,717	0,838
Adventic	A	6	0,871	0,904
Costspec	A	10	0,541	0,611
Prodbrut	A	4	0,592	0,790
Subsidys	A	4	0,664	0,804

Après un recodage approprié des blocs , les statistiques d'unidimensionnalité se sont améliorées :
elles sont toutes supérieures à *0,7* (seuil d'acceptation de l'hypothèse d'unidimensionnalité)
sauf pour

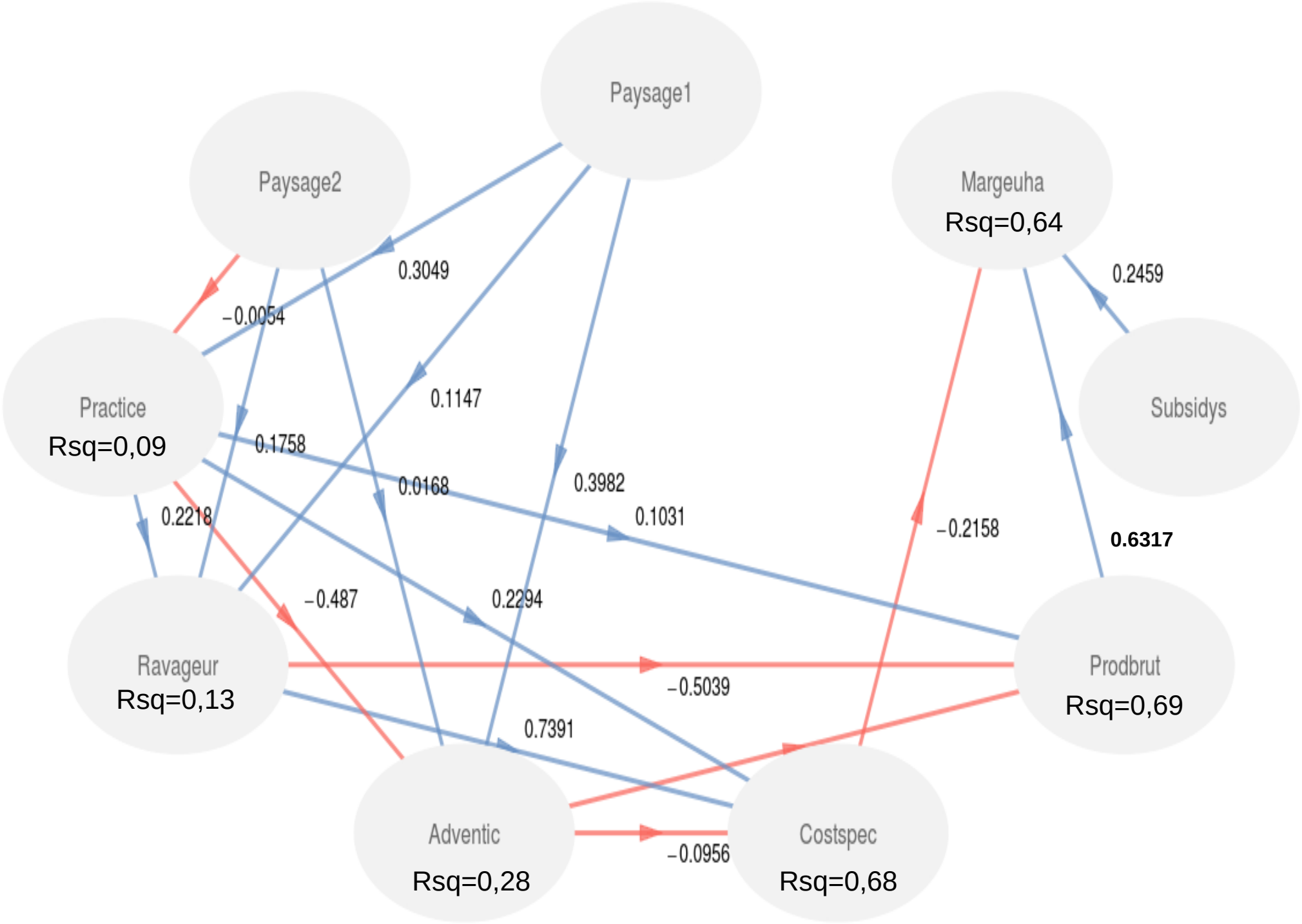
le bloc *Paysage1* et les blocs *Costspec*, *Prodbrut*, *Subsidys* avec des valeurs inférieures à *0,7* pour l'**Alpha de Cronbach**,

cependant corrects pour le **Rho de Dillon-Goldstein** (*>0,6*) sauf pour *Paysage1 (0,215)*.

Peerless : schéma interne estimé, effets directs et indirects



Peerless : modèle interne amendé, effets totaux (GoF=0,4565)



PLS-PM : lexique

Lexique de l'approche PLS-PM

- *loadings* : corrélation entre variable mesurée (VM) et variable latente (VM) ;
- *MV* : variable manifeste ou mesurée (tableau des données observées) ;
- *LV* : variable latente (facteur d'une analyse factorielle, composante principale d'une analyse en composantes principales, pseudo-composantes PLS) ;
- *Formative* : mode de relation entre la VM et la VL ;
- *Réflectif* : mode de relation entre la VM et la VL ;
- *A* : mode réflexif ;
- *B* : mode formatif ;

PLS-PM : références bibliographiques

Sur l'approche PLS-PM :

Sanchez, G. (2013) *PLS Path Modeling with R.*Trowchez Editions. Berkeley, 2013.
Libre accès à: http://www.gastonsanchez.com/PLS_Path_Modeling_with_R.pdf

Sanchez, G. Trinchera, L., Russolillo G. (2017). *plspm: tools for partial least squares path modeling (PLSPM).* Logiciel R, version 0.4.9.

Sanchez, G. Trinchera, L., Russolillo G. (2017) *Introduction to the R package plspm,* 10 p.

Tenenhaus M., Vinzi V.E., Chatelin Y.-M., Lauro C. (2005). PLS path modeling. *Computational Statistics & Data Analysis* n°48, pp. 159–205.

Sur la régression PLS et ses différentes applications :

Tenenhaus M. (1998). *La régression PLS : théorie et pratique*, éditions Technip, 264 p.

Vinzi, E.V., Chin, W.W., Henseler, J., Wang, H. (2010) *Handbook of Partial Least Squares : Concepts, Methods and Applications*, Springer, 798 p.

Sur l'application aux données agronomiques et environnementales :

Mezerette F. (2016) *Quels effets de la gestion agricole et du paysage sur l'abondance de bioagresseurs et le rendement ?*, Mémoire de Master , Université Rennes 1, 33 p.

Quinio M. & *alii* (2017) Separating the confounding effects of farming practices on weeds and winter wheat production using path modelling, *Europ. J. Agronomy,* n°82, pp.134-143 .