

The logo for ARVALIS features a stylized graphic on the left composed of three overlapping shapes in yellow, teal, and blue. To the right of this graphic, the word "ARVALIS" is written in a bold, teal, sans-serif typeface. A thin, teal horizontal line is positioned directly beneath the text.

ARVALIS



Intérêt des DAG pour l'analyse causale des données observationnelles

Application aux foreurs du maïs

François PIRAUX

Doriane HAMERNIG

Jean-Baptiste THIBORD

Titre de l'intervention



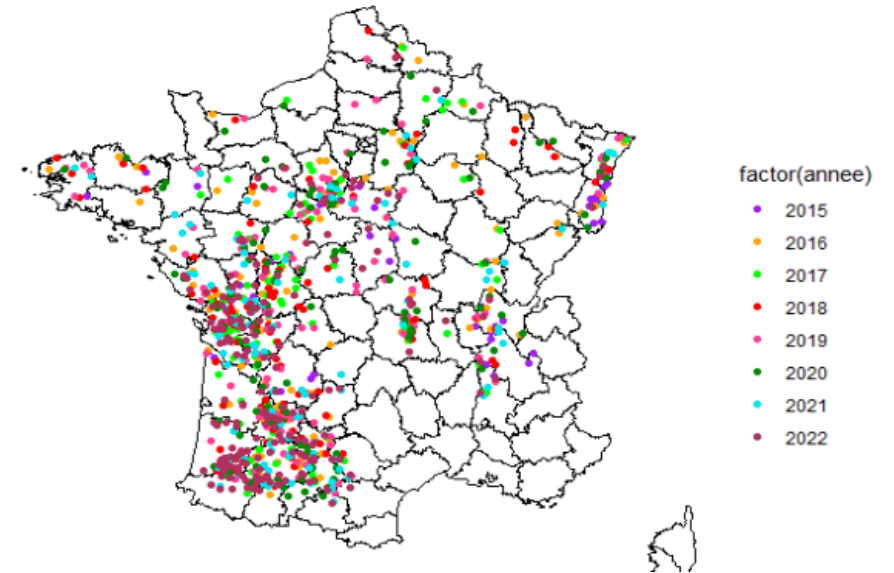
Problématique agronomique

- Foreurs du maïs
- Quels sont les facteurs causaux de l'abondance des foreurs ?
 - Facteurs agronomiques : lesquels ? Avec quel impact ?
 - Facteurs climatiques : lesquels ? Avec quel impact ?



Bdd abondance foreur

- Données d'observation
 - 2015-2022
 - 1875 parcelles enquêtées
- Variables
 - Pratiques culturales : 109 variables
 - Sol – paysage : type de sol, pente, haie, ...
 - Climat année n et n-1 : Tmoy, Tmin, Tmax & Pluie journalières
 - Variable à prédire : abondance de foreurs à l'automne (oct)



Localisation des parcelles enquêtées



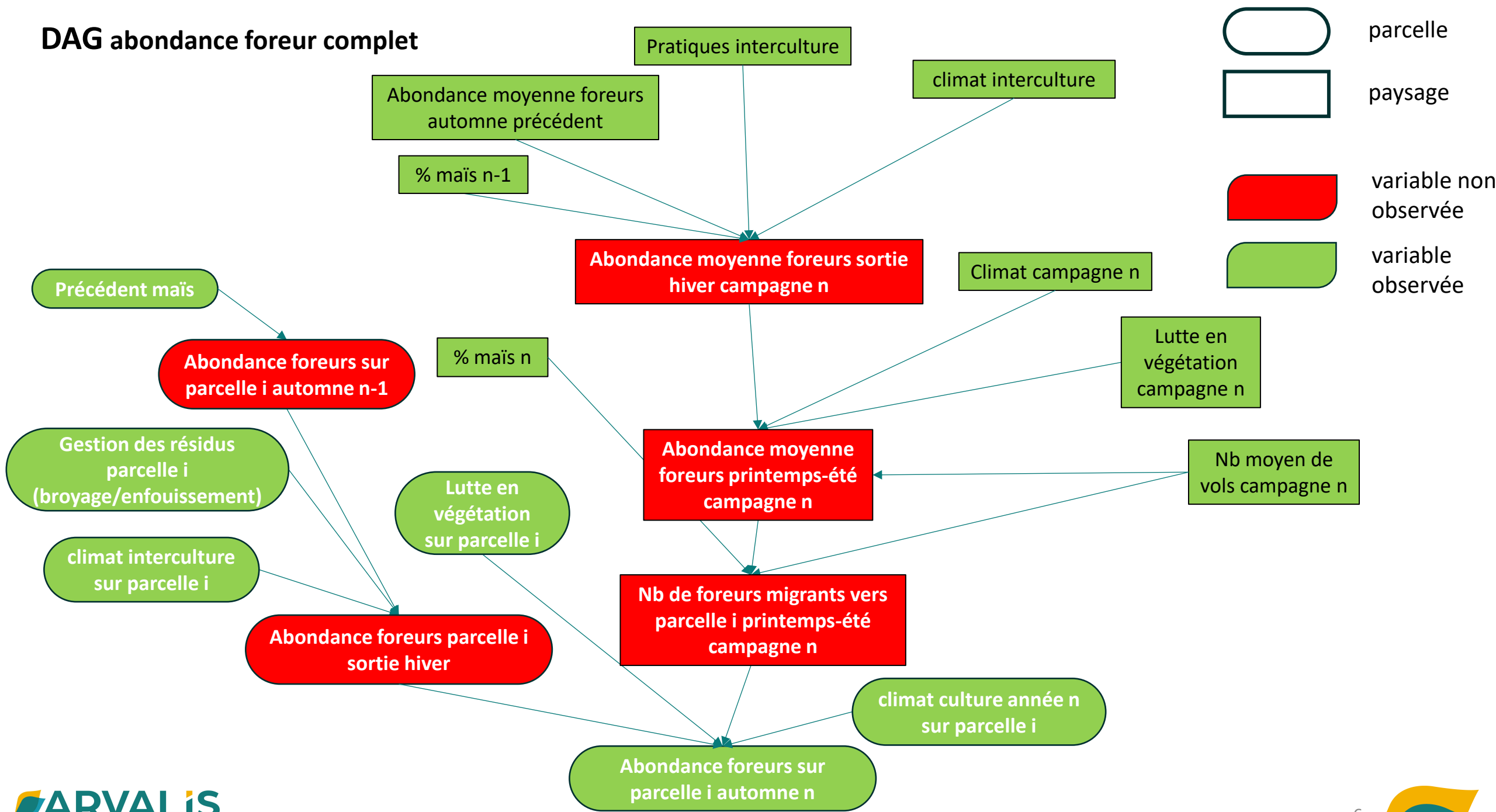
Méthodologie

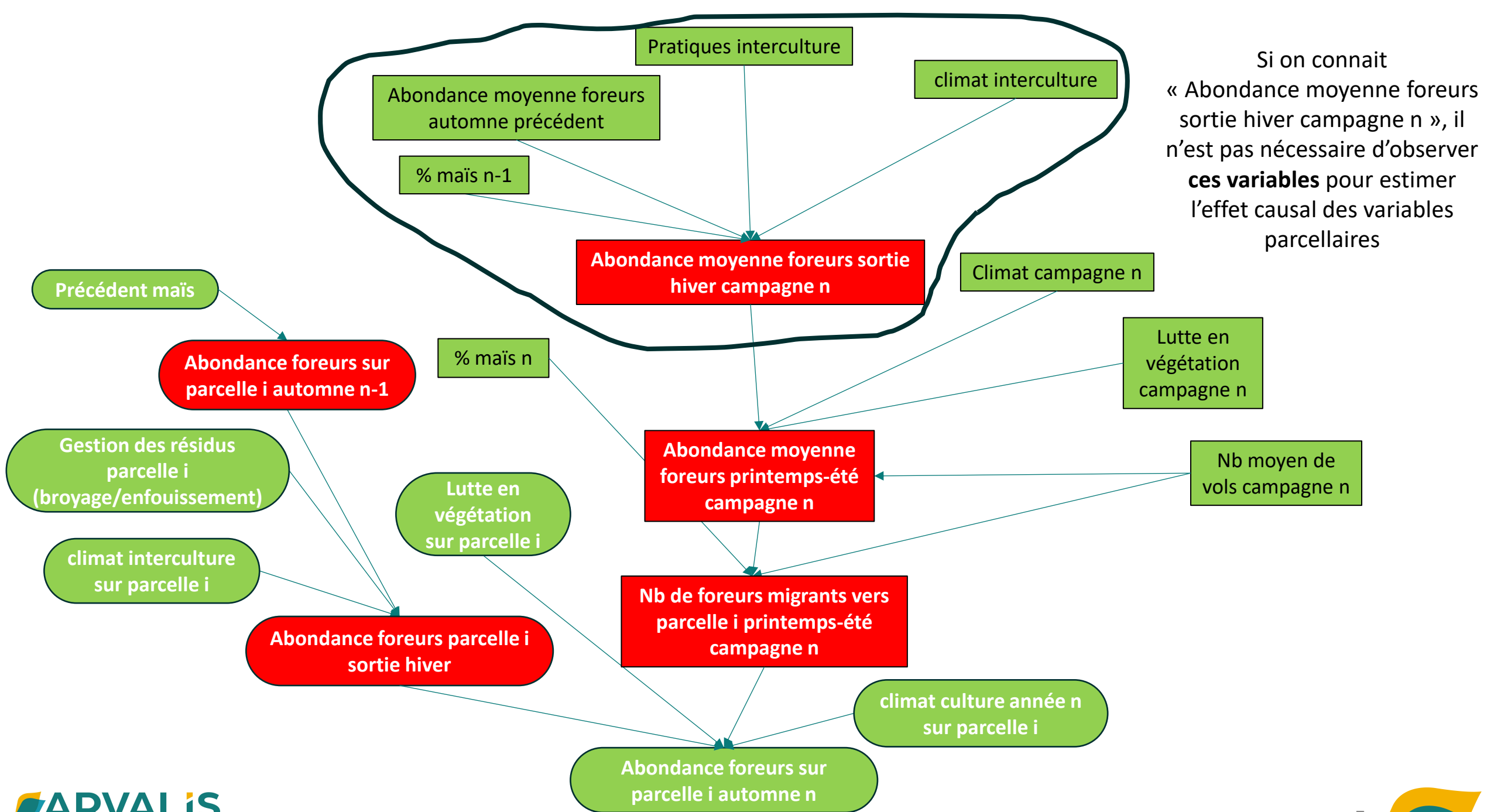
- Causalité à partir de données d'observation, méthodes d'analyse envisagées :
 - Standardisation (G – computation)
 - Score de propension
 - Double robustesse
- La standardisation
 - requiert un modèle prédisant la variable d'intérêt en fonction de la variable explicative dont on veut estimer l'effet causal (A), et des facteurs de confusion (L) :

$$Y = g(A, L)$$

- DAG
 - On pourrait être tenté d'inclure dans les analyses un maximum de variables de contrôle (L) afin d'éviter tout risque de confusion d'effets. Cette pratique est cependant à proscrire*.
 - Les DAGs sont particulièrement utiles pour distinguer les associations causales des associations non causales, et pour identifier les variables qui doivent être mesurées et contrôlées afin d'obtenir des estimations sans confusions d'effets.

DAG abondance foreur complet



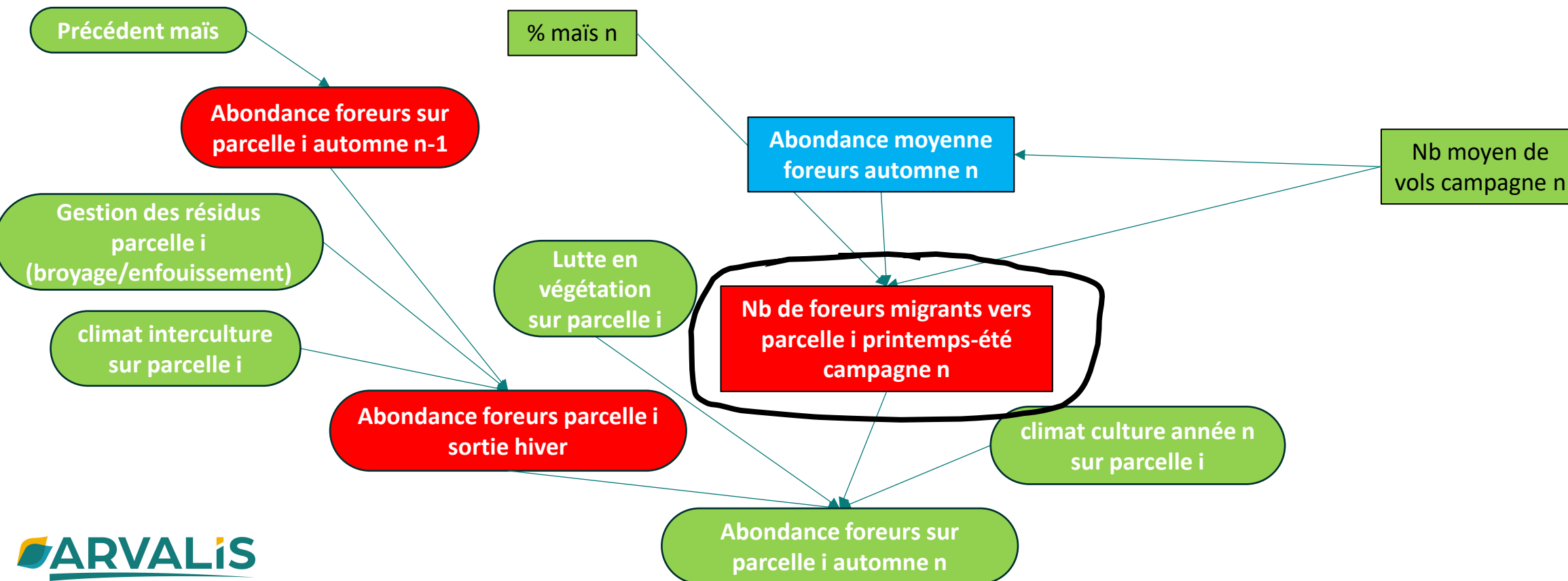


AG abondance foreur complet

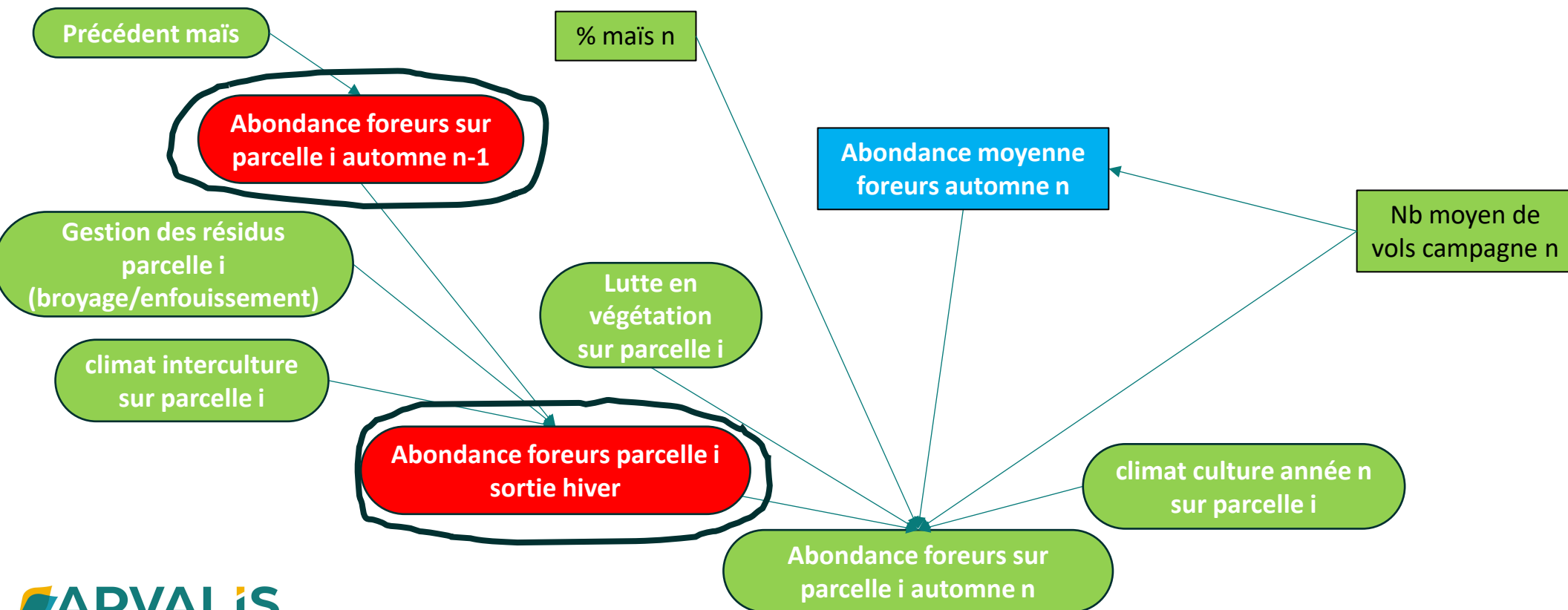
« Nb de foreurs migrants vers parcelle i printemps-été campagne n » n'est pas observée.

→ On considère un lien direct de « % maïs n », « Abondance moyenne foreurs automne n » et

« Nb moyen de vols campagne n » vers « Abondance foreurs sur parcelle i automne n »



« **Abondance foreurs parcelle i sortie hiver** » et « **Abondance foreurs sur parcelle i automne n-1** » ne sont pas observées
→ On considère un lien direct de « **Précédent maïs** », « **Gestion des résidus parcelle i** » et « **climat culture année n sur parcelle i** » vers « **Abondance foreurs sur parcelle i automne n** »



Variables retenues pour la G – computation

« % maïs n » est estimé par segment à partir des données RPG

% maïs n

« Abondance moyenne foreurs automne n » correspond à une moyenne d'abondance par segment

Abondance moyenne foreurs automne n

« Nb moyen de vols campagne n » est estimé par une somme de degrés-jours

Nb moyen de vols campagne n

Précédent maïs

Gestion des résidus
parcelle i
(broyage/enfouissement)

climat interculture
sur parcelle i

Lutte en
végétation
sur parcelle i

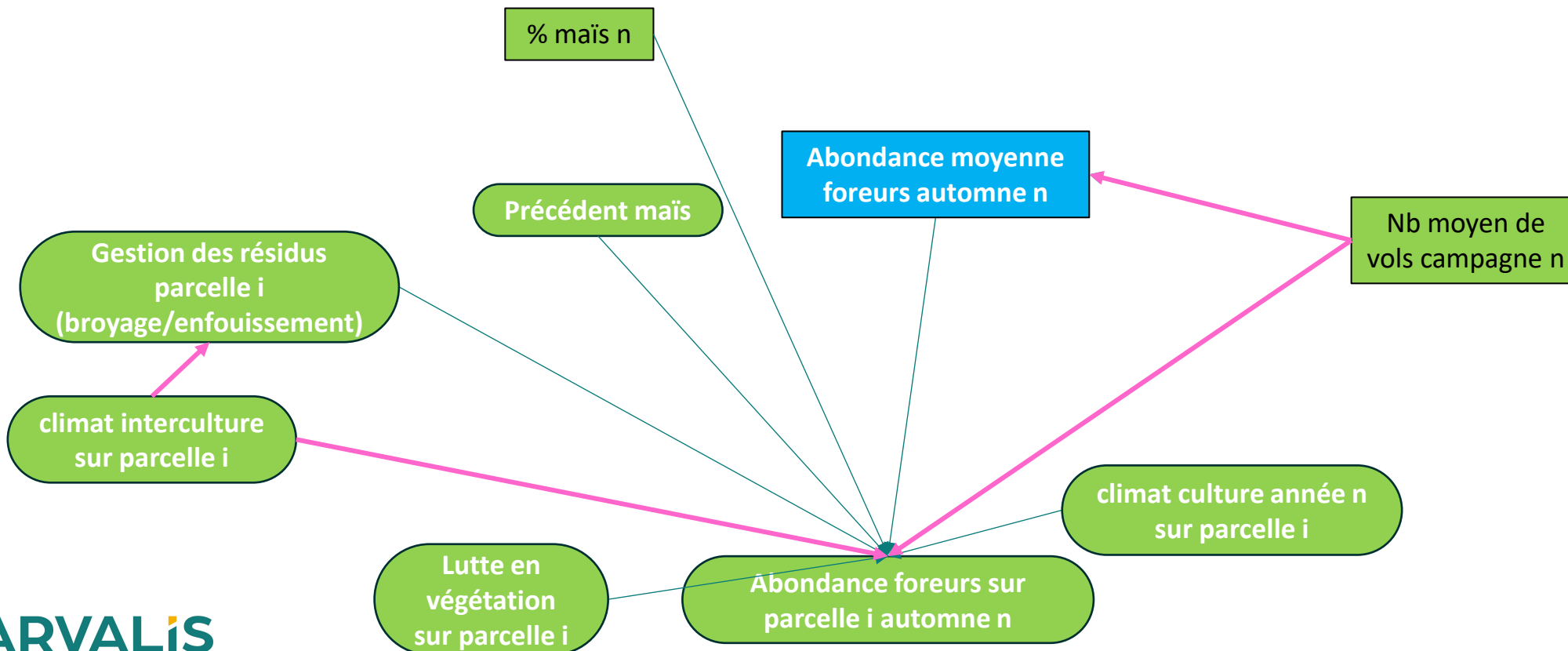
climat culture année n
sur parcelle i

Abondance foreurs sur
parcelle i automne n

Identification de confusions d'effets potentielles

- Le DAG permet d'identifier des confondeurs: variables qui agissent à la fois sur l'exposition et le risque:

Ex: « Nb moyen de vols campagne n », « climat interculture parcelle i »

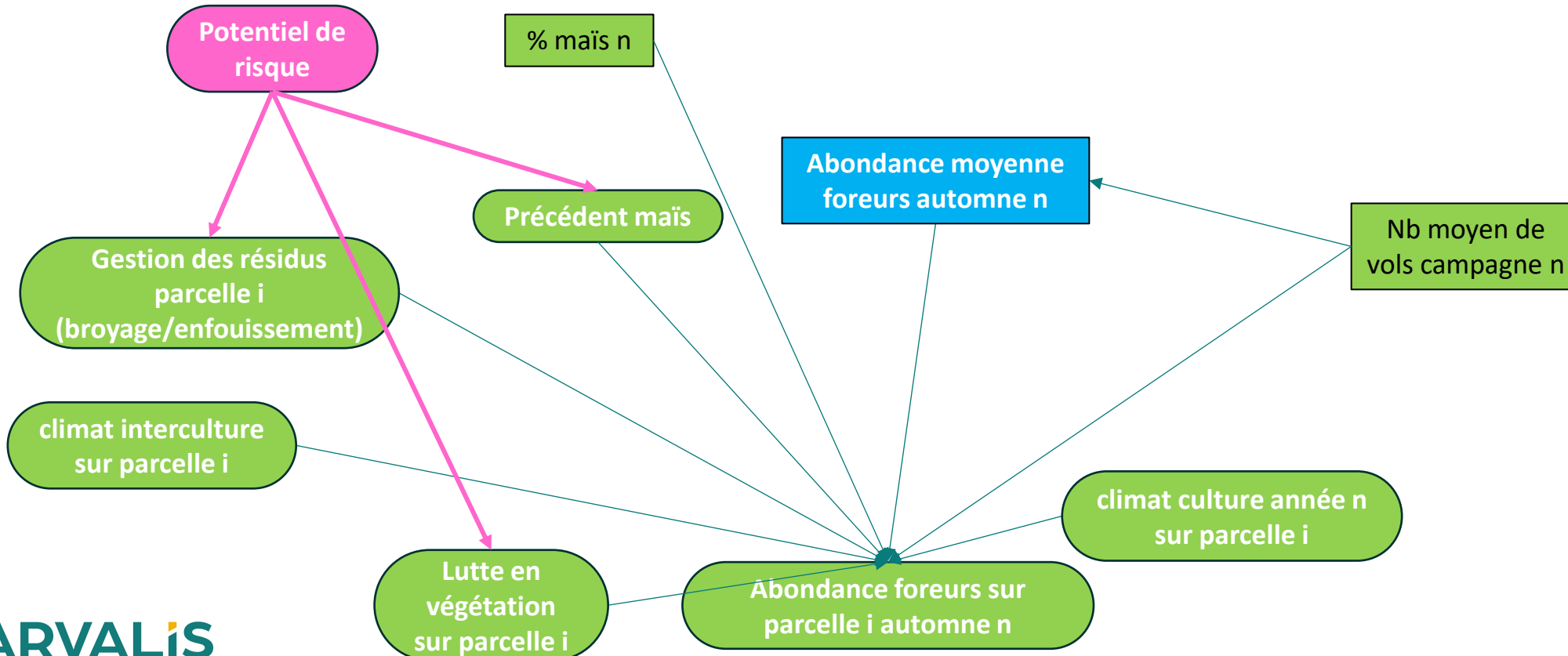


Identification de confusions d'effets potentielles

- On travaille sur des données observationnelles et il existe donc sans doute des corrélations entre les variables agros étudiées qui peuvent être dues à :
 - Une cause commune : ex: **potentiel de risque**
 - Un hasard d'échantillonnage

Exemple:

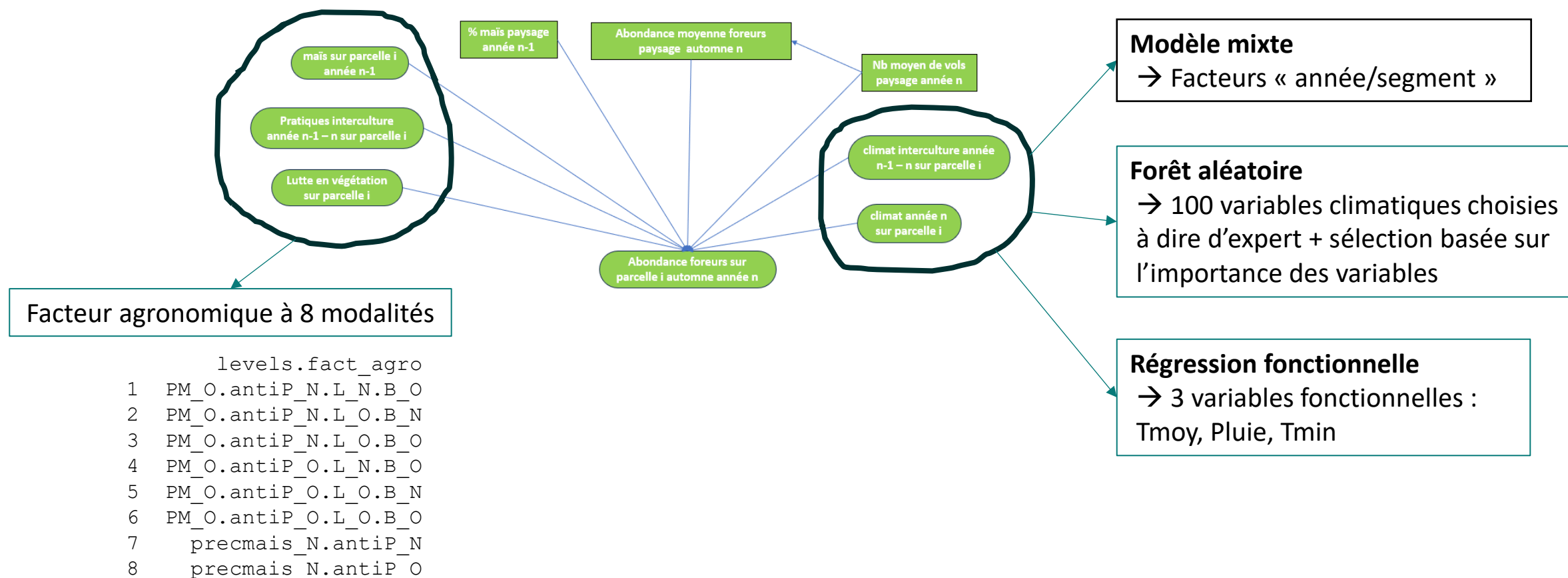
=> Dans le jeu de données, le labour est plus fréquent quand il y a un précédent maïs que lorsqu'il n'y en a pas



DAG foreur

Quelques résultats : standardisation

• Modèles testés



DAG foreur

Quelques résultats : standardisation

- Effet antipyrrale

```
      levels.fact_agro
1 PM_O.antiP_N.L_N.B_O
2 PM_O.antiP_N.L_O.B_N
3 PM_O.antiP_N.L_O.B_O
4 PM_O.antiP_O.L_N.B_O
5 PM_O.antiP_O.L_O.B_N
6 PM_O.antiP_O.L_O.B_O
7   precmais_N.antiP_N → Précédent non maïs, pas d'antipyrrale
8   precmais_N.antiP_O → Précédent non maïs, antipyrrale
```



DAG foreur

Quelques résultats : standardisation

- Prédiction du modèle sur une grille de référence = jdd
- Exemple Modèle Mixte

```
lmer(log(abondance) ~ value_moy_sgmt + pct_mais + nb_vols + fact_agro + (1|ANNEE/CODE_SEGMG))
```

jdd copie 1		value_moy_sgmt	pct_mais	nb_vols	fact_agro	ANNEE	CODE_SEGMG	pred_factagro7
	1	0.24740741	0.060	1.296618	precmais_N.antiP_N	2011	NCBG	-1.662375
	2	0.24740741	0.060	1.230580	precmais_N.antiP_N	2011	NCBG	-1.737150
	3	0.24740741	0.060	1.240371	precmais_N.antiP_N	2011	NCBG	-1.726064
	4	0.24740741	0.060	1.206752	precmais_N.antiP_N	2011	NCBG	-1.764131
	6	0.08444444	0.089	1.453923	precmais_N.antiP_N	2011	COLH	-2.194597
	7	0.08444444	0.089	1.453923	precmais_N.antiP_N	2011	COLH	-2.194597
	...							
jdd copie 2		value_moy_sgmt	pct_mais	nb_vols	fact_agro	ANNEE	CODE_SEGMG	pred_factagro8
	1	0.24740741	0.060	1.296618	precmais_N.antiP_O	2011	NCBG	-3.051943
	2	0.24740741	0.060	1.230580	precmais_N.antiP_O	2011	NCBG	-3.126719
	3	0.24740741	0.060	1.240371	precmais_N.antiP_O	2011	NCBG	-3.115632
	4	0.24740741	0.060	1.206752	precmais_N.antiP_O	2011	NCBG	-3.153699
	6	0.08444444	0.089	1.453923	precmais_N.antiP_O	2011	COLH	-3.584165
	7	0.08444444	0.089	1.453923	precmais_N.antiP_O	2011	COLH	-3.584165
	...							

$\text{mean}(\text{pred_factagro8} - \text{pred_factagro7})$



DAAG foreur

Quelques résultats : standardisation

- Résultats

Modèle	Estimation effet antipyrals	IC bootstrap 95%
Modèle mixte	-1 . 39 (4 . 0)	(-2 . 01 , -0 . 77)
Forêt aléatoire	-0 . 89 (2 . 4)	(-1 . 34 , -0 . 55)
Régression fonctionnelle	-1 . 50 (4 . 5)	(-2 . 03 , -0 . 95)



Intérêt d'avoir fait un DAG

- L'approche « DAG » permet rendre plus transparents les choix de modélisation
- Elle facilite également le dialogue entre spécialiste thématique et data scientist
- Permet d'identifier les variables causales et les confondeurs à inclure dans le modèle



Limites

- On n'est jamais certains d'avoir identifié tous les confondeurs (modification du DAG à chaque fois qu'on le reprend)
- Résultats de la standardisation différents en fonction du modèle
- Prise en compte du climat quand on a peu d'expertise sur ces variables

