

MODÉLISATION DU SCOPE 3

BARRAUD Lorenzo

BRIAND Léo

SOMAVO Gloria

VAUTIER Samuel

square[®]
management

adWay



MASTER
ESA
Université d'Orléans

Crédit Agricole renonce au financement de nouveaux projets d'énergies fossiles

Les Echos, le 14/12/23

La Banque Postale s'engage à ne plus financer les énergies fossiles d'ici à 2030 

Les Echos, le 14/10/21

Sommaire

Contexte

Exploration des données

Modélisation

Que retenir ?

Contexte



Les banques font face à un nouvel ensemble de **risques** qui s'ajoutent aux défis économiques, réglementaires et de réputation. L'Union européenne a mis en place un plan d'action pour une **finance durable**, et les législations nationales contribuent à une "avalanche réglementaire" imposant aux banques de se conformer à diverses directives.

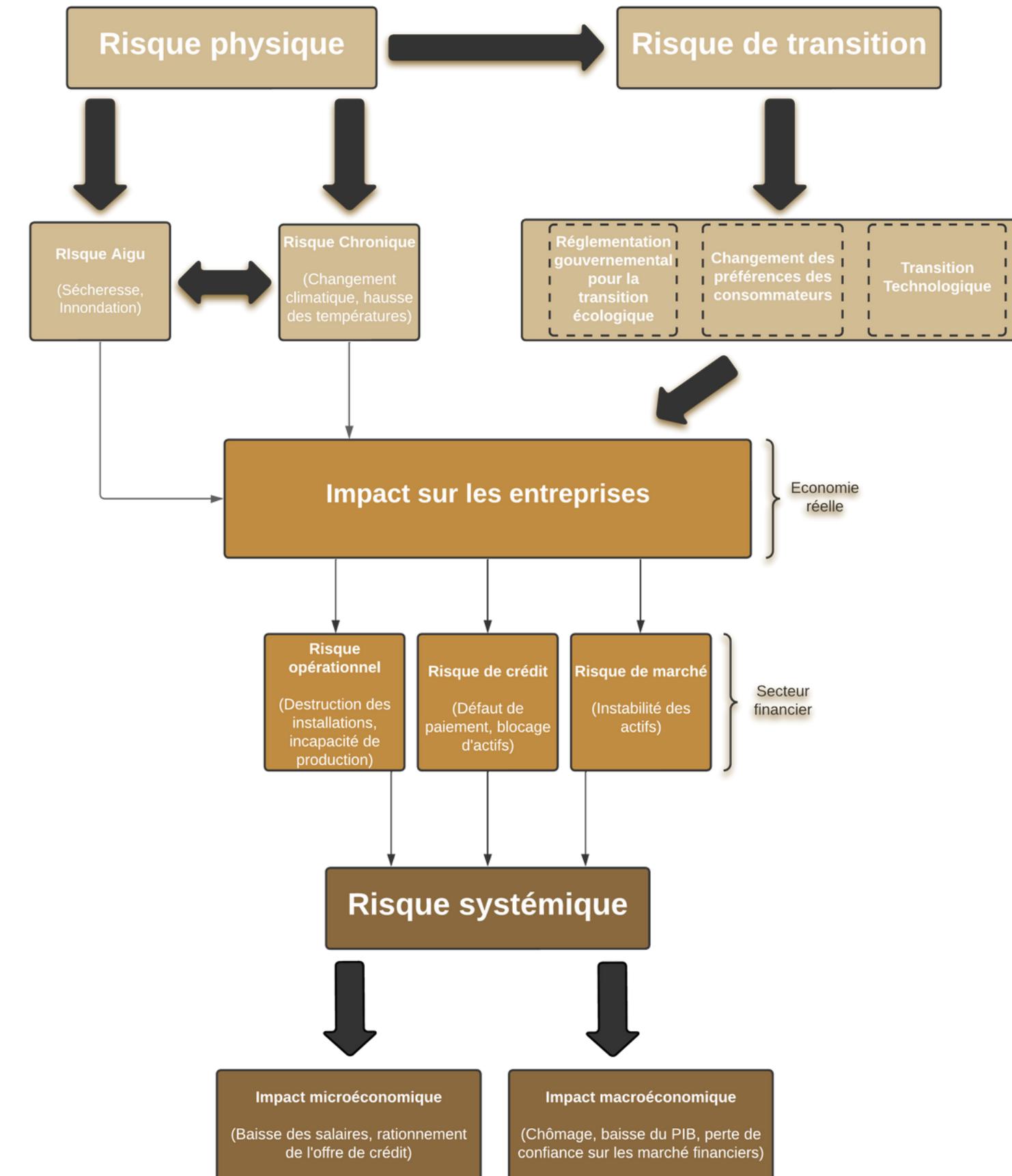


En 2020, la BCE a publié un rapport intitulé "**Guide relatif aux risques liés au climat et à l'environnement**". Dans ce dernier, l'institution a souligné que, nous citons : "Lorsqu'ils déterminent leurs objectifs stratégiques, les établissements devraient notamment tenir compte des risques que présente la transition vers une économie plus durable, à **faible intensité carbone**, pour leurs portefeuilles de prêts et de négociation."

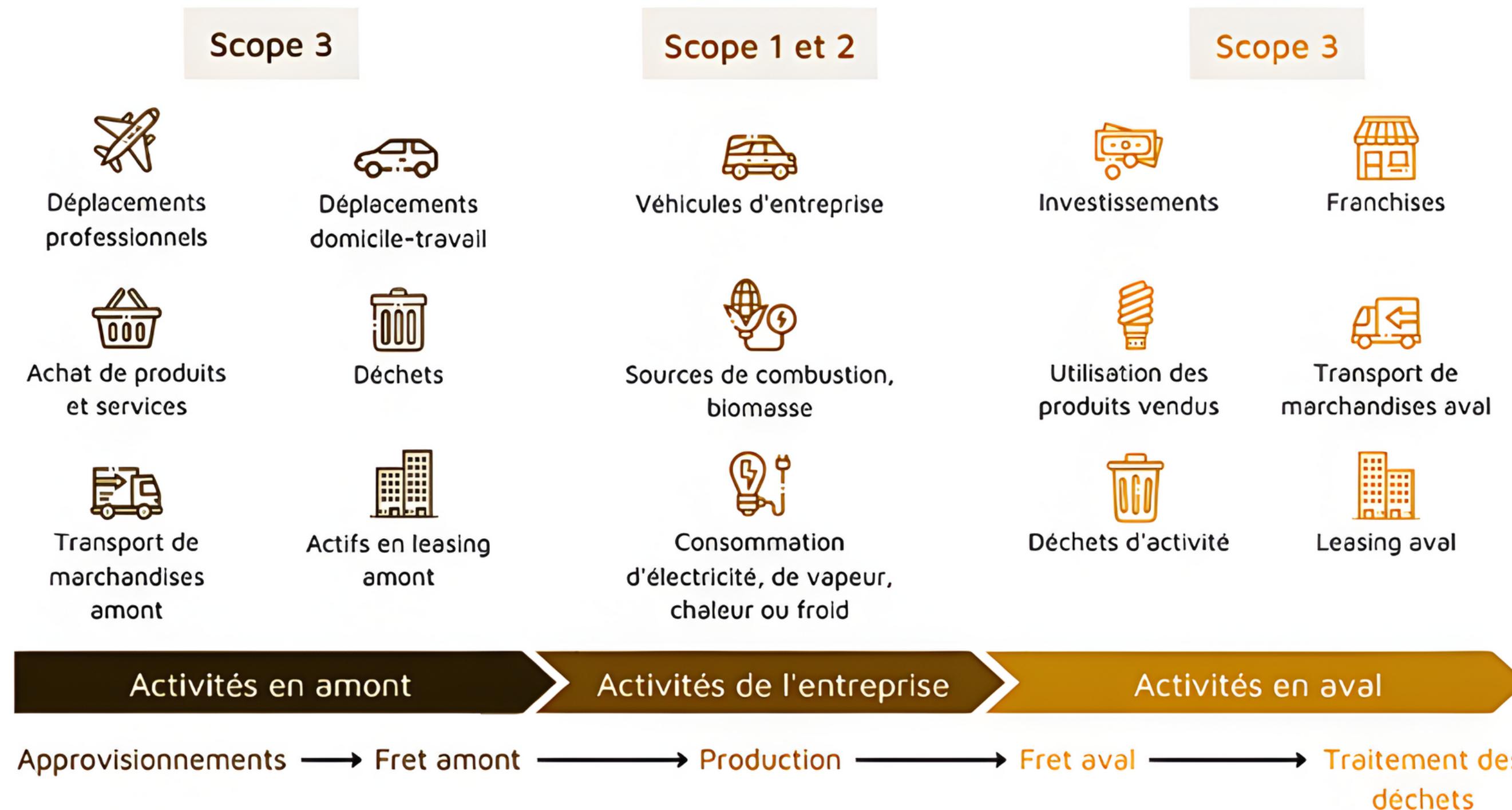


Il est nécessaire pour les établissements financiers de connaître l'empreinte carbone de leur portefeuille afin d'apprécier leur **risque de transition**.

Cartographie des risques



Les scopes





Objectif 1

Estimer le Scope 3 dans
un contexte **Big Data**



Objectif 2

Proposer de nouvelles
méthodes d'estimation du scope 3



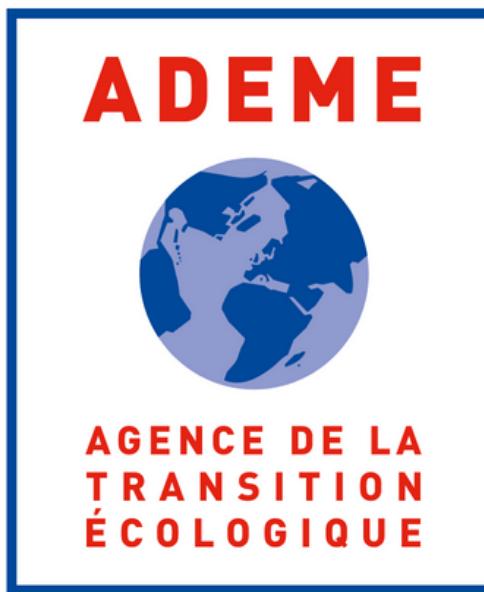
Objectif 3

Poser **les limites** et donner
des pistes d'ouvertures à ces
méthodes

A photograph of an industrial landscape at sunset. In the foreground, two dark silhouettes of smokestacks stand against a sky filled with large, billowing plumes of smoke. The smoke is illuminated from behind by the setting sun, appearing in shades of orange, yellow, and red. The sky is a mix of dark blues and blacks, with some lighter, wispy clouds visible. The overall atmosphere is one of pollution and industry.

EXPLORATION DES DONNÉES

Présentation des données



Base ADEME - Bilan GES

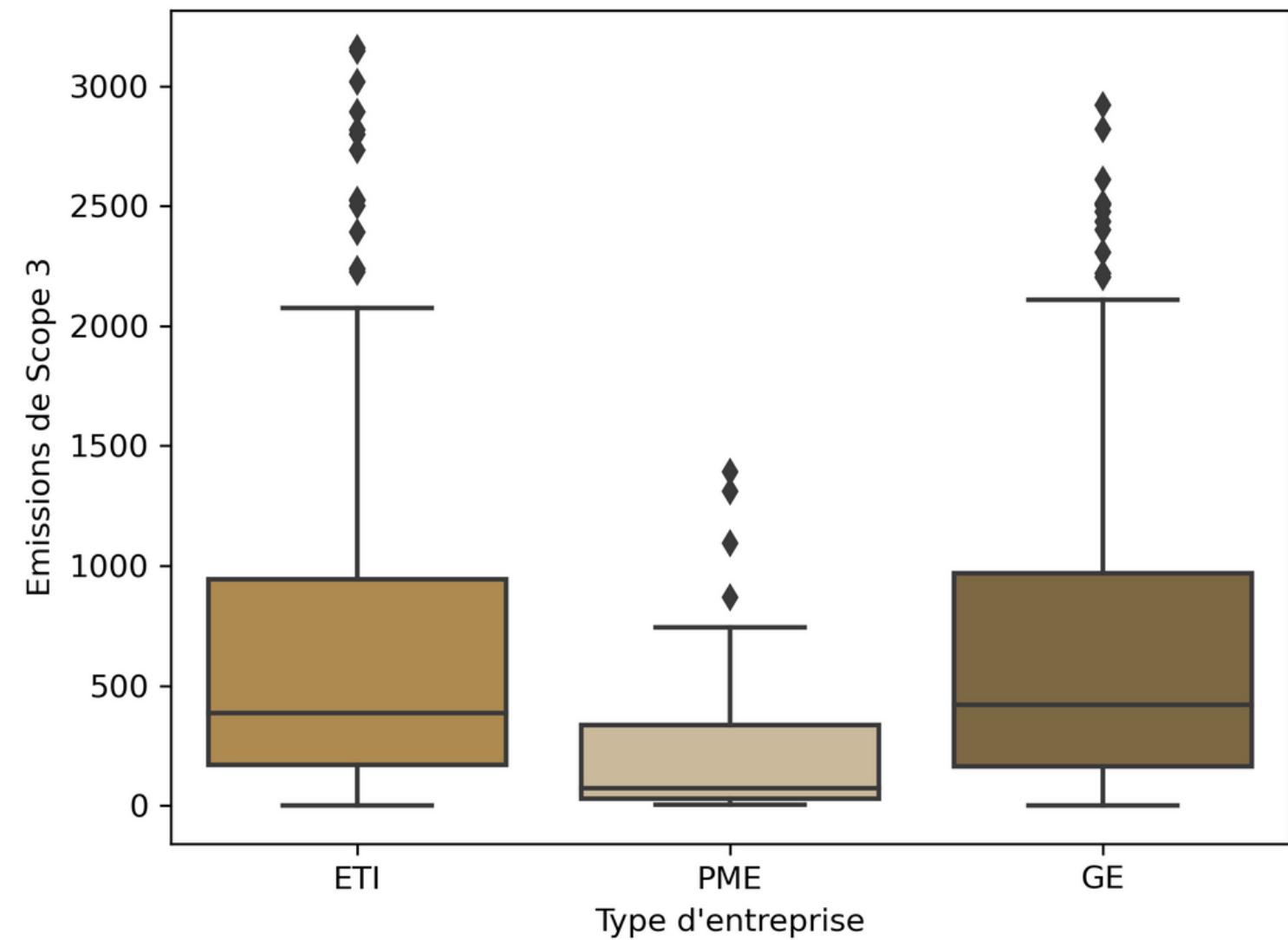


Base Sirene



Base DIANE

La variable Scope 3



Aux vues du faible nombre de PME comparativement aux GE et aux ETI, il n'est pas pertinent d'interpréter les scope 3 pour ce groupe.

Il ne semble pas y avoir de forte différence statistique entre le scope 3 pour les ETI et les GE.

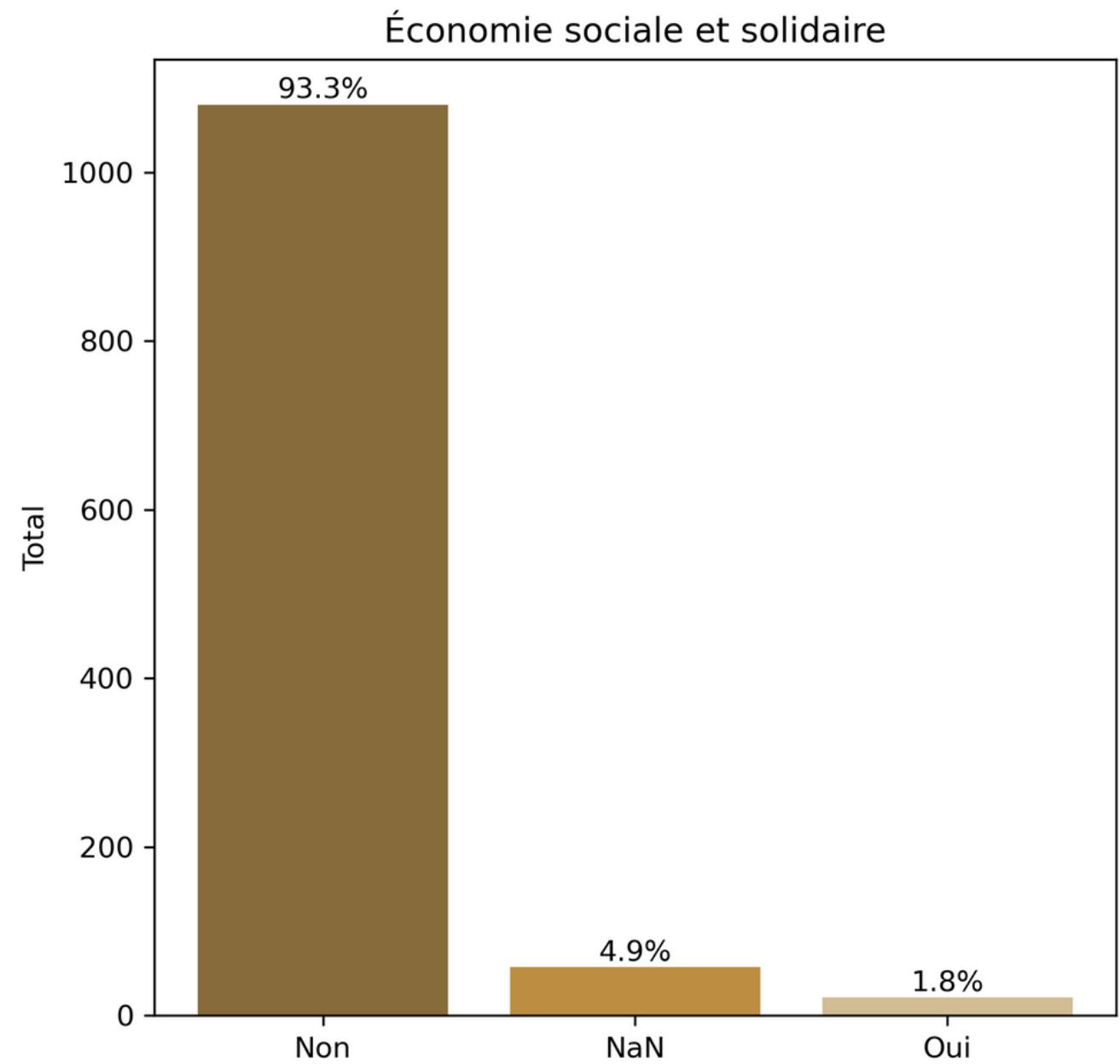
Valeurs manquantes

Seules deux variables comportent des données manquantes :

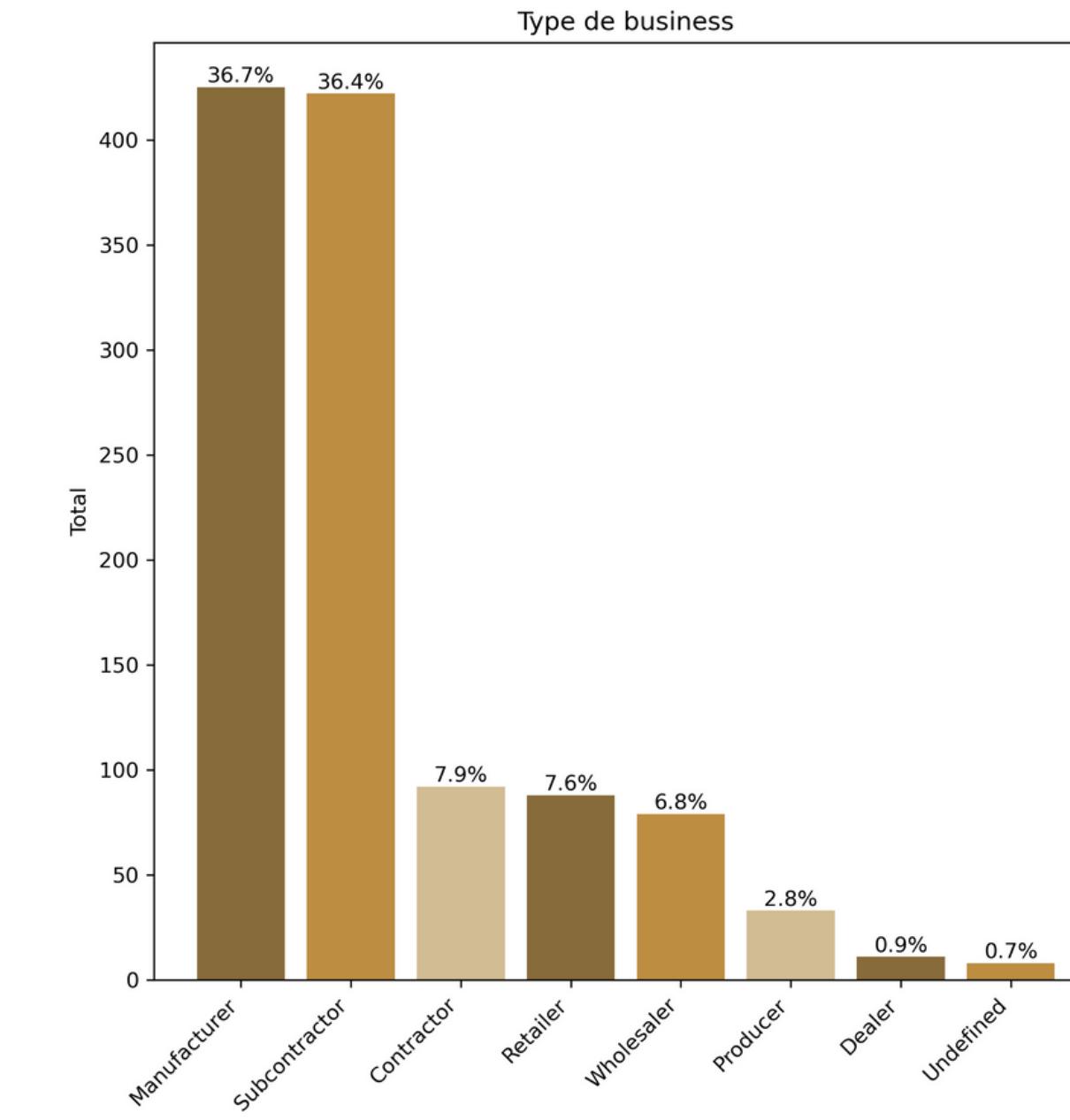
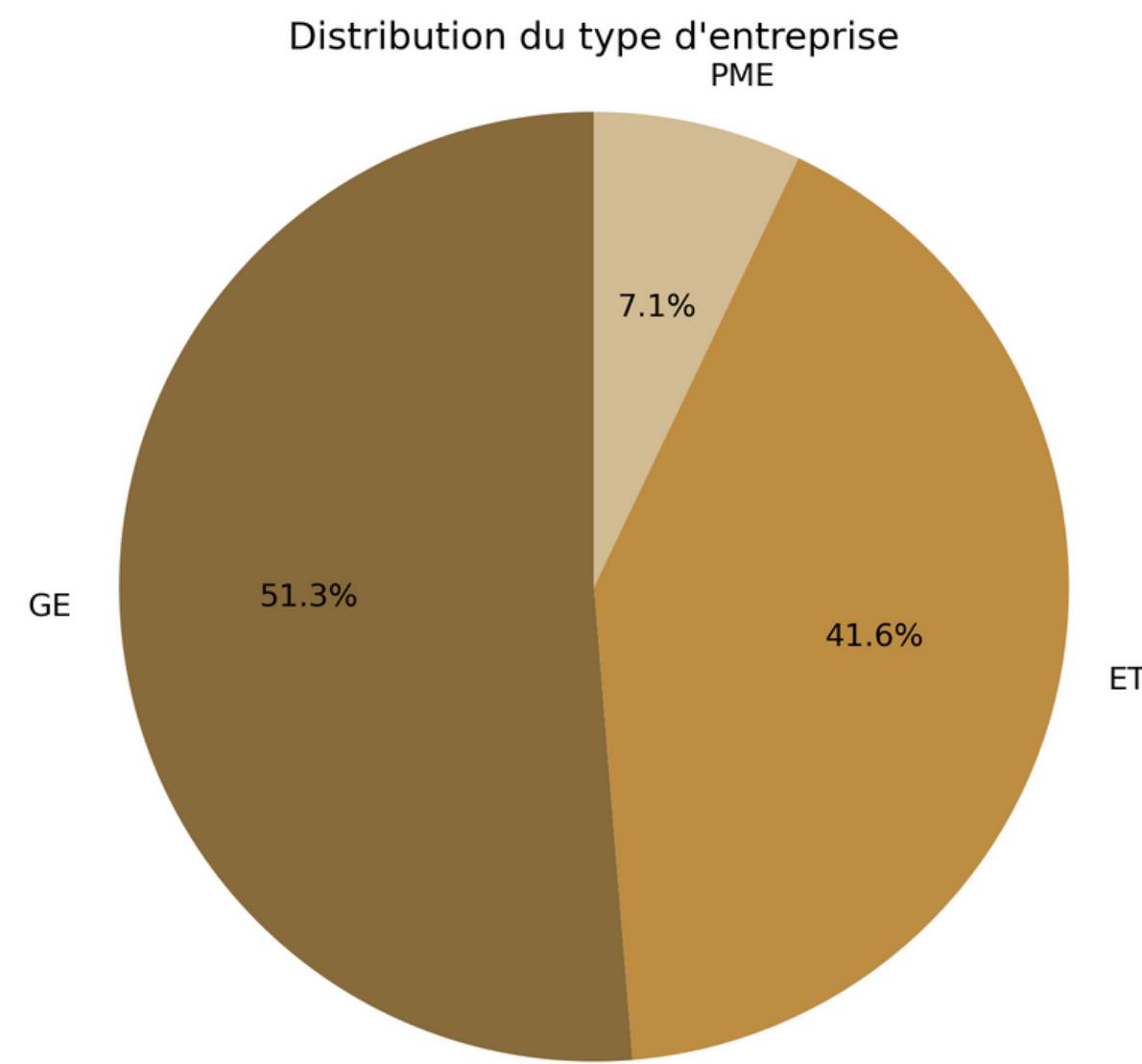
- **economie sociale et solidaire** : 5%
- **total_scope_3** : 60%

Nous avons décidé de ne pas procéder à l'imputation des valeurs manquantes pour le **Scope 3** et de **conserver uniquement les observations avec des valeurs**.

Cette décision entraîne une perte significative d'observations, réduisant la base initiale d'environ 1158 observations à **479 observations**.



Les prédicteurs



- Suppression des doublons : **1158 codes SIREN uniques**
- **129 variables** propres à l'entreprise : 120 numériques et 9 catégorielles

Nous retrouvons des variables comme le chiffre d'affaires, le **nombre d'employés**, le nombre d'actionnaires ainsi que les **émissions de Scope 1, 2 et 3**.

MODÉLISATION



CONTEXTE BIG DATA

Sélection des variables

Liste des variables sélectionnées		
Nombre de salariés	SCOPE 1 de l'entreprise (tonnes de CO ₂ e)	Bénéfice d'exploitation
Créances nettes	Revenus de dividendes	Actifs financiers : Ventes
Dettes financières	Revenues financiers	Dettes intra-groupes : Total brut
Subventions d'investissement	Trésorerie nette	Immobilisations nettes
Actions nettes	Flux de trésorerie d'exploitation	Autres immobilisations incorporelles : Montant au début de la période
Provisions	Achats (marchandises), matières premières et autres	Achats de marchandises
Achats de matières premières	Salaires et avantages : Total brut	Ventes de produits manufacturés
Capital social	Immobilisations corporelles : Montant au début de l'exercice	Total : Total brut
Total des provisions pour risques et charges : Montant au début de l'exercice	Transferts de charges	TVA
Salaires et traitements	Fonds de roulement	

Avant tout, nous avons procédé à un échantillonnage stratifié sur la catégorie d'entreprise :

- Apprentissage : 75%
- Validation : 25%
- Test : PME

Sélection par **Elastic-net** afin de conserver les variables les plus importantes tout en gérant les corrélations entre prédicteurs.

Obtention d'un **modèle plus parcimonieux** tout en maintenant une **performance prédictive élevée**.

Random Forest : présentation

Mise en place de deux modèles :

- Modèle restreint avec les variables sélectionnées
- Modèle complet avec toutes les variables

Validation croisée afin d'obtenir les **hyperparamètres** optimaux.

Les résultats amènent à :

- Nombre d'arbres : 500
- Profondeur maximale : 5

Prévisions sur les échantillons d'apprentissage et de validation. Les résultats montrent un **sur-ajustement très élevé** sur l'échantillon d'apprentissage et des mauvaises prévisions sur l'échantillon de validation.

Modèle restreint		
Métrique	Ensemble d'apprentissage	Ensemble de test
R ²	0.61	0.06
MSE	190 516	416 723
MAE	332.9	457

Modèle complet		
Métrique	Ensemble d'apprentissage	Ensemble de test
R ²	0.65	0.03
MSE	171 958	429 578
MAE	325.7	470.5

Random Forest : présentation

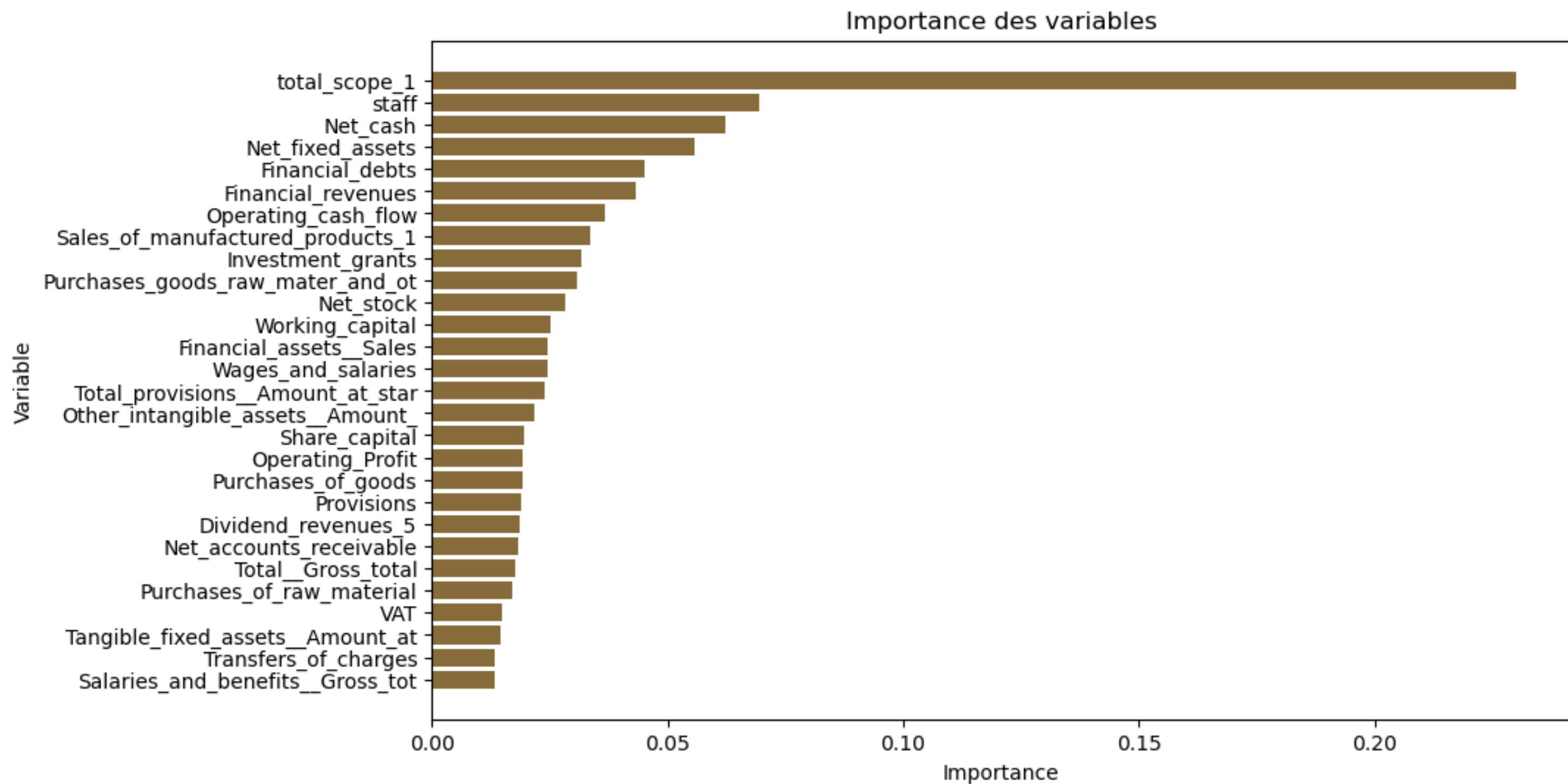
Métrique	Ensemble de test
R ²	0.33
MSE	65 625.6
MAE	196

Globalement les résultats sont mauvais. En effet, la moyenne des erreurs (MAE) est d'environ 196 alors que la moyenne du scope 3 pour les PME est de 230.

Les erreurs sont conséquentes comparativement aux vraies valeurs de la variable cible. Nous pourrions expliquer cela par la très petite taille de la population. Elle ne permet pas au modèle d'apprendre suffisamment des caractéristiques de ces entreprises.

En effet, l'échantillon de PME n'est pas représentatif de l'ensemble des PME en France, du point de vue de l'estimation de GES, et par conséquent, les performances obtenues sur cet échantillon sont peu fiables.

Random Forest : interprétation



L'analyse des importances des variables a été menée pour chaque modèle, mettant en lumière les **caractéristiques les plus influentes**.

Comme le montre le graphique ci-contre, les variables les plus significatives quant à la prévision des émissions de scope 3 des entreprises sont le **scope 1**, le **staff**, le **net cash**.

XGBoost : présentation

Mise en place de deux modèles :

- Modèle restreint avec les variables sélectionnées
- Modèle complet avec toutes les variables

Validation croisée afin d'obtenir les hyperparamètres optimaux.

Les résultats amènent à :

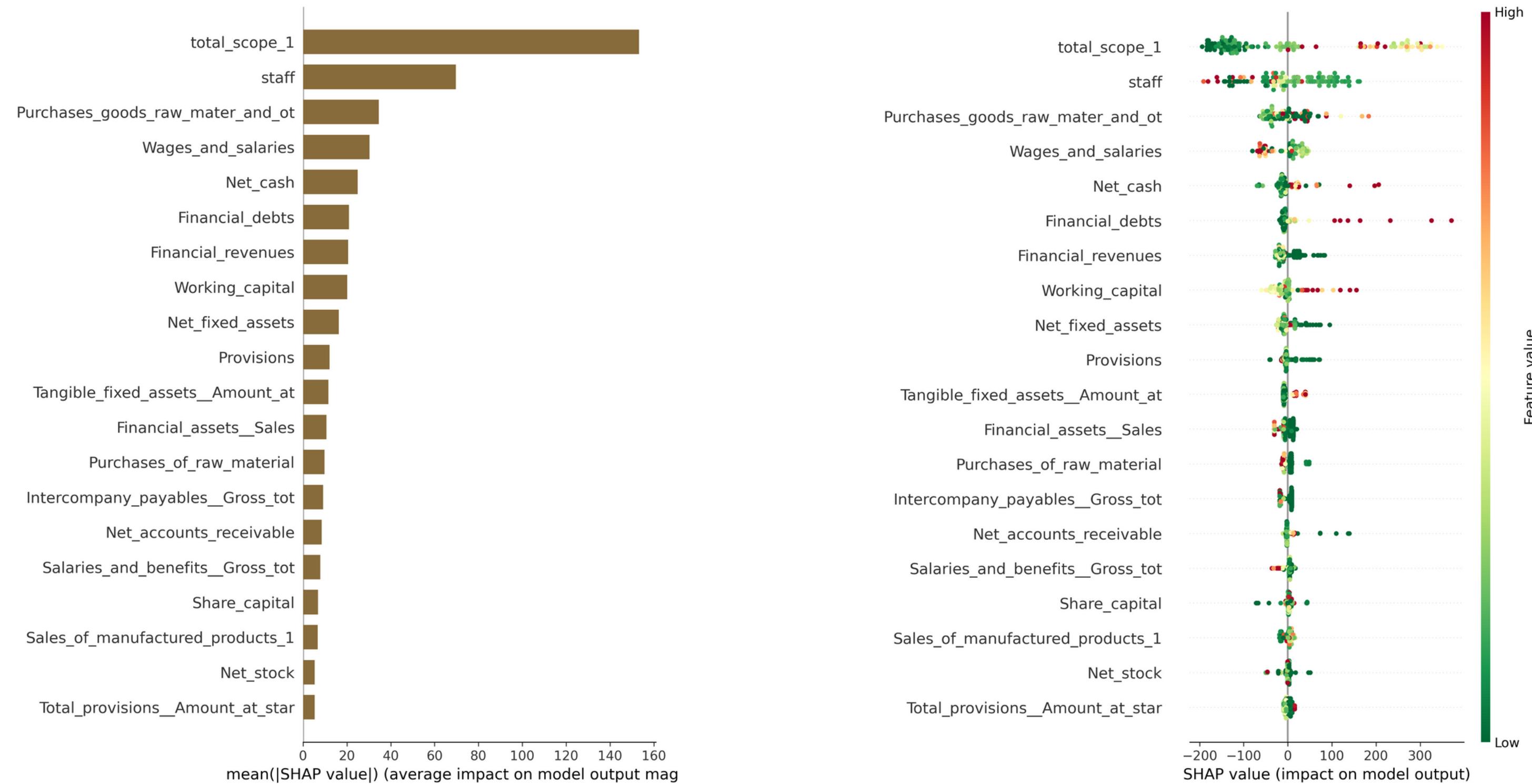
- Nombre d'arbres : 300
- Profondeur maximale : 3
- Taux d'apprentissage : 0.01

Prévisions sur les échantillons d'apprentissage et de validation. Les résultats montrent un **sur-ajustement très élevé** sur l'échantillon d'apprentissage et des mauvaises prévisions sur l'échantillon de validation.

Modèle restreint		
Métrique	Ensemble d'apprentissage	Ensemble de test
R ²	0.54	0.07
MSE	224 930	411 939
MAE	354.5	459.6

Modèle complet		
Métrique	Ensemble d'apprentissage	Ensemble de test
R ²	0.61	0.12
MSE	189 347	389 099
MAE	334.6	448.7

Shapley Values





VERS DE NOUVELLES METHODES



Prise en compte des effets spatiaux

Limites du Machine Learning classique sur ce type de données :

- Interprétabilité difficile : l'inférence n'est pas disponible
 - Des effets entre les observations ne sont pas pris en compte
-

N'existerait-il pas une méthode de Machine Learning qui permettrait de prendre en compte avec pertinence les dépendances entre chaque observations ?

En l'occurrence, nous avons décidé d'implémenter un modèle qui répond à ces attentes : le **GPBoost**.

Le modèle GPBoost

- Modèle développé par des chercheurs de l'Université de Lucerne en 2022.
- Ce modèle combine 3 outils de modélisation cruciaux :

Processus Gaussien

Mixed effects

Boosting

- Il permet de prendre en compte les dépendances entre les observations, ici en l'occurrence, les établissements.

Le modèle GPBoost

GP Boost

$$y = f(x) + Zb + \epsilon$$

fixed $\hat{}$

predictors

random effects

grouped
random

Type of business

Le modèle GPBoost

- Prise en compte des dépendances entre les observations, il procède à travers un covariogramme, fonction d'une distance entre deux points s et s' , notée : $h = ||s - s'||$.
- Covariogramme de forme exponentiel
 - Implications : “Cela signifie que la corrélation entre les valeurs diminue progressivement à mesure que la distance spatiale augmente”.
- Obtention de variables de géolocalisation : Longitude, Latitude.

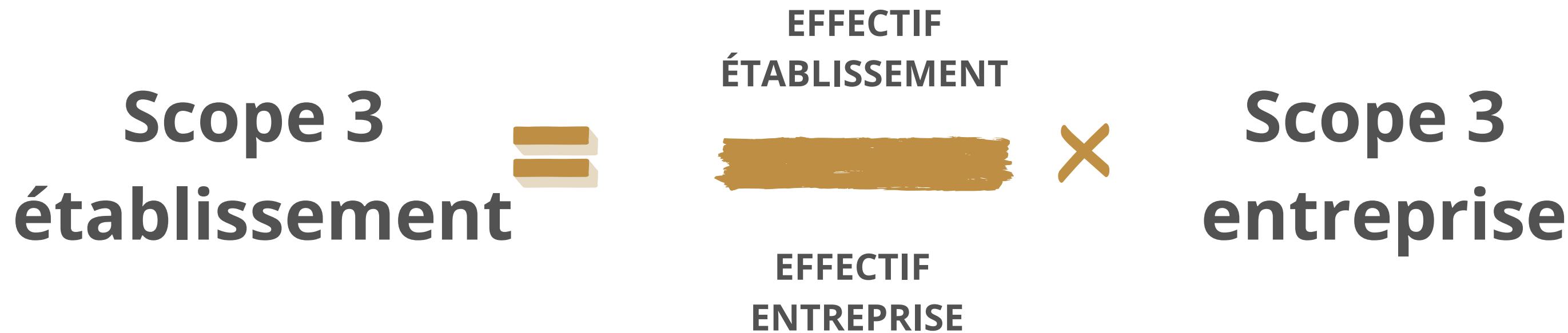


data.gouv.fr

Base Sirene des entreprises et de leurs établissements (SIREN, SIRET)

Le modèle GPBoost

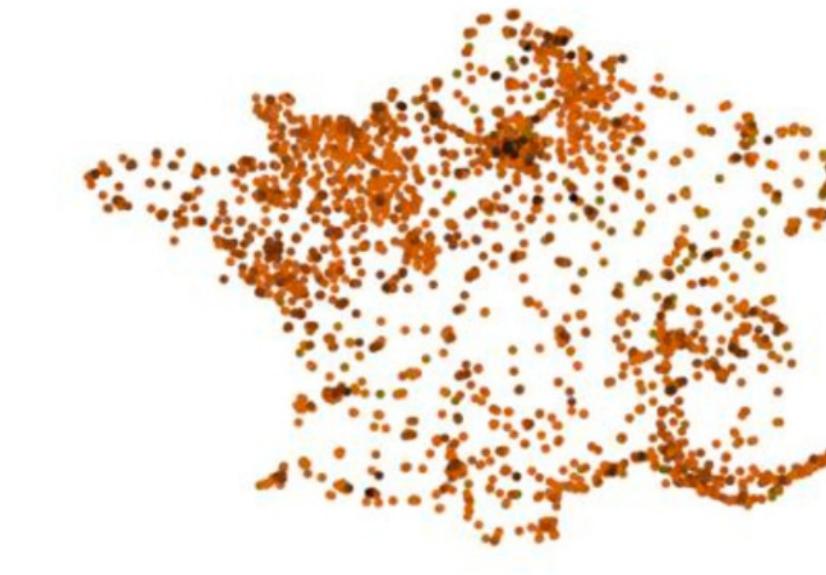
- Dans la base initiale, nous disposons de code SIREN.
- Avec cette nouvelle base jointe, nous disposons des SIRET (établissements) de ces entreprises avec leurs données de géolocalisation et leurs effectifs.
- Hypothèse forte : Nous avons donc défini une pondération pour quantifier le Scope 3 et ses prédicteurs en fonction de leurs effectifs.



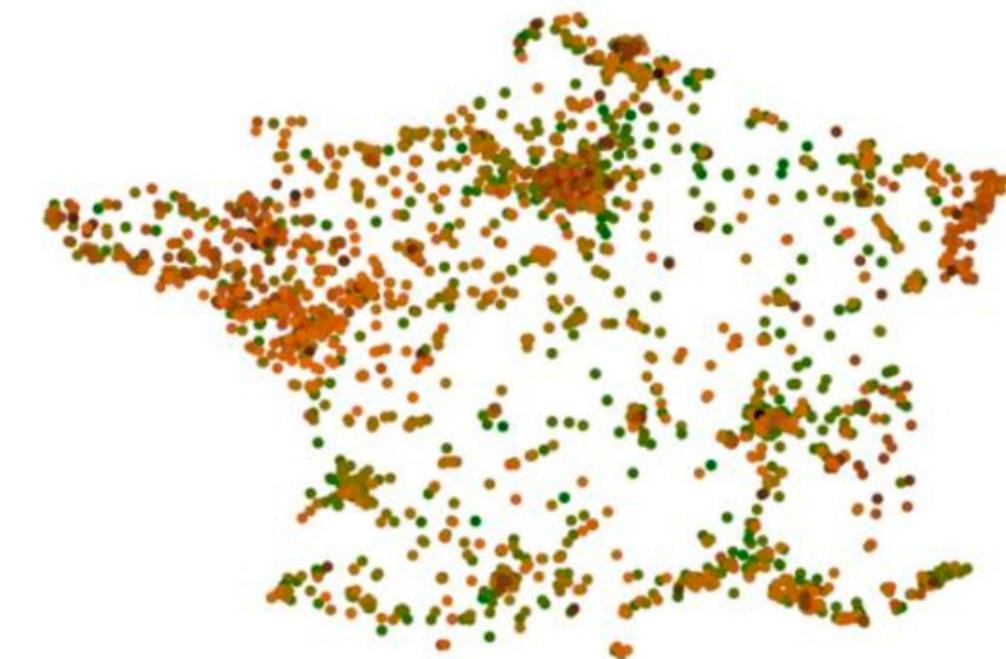
Géolocalisation des établissements



Transport et entreposage



Commerce ; réparation
d'automobiles et de motocycles



Hébergement et restauration

Résultats du GPBoost

Covariance parameters (random effects):

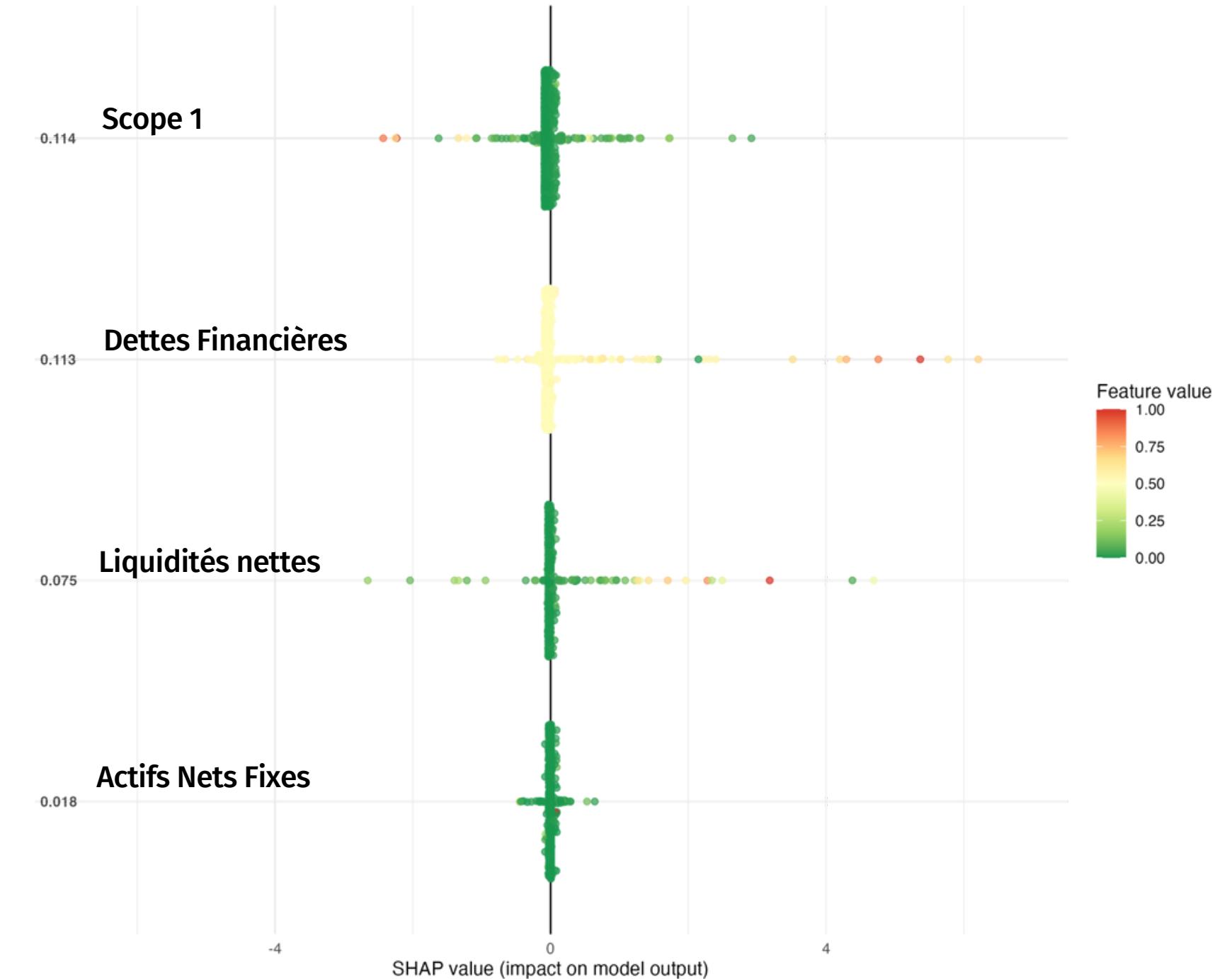
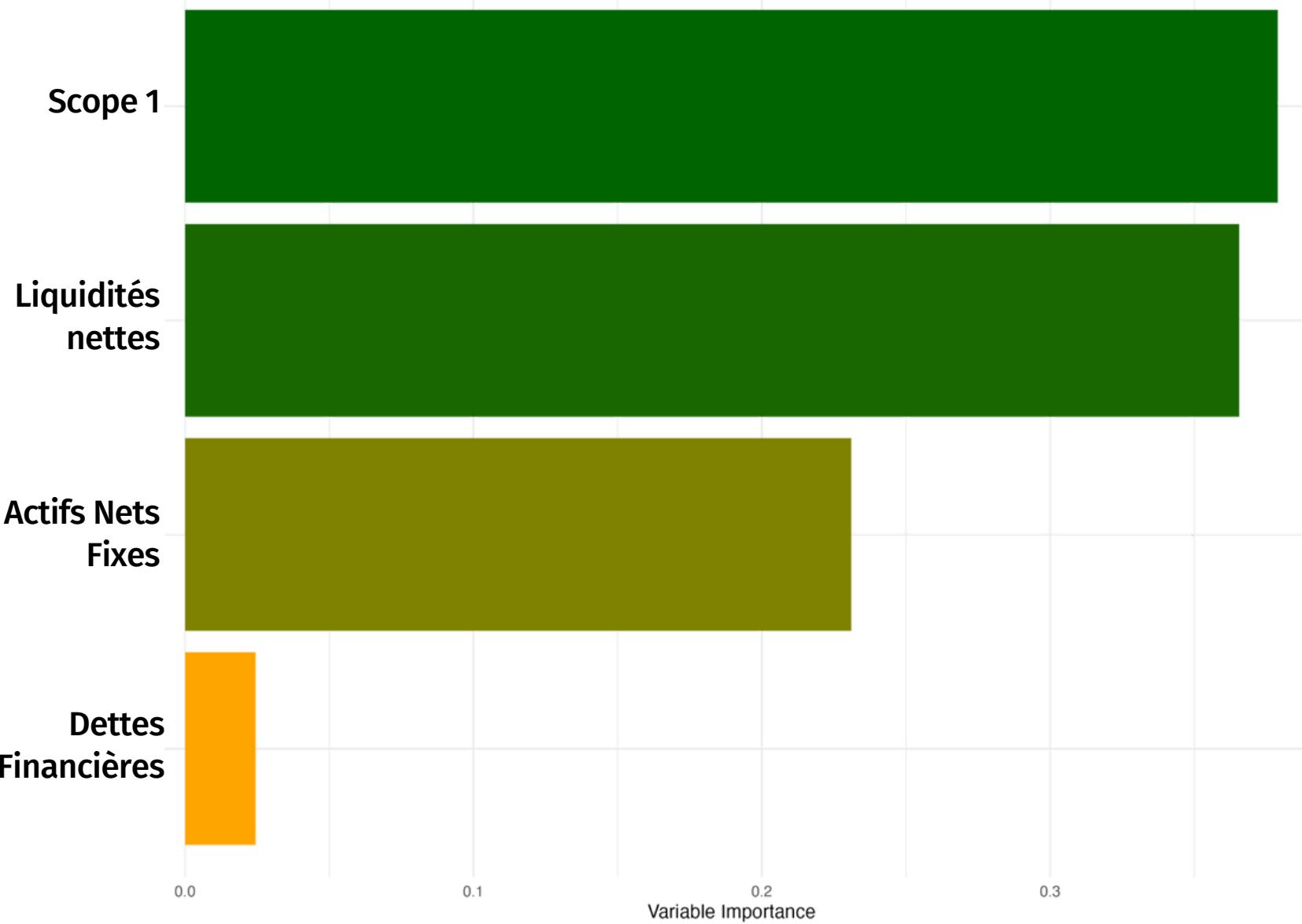
	Param.	Std. dev.
Error term	4155.7510	108.1866
GP var	9.2970	16.2316
GP range	1.5893	3.4692

Linear regression coefficients (fixed effects) :

	Param.	Std. dev.	z value	P(> z)
Constante	1.3922	1.6675	0.8349	0.4038
Scope 1	0.0783	0.0029	27.4404	0.0000
Dettes Financières	0.0000	0.0000	0.5188	0.6039
Actifs Nets Fixes	0.0000	0.0000	-13.1547	0.0000
Liquidités Nettes	0.0000	0.0000	22.8561	0.0000

R2	MAE (test)	MSE (test)
0.1346127	43.09404	21150.28

GPBoost, interprétations



ÉCONOMÉTRIE SPATIALE

Modèle SLX

- La scope 3 englobe les émissions indirectes d'une entreprise sur toute sa chaîne de valeur, incluant la production de matières premières, la sous-traitance, le transport, et l'usage des produits par les clients, correspondant à certaines émissions du scope 1 et 2 de ses partenaires.

$$Y = \beta_0 + \beta_1 X + \beta_2 W \times X + \epsilon$$

- Le modèle SLX étend l'analyse des variables indépendantes en incluant les effets des entités voisines via des variables spatialement laguées. Ces dernières sont obtenues en calculant une moyenne pondérée des valeurs adjacentes, à l'aide d'une matrice de poids spatiaux (W) qui détermine l'interaction et la proximité entre les unités.

Performances du modèle SLX

Base_ADEME_SIREN_Diane dispose d'informations spécifiques aux entreprises qui pourrait rendre plus opérationnelle la démarche des banques.

Métrique	Ensemble test
R2	0.24
MSE	365601.7
MAE	424.9503

- Le modèle spatial explique 24% de la variabilité du scope 3.
- La MSE et la MAE élevées révèlent un pouvoir prédictif faible de ce modèle.



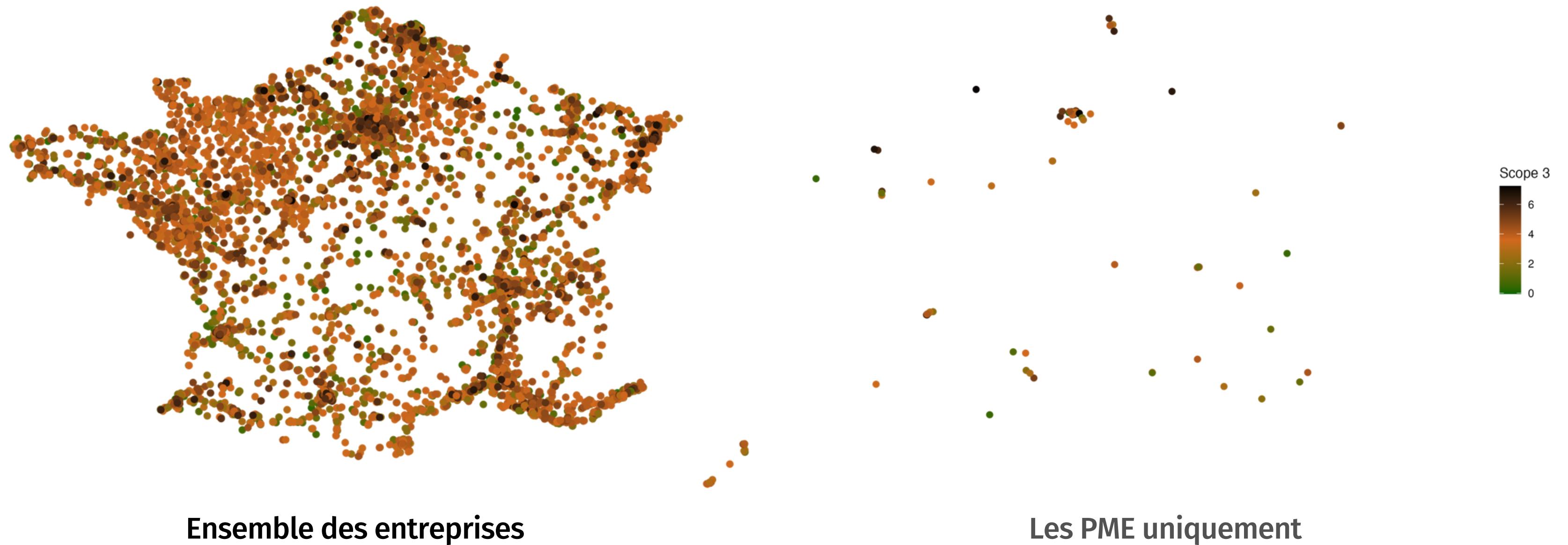
QUE RETENIR ?



Conclusion

- Les modèles de Machine Learning donnent de mauvaises performances en raison d'un **manque de données**.
- Les nouvelles méthodes comme le GPBoost sont des **alternatives crédibles** en combinant la puissance de Machine Learning et une capacité d'interprétation.
- L'économétrie spatiale fournit une réponse face à de l'information manquante notamment en prenant en compte les **dépendances sectorielles**.

Cartographie des établissements selon leur Scope 3



Limites de notre analyse

- Les données liées au Scope 3 des PME et TPE sont **rares** ce qui réduit drastiquement la capacité prédictives des modèles, tant d'économétrie classique que de Machine Learning.
- Limitation en terme de **puissance de calcul**, notamment sur l'utilisation des modèles de Machine Learning.
- **Le Scope 3 n'est pas le seul indicateur** à prendre en compte pour mesurer degré d'exposition d'un portefeuille au risque de transition.

Contacts



Lorenzo BARRAUD

barraudlorenzopro@gmail.com



Léo BRIAND

leobriand35@gmail.com



Gloria SOMAVO

somavogloria@gmail.com



Samuel VAUTIER

s37.vautier@gmail.com



The background of the slide features a dark, hazy industrial scene with several tall, dark smokestacks emitting large plumes of white and grey smoke against a dark, orange-tinted sky. A bright, circular light source, resembling the sun, is visible in the upper right corner.

square[®]
management

adWay



MASTER
ESA
Université d'Orléans